Review

JOURNAL OF UNIVERSITY OF BABYLON
For Pure and Applied Sciences (JUBPAS)

Vol.31; No.2.| 2023

Most Recent Malicious Software Datasets and Machine Learning Detection Techniques: A Review

Zahraa Najah[1]*, Wesam Sameer Bhaya[2]

[1]College of Information Technology, University of Babylon, zahraanajah.sw.msc@student.uobabylon.edu.iq, Babil, Iraq.

[2]College of Information Technology, University of Babylon, wesambhaya@uobabylon.edu.iq, Babil, Iraq.

***Corresponding author email:** zahraanajah.sw.msc@student.uobabylon.edu.iq; mobile:07805144316

أحدث مجموعات بيانات البرامج الضارة وطرق اكتشاف التعلم الآلي: مراجعة

زهراء نجاح*[1]، وسام سمير بهيه[2]

1 كلية تكنلوجيا المعلومات، جامعة بابل، zahraanajah.sw.msc@student.uobabylon.edu.iq ، بابل، العراق

2 كلية تكنلوجيا المعلومات، جامعة بابل، wesambhaya@uobabylon.edu.iq ، بابل، العراق

Received:	11/1/2023	Accepted:	16/4/2023	Published:	30/6/2023

ABSTRACT

Background:

Within the context of cyber security, it has become crucial to monitor systems and analyze data to maintain data security and integrity. Recently, it has become important to create a system for analyzing and classifying data, to prevent any malicious programs such as malware.

Materials and Methods:

The latest malware dataset and the latest machine-learning techniques were used to detect malware, based on dynamic feature identification.

Results:

The results showed that the FFNN algorithm was the best algorithm for the sorel20M dataset based on the research work discussed in this paper.

Conclusion:

The continuous increase in the number and types of attacks has led to a huge expansion in the variants of malware samples. Therefore, malware needs to be categorized into groups according to their behavior, influence, and characteristics. Given the fact that research and training are essential elements of cyber security, its constantly changing nature poses a great challenge. This study mainly aims to demonstrate the most recent malware dataset and modern machine-learning techniques of malware detection, based on dynamic feature selection.

Keywords:

cyber security, malware, data set, machine learning.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

الخلاصة

مقدمة:

في سياق الأمن السيبراني ، أصبح من الضروري مراقبة الأنظمة وتحليل البيانات للحفاظ على أمن البيانات وسلامتها. في الآونة الأخيرة ، أصبح من المهم إنشاء نظام لتحليل البيانات وتصنيفها ، بهدف منع أي برامج ضارة مثل البرامج الضارة.

طرق العمل:

تم استخدام أحدث مجموعة بيانات للبرامج الضارة وتقنيات التعلم الآلي الحديثة للكشف عن البرامج الضارة ، بناءً على اختيار الميزات الديناميكية.

الإستنتاجات:

أدت الزيادة المستمرة في عدد وأنواع الهجمات إلى توسع هائل في متغيرات عينات البرامج الضارة. لذلك ، يجب تصنيف البرامج الضارة إلى مجموعات وفقًا لسلوكها وتأثيرها وخصائصها. بالنظر إلى حقيقة أن البحث والتدريب عنصران أساسيان للأمن السيبراني ، فإن تغيير الطبيعة باستمرار يشكل تحديًا كبيرًا. تهدف هذه الدراسة بشكل أساسي إلى توضيح أحدث مجموعة بيانات للبرامج الضارة وتقنيات التعلم الآلي الحديثة للكشف عن البرامج الضارة ، بناءً على اختيار الميزات الديناميكية

الكلمات المفتاحية:

الأمن السيبراني ، البرامج الضارة ، مجموعة البيانات ، التعلم الآلي.

INTRODUCTION

Malware is a term used to describe programs or malicious codes that are developed to harm computer systems. Malware is used for disrupting system services, gaining access to systems, denying services, or stealing or modifying confidential data [1][2].

Malware can be classified into different types such as viruses, Trojans, spyware, and adware. The unstable growth of malware and good ware, and the increase in different families of malware create the demand for practically studying the classification of malware [3][4]. Generally, there are two types of malware analysis: static and dynamic. The analysis is considered to be static in case it is not run on a system, otherwise, it is considered to be dynamic [5].

In static analysis, the executable file undergoes analysis based on its structure without being executed within a controlled environment. There are numerous static characteristics of executable files, including memory compactness and various memory sections. One of the ways of extracting static characteristics out of executable files is the portable executable python library (PEFILE) [6].

As for dynamic analysis, the analysis of malware is performed within a dynamically controlled environment. During the execution of the malware, the registry keys witness a malicious change, after which the privilege mode of the OS is taken. The latter causes everything to change in the operating system [7]. The software can fully access the resources to be executed within the environment [8]. The software could alter the registry keys on the computer, as well as activate the debugger mode. After executing the malware, the environment is reverted to its past state, according to the snapshot made at the beginning of the setup. The behavior of the software is logged by the agent in the controlled environment. [6].

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

The main problem in any system is the efficiency of the system and its ability to detect malicious files accurately. Given the increasing number of ML algorithms, it becomes rather challenging to identify the ideal algorithm. In any case, this research discusses a group of research papers and will compare a group of databases that have been classified through ML algorithms.

ML techniques are used to identify and categorize malware into its categories and families, to separate the instances that exhibit new activity for in-depth study. This section discusses the literature works related to these approaches.

The authors in [9] present a flexible architecture that allows users to distinguish between clean and malware files using machine learning methods. In their research, one-sided perceptrons and kernelized one-sided perceptrons were used to reduce false positives and distinguish between malware and clean files.

In [10], it is suggested to figure out how the malware samples should behave. Many algorithms are used throughout this process in particular, including K-NN, Decision Trees, SVM, Naive Bayes, and Random Forest.

The work in [11] presents a methodology in which machine learning techniques are used to process data, classify malware, and detect new malware. Opcode n-grams, feature extraction, and grayscale images are all used in data processing. Malware classification is displayed in form of a decision-making model. The feed-forward neural network (FFNN) technique is used by the detection module to categorize malware families.

Most of the previous research works test the efficiency of a particular algorithm using only one database. In some cases, they test the efficiency of a group of algorithms on a specific database, while some algorithms differ in their efficiency from one database to another. An example of this is the FFNN algorithm. One of its advantages is that it can deal with large data, unlike the SVM algorithm, which is less efficient in the case of large data. Therefore, it has become necessary to test the algorithm on more than one type of data, which is what this paper aims at.

Materials and Methods

- **Data Sets and Machine Learning Methods**

  In its general meaning, the dataset is defined as “a collection of data”. Normally, data is represented in database tables, in which the columns indicate unique variables and every row stands for a specific record for a respective dataset. Furthermore, each variable is listed with different values [12].

  During the past years, different datasets have been used due to the increase in the number and type of attacks. Therefore, it has become necessary to generate and update the datasets for reducing attacks and improving security. Table 1 shows an example of a collected sample of a dataset [13][14].

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

Table 1. Samples of datasets collected in 2017

Profile	Period of collection	Malware	Benign	Unlabeled	Sum
Profile1	Jan - Feb	60K	50K	60K	170K
Profile2	Mar – Apr	60K	50K	60K	170K
Profile3	May – Jun	60K	50K	60K	170K
Profile4	Jul – Aug	60K	50K	60K	170K
Profile5	Sep-Oct	60K	50K	60K	170K
Profile6	Nov -Dec	100K	100K	0	200K
All		400K	400K	300K	1.1M

To educate machines on how to handle data more effectively, machine learning (ML) is used. The main reason for using ML is the fact that data cannot be evaluated or extrapolated [15]. To solve data challenges, machine learning uses a variety of algorithms. The type of algorithm used relies on many factors, such as type of the problem to be solved, the number of variables, the most suitable model type, and others [16].

This part contains two main sections. The first section demonstrates the most popular datasets used over the past years, while the second section will focus on the methods used for malware classification for each dataset.

- **Sorel- 20M Dataset**

This dataset is considered to be of a larger scale and consists of metadata and features that are extracted in advance. It also contains labels of high quality that are collected through different sources. Furthermore, 20M malware samples are found in this dataset, in addition to information about vendor detections at the collection time. It also involves tags about the aforementioned information with samples serving as extra targets [17]. Approximately, sorel provides 10 million malware samples whereby the optional_headers.subsystem and file_header. machine flags are set to zero, to be utilized when exploring features and detection strategies [18].

- **Soerel Detection Techniques**

There are two baseline machine learning algorithm models used on the sorel dataset. First of all, the feed-forward neural network (FFNN) model is used. The weights of input data are a key element in input data and data classification. Pre-training and data pre-processing are considered to be key components in creating effective methods for achieving quick training and high classification accuracy [19].

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

Secondly, the LightGBM gradient boosted the decision tree model. This is a commonly adopted ML algorithm that is known for being accurate, interpretable, and efficient. GB DT has been found to achieve state-of-the-art performances when executing different ML tasks, like multi-class classification, click predicting, and learning to rank[20]. However, the higher feature dimension and increased data size cause it not to be considered efficient or scalable to a satisfactory extent. This is mostly due to the need for teaching features for scanning every data instance to estimate the information gained via every possible split point. This procedure tends to cost a lot of time [21].

- **Soerel Classification Result**

This section shows the Soerel data set classification result after implementing the FFNN algorithm and the LightGBM algorithm.

Table 2.The Distribution of Behavioural Tags in the Training Set.

Adware	Flooder	ransomware	Dropper	spyware	packed	Installer	Worm
200M	500K	750K	2.75M	3.75M	2.5M	500K	2M

Table 3.The Distribution of Behavioural Tags in the Recommended Validation Set

Adware	Flooder	ransomware	Dropper	spyware	packed	Installer	Worm
100.5K	10K	100K	200.75K	200.5K	300K	100K	500K

Table 4.The Distribution of Behavioural Tags in the Recommended Test Set.

Adware	Flooder	ransomware	Dropper	spyware	packed	Installer	Worm
200K	1K	200K	400.5K	400.25K	500K	500K	500K

- **Ember Dataset**

One of the most popular malware datasets used with ML model training for detecting malicious executable files is the Ember dataset. It consists of 1.1 million extracted binary files, which include more than 900k training samples (300k benign, 300k unlabeled, 300k malicious)[13].

- **Ember Detection Techiques**

Gradient boosting decision tree (GBDT) is a common algorithm for ML models. The reason for its popularity is its efficiency, interpretability, and accuracy [22]. However, the big data terms (number of instances and number of features) cause several challenges to GBDT in the tradeoff between efficiency and accuracy. Therefore, GBDT needs every feature and scans all data to estimate information gain. This will increase implementation time when handling big data[23]

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

- **Ember Classification Result**

This section shows the results of the Ember dataset classification using Light GBM, as shown in the table below.

Table 5. Ember Dataset Classification Using Light GBM

Label	1	2	3	4	5	6	7	8	9	10	11
Unlabeled	28K	32K	10K	40K	30K	15K	32K	15K	18K	25K	0K
Benign	15K	25K	25K	15K	20K	15K	30K	10K	12K	25K	72K
Malicious	30K	25K	12K	50K	50K	10K	30K	32K	30K	29K	55K

- **Bodmas**

It can be described as an open PE malware data set used in facilitating research efforts within ML-based malware analyses. It involves 57,293 malware samples and 77,142 benign samples gathered between August 2019 and September 2020, using family information that underwent careful curation (581 families)[14]. A preliminary analysis is performed for illustrating the effect of concept drift and discussing the ways through which the data set could contribute to current as well as future research efforts.

- **Bodmas Detection Techiques**

In the BODMAS dataset, the Gradient Boosted Decision Tree (GBDT) classifier is used like the Ember dataset.

- **BodmasClaasification Result**

This section illustrates the results of the LightGBM algorithm on the BODMAS data set, as presented in the table below.

Table 6. BODMAS Dataset Classification Using Light GBM Algorithm Results

Phase	BODMAS	
Validation	Benign	Malware
Test 10/19	3925	4549
Test 11/19	3718	2494
Test 12/19	6120	4039
Test 01/20	5926	4510
Test 02/20	3703	4269
Test 03/20	3577	4990
Test 04/20	5201	4640
Test 05/20	6121	5449

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

- **New Dataset for Dynamic Malware Classification:**

This research introduces two types of datasets. The first dataset of 9795 samples was obtained from simple sharing, and the second dataset was obtained via virus share. The new dataset researches also analyze the performance in terms of the balance and imbalance of the multi-class malware using RF, SVM, Histogram-based gradient boosting, and XGBoost[24].

- **Histogram-based Gradient Boosting (HGB):**

It is a popularly used ML algorithm known for its various implementation domains, which makes it easy to manage concerning how complex the model is, using tree depth and the number of trees [24]. HGB is found to be significantly faster when applied to larger datasets and provides native support for missing values found in the dataset, which is a key feature b[25].

Table 7. HGB Algorithm Result

Data type	HGB
Adware	89%
Agent	91%
Backdoor	95%
Trojan	75%
Virus	95%
Worms	92%
Total	89.5%

- **Random Forest (RF)**

RF is a popular machine-learning algorithm. It has developed into a widely used non-parametric approach that may be adopted in classifying or regression issues [25]. To obtain more precise predictions, the RF algorithm builds numerous decision trees and then combines them [26].

- **Random Forest Result**

This section shows the results of the RF, as illustrated in the table below.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

Table 8. Random Forest Result

Data type	HGB
Adware	89%
Agent	91%
Backdoor	95%
Trojan	75%
Virus	95%
Worms	92%
Total	89.5%

- **Support Vector Machine (SVM)**

SVM is a supervised ML algorithm that is commonly applied for classifying tasks in complex data sets. The benefit of adopting SVM is its increased efficiency when applied in high-dimensional spaces [26]. Besides, it provides more accurate results and can be used with a larger number of independent variables [1].

Table 9. SVM Algorithm Results

Malware type	(SVM)
Adware	96%
Agent	91%
Backdoor	94%
Trojan	76%
Virus	97%
Worms	92%
Total	91%

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

As a result of classifying the new dataset using the HGB algorithm, the RF algorithm, and SVM, the obtained accuracy rates were 89.5% for both the HGB and RF algorithms, whereas the accuracy of SVM was 91%. Therefore, the SVM is considered to be more efficient than the RF and HGB algorithms.

Results and Discussion

This paper discussed and reviewed a set of research works that dealt with the topic of datasets and machine-learning algorithms used for malware classification. The first research explained how the Sorel dataset is classified using the FFNN algorithm, where the results were accurate and efficient. The Ember database was classified using the GBDT algorithm, which showed the ability to classify the data in a semi-efficient manner. The disability of GBDT was due to the time consumed by the algorithm because it requires all properties of the elements in the database. This leads to a slowdown in the work of the algorithm. This is also the case with the BODMAS database when classified using the GBDT algorithm. As for the new dataset, it was classified by three algorithms: Random Forest, SVM, and HGB. As a result, the SVM algorithm was the best algorithm obtaining an accuracy of 91%. This led to the conclusion that the best algorithm is FFNN after implementing it on Sorel 20M. The reason for choosing this algorithm is that it has been tested on a dataset that is considered to be relatively larger as compared to the rest of the databases. It also contains a larger number of types, so it can be concluded that it is more efficient.

Table 10. FFNN Result

FFNN	
Malware type	**Result**
Adware	200M
Flooder	500K
Ransomware	750K
Dropper	2.75M
Spyware	3.75M
Packed	2.5M
Installer	500K
Worm	2M

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

Conclusion

Malware classification is a significant field of study. This paper has reviewed a group of research papers that dealt with the topic of malware classification using machine learning algorithms. The malware data was obtained from popular malware datasets which are the Ember dataset and the Sorel dataset. Moreover, the datasets were collected by researchers such as BODMAS and the new dataset. It was found that the FFNN algorithm was the best algorithm for the sorel20M dataset based on the research work discussed in this paper.

Conflict of interests.

There are non-conflicts of interest.

References

[1] B. Gencaydin, C. N. Kahya, F. Demirkiran, B. Duzgun, A. Cayir, and H. Dag, "Benchmark Static API Call Datasets for Malware Family Classification," *Proc. - 7th Int. Conf. Comput. Sci. Eng. UBMK 2022*, pp. 137–141, 2022, doi: 10.1109/UBMK55850.2022.9919580.

[2] B. Narayanan and V. S. P. Davuluru, "Ensemble Malware Classification System Using Deep Neural Networks," *Electronics*, vol. 9, p. 721, Apr. 2020, doi: 10.3390/electronics9050721.

[3] R. Ronen, M. Radu, C. Feuerstein, E. Yom-Tov, and M. Ahmadi, "Microsoft malware classification challenge," *arXiv Prepr. arXiv1802.10135*, 2018.

[4] C. C. Uchenna, N. Jamil, R. Ismail, L. K. Yan, and M. A. Mohamed, "Malware threat analysis techniques and approaches for iot applications: A review," *Bull. Electr. Eng. Informatics*, vol. 10, no. 3, pp. 1558–1571, 2021, doi: 10.11591/eei.v10i3.2423.

[5] K. Sethi, R. Kumar, L. Sethi, P. Bera, and P. K. Patra, "A novel machine learning based malware detection and classification framework," in *2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, 2019, pp. 1–4.

[6] M. Ijaz, M. H. Durad, and M. Ismail, "Static and Dynamic Malware Analysis Using Machine Learning," *2019 16th Int. Bhurban Conf. Appl. Sci. Technol.*, pp. 687–691, 2019, doi: 10.1109/IBCAST.2019.8667136.

[7] M. Hassen, M. M. Carvalho, and P. K. Chan, "Malware classification using static analysis based features," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–7.

[8] C. V. Gonz, "Towards Explaining the Effects of Data Preprocessing on Machine Learning," *2019 IEEE 35th Int. Conf. Data Eng.*, pp. 2086–2090, 2019, doi: 10.1109/ICDE.2019.00245.

[9] O. Bawazeer, T. Helmy, and S. Al-Hadhrami, "Malware Detection Using Machine Learning Algorithms Based on Hardware Performance Counters: Analysis and Simulation," *J. Phys. Conf. Ser.*, vol. 1962, no. 1, 2021, doi: 10.1088/1742-6596/1962/1/012010.

[10] K. Chumachenko, "Machine learning methods for malware detection and classification," 2017.

[11] L. Liu, B. Wang, B. Yu, and Q. Zhong, "Automatic malware classification and new malware detection using machine learning," *Front. Inf. Technol. Electron. Eng.*, vol. 18, no. 9, pp. 1336–1347, 2017.

[12] A. A. R. Melvin *et al.*, "Dynamic malware attack dataset leveraging virtual machine monitor audit data for the detection of intrusions in cloud," *Trans. Emerg. Telecommun. Technol.*, vol. 33, no. 4, pp. 1–19, 2022, doi: 10.1002/ett.4287.

[13] Y. Oyama, T. Miyashita, and H. Kokubo, "Identifying useful features for malware detection in the ember dataset," in *2019 seventh international symposium on computing and networking workshops (CANDARW)*, 2019, pp. 360–366.

[14] L. Yang, A. Ciptadi, I. Laziuk, A. Ahmadzadeh, and G. Wang, "BODMAS: An open dataset for learning based temporal analysis of PE malware," in *2021 IEEE Security and Privacy Workshops (SPW)*, 2021, pp. 78–84.

[15] B. Mahesh, *Machine Learning Algorithms -A Review*. 2019. doi: 10.21275/ART20203995.

[16] D. Dhall, R. Kaur, and M. Juneja, "Machine learning: a review of the algorithms and its applications," *Proc. ICRIC 2019*, pp. 47–63, 2020.

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

[17] D. Serpanos, P. Michalopoulos, G. Xenos, and V. Ieronymakis, "Sisyfos: A modular and extendable open malware analysis platform," *Appl. Sci.*, vol. 11, no. 7, p. 2980, 2021.
[18] R. Harang and E. M. Rudd, "SOREL-20M: A large scale benchmark dataset for malicious PE detection," *arXiv Prepr. arXiv2012.07634*, 2020.
[19] R. Asadi and S. A. Kareem, "Review of feed forward neural network classification preprocessing techniques," in *AIP Conference Proceedings*, 2014, vol. 1602, no. 1, pp. 567–573.
[20] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[21] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electron. Commer. Res. Appl.*, vol. 31, pp. 24–39, 2018.
[22] R. Kumar and S. Geetha, "Malware classification using XGboost-Gradient boosted decision tree," *Adv. Sci. Technol. Eng. Syst*, vol. 5, pp. 536–549, 2020.
[23] H. S. Anderson and P. Roth, "Ember: an open dataset for training static pe malware machine learning models," *arXiv Prepr. arXiv1804.04637*, 2018.
[24] D. Tang, S. Zhang, Y. Yan, J. Chen, and Z. Qin, "Real-time Detection and Mitigation of LDoS Attacks in the SDN Using the HGB-FP Algorithm," *IEEE Trans. Serv. Comput.*, 2021.
[25] C. D. Morales-Molina, D. Santamaria-Guerrero, G. Sanchez-Perez, H. Perez-Meana, and A. Hernandez-Suarez, "Methodology for malware classification using a random forest classifier," in *2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, 2018, pp. 1–6.
[26] E. Gandotra, D. Bansal, and S. Sofat, "Malware analysis and classification: A survey," *J. Inf. Secur.*, vol. 2014, 2014.

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية