A comprehensive study on high performance malware classifiers based on machine learning algorithms

Saifaldeen H. A.
Computer Sciences Dept.University of Mosul
Saifaldeenhasanali@gmail.com
Karam H. Thanoon
Computer Sciences Dept. University of Mosul
karamhatim@uomosul.edu.iq

Abstract

Malware as a malicious software has been developed and became an interest issue that take great attention of the researchers and companies that delt with data security. Therefore, determining the classes of these malwares is very important to detect further newer or modified versions that are continuously developed. Many classifiers had been developed to implement them to build newer detectors to that are able to secure data. This paper, as comprehensive study illustrates the performance of the most common malwares and highest performance classifiers that were recently proposed. The study shows that there are three proposed classifiers with highest accuracy. They are: Random forest, Support Vector Machine and x-gradient boots. The number of features had a great effect in classification process in which the accuracy decreased if the features number increased. Additional techniques may enhance classification accuracy such as New Feature Engineering.

Keywords: Machine learning, Malware Classification, Classifier, Random Forest, Support Vector machine, Gradient Boost.

1. Introduction

Malwares are the most intractable subjects in the cyber security. In such development of internet connection and applications, more security and privacy issues had faced the users of different types of internet websites and applications. Those issues affected on performance of the users' devices and internet connections. Thus, more applications had been developed to detect malwares which had some similarity in their features. The classifiers managed the similarity of the malwares to classify them in separated families[1].

Classifying malwares is complex procedures that implement more computing resources such as Machine Learning techniques. Basically, ML is the backbone of the Malware classification methods. Other developers had interest in Deep Learning based method which are out of this paper subject. Many classifiers had been proposed that are based on ML to classify malwares in families. These classifiers are variant according to their proposed ML algorithms. This variance must be evaluated to consider the most effective classifiers. But, each of classifier had several versions that represent an improvement in its performance [2].

Several survey or comparative studies had been proposed to evaluate the common classifiers that are based ML algorithms. In[1], Random Forest(RF) and Support Vector Machine(SVM) had been illustrated in a complex mythology to evaluate different classifiers with a proposed criteria. In [3], the RF, gradient Boost and SVM had been implemented in a comparative study to evaluate their performances. Six classifiers had been implemented with 12 detectors. In [4], the authors hasd implement RF, GB and SVM and other classifiers in Adversarial Conditions.

This paper illustrates recently proposed classifiers that had higher performance according to previously presented studies [1][3][4]. It aims to achieve the major goal via surveying the recent studies that implements the higher performance classifiers. It also aimed to evaluate the specifications and performance of these classifiers' versions.

2. Malware

Malwares are the most affective threats via launching a cyber-attack that aims the security and privacy of information. Malware, simply is a malicious software such as virus, rootkit, worm, ransomware and backdoors, that has malicious activities in which devices and connections are affected[5]. The malwares can be classified in the following major families:

a) Virus:

It is firstly known family of malwares in early of 80s of the 20thcentury. The virus is a malicious software that causes damages of user's devices and affects of connections. There are many types of viruses that variant according to their effects and type of attack[6].

b) Trojan Horse:

Trojansare packages of programs that are not predictable and has no data.They are implanted in the host(victim) devices. As a client server program, trojans are remotely activated. When a device is infected, Trojan is a database used to transfer a significant function that may harm user's devices. Those functions are concealed and unrecognized. Its package performs unknown functions that transfer users data to the trojans builder [7].

c) Spam

Spam is a malicious software embedded in anemail that sent by intruder to anyone via internet transmission. The spam showed unrelated information, but the malicious software gathers some user files that related with user behavior and interests. The effects of spam represented with the delay time according to spam activity that slows the internet speed because unauthorized traffic. The crowded traffic slows internet response [7].

d) Worm

The worm is malicious software that is similar to Trojan Horse but it can duplicate it self on the infected devices. The successful implanted worm duplicates its script via sudden discovered vulnerabilities and network connection. The duplications of worm scripts are performed separately without control of the original script. Worms decrease the user's device response [7].

e) Spyware

Spyware is a malicious software that records the activity of any network connections. It may detect any personal information such as Personal Identification Number (PIN) of user's accounts, or any password of other internet webpages and applications. In which information will be directed to the intruder. Most spyware are embedded into adult webpages or other downloading sites. The spyware is concealed into the desired software that infects user's devices and reduced their performance. Spyware must be treated accurately because it may remain in the device even when recovery stage [5].

f) Adware

Adware is malicious package or product that its script concealed into the website or applications contents. The script of website or application activated adware. But some devices prevent its activity. Thus, some adware has capability to overcome that prevention and activated during running applications[6].

g) Backdoor

Backdoor is a malicious software that enables the attackers to overcome authentication processes that are required during accessing to servers and accounts. This software is activated remotely in the victim device to overcome required authentication information. Then, the Backdoor allows the attackers to manage files and servers without performing any authentication and gives them the ability to control victim remotely and update its information[8].

h) Downloader

Downloader is a malicious software the concealed into executable files that are downloadedin a screed state. While a part of these files contain request to download more parts to build a malicious software in the victim's device. In this case, the detection process is more difficult because downloaded process are not malicious software or their complements[9].

i) Dropper

Dropper is a malicious software that concealed into executable files. The Dropper has payloads that are installed and have malicious effects on victim's device. Droppers may conceal different types of malware that will infect the victim's device [10].

j) Ransomware

Ransomware is a malicious software prevent victim from accessing his own files or may locks his device unless the attacker receive amount of money as a ransom for its malicious effects. Usually, ransom is paid in bitcoins that had nourished these schemes at the last two decades [11].

3. Machine Learning

The term Machine learning belongs to the middle of the 20th century. It showed the capability of machine (Computer) to learn and gain knowledge. It illustrates the methods of computer to acquire knowledge from Data to solve statistical problems. Thus, ML methods were used in several statistical issues such as Regression. But these methods were built depending on statistical rules and laws. Later, ML developed and mimicked the human way of thinking that enables computers to learn real information from data and solve more complicated problems. ML have the benefit of the computational power of the PC to deal with big data and uncover the nested structure in big data.

The researchers in this field considered as data mining tool. They recognized the three essential problems:

- Regression (Predicting a continuous outcome variable),
- Classification (Predicting a categorical outcome variable),
- and Clustering (Dividing a population of individuals into k subpopulations such that individuals within a group are as similar as possible, and the individuals among the groups are as dissimilar as possible).

They built and developed their algorithms on PC for a nonstatistical, assumption-free nonparametric approach.The Datamining field has many specialtiessuch as: support vector machines, neural networks,fuzzy logic, genetic algorithms and programming, information retrieval, text processing,knowledge acquisition, inductive logic programming, expert systems, and dynamic programming. All thesefields and methodshave the same objective [12].

4. Malware Classification

Incremental usage of the Internet complicates the issues of cybersecurity especially in commercial and e-payment fields that faced by vulnerable of Malwares. To avoid the effects of these malwares, researchers had to discover the type of malwares to be detected and quarantined. Millions of malwares have been developed and many detecting algorithms had been built to detect them. The large number of malwares led to implement many classification algorithms to classify these incremented numbers of malwares in order to deal with their malicious effects.

At the beginning, many traditional methodologies had been implemented to classify malwares. Gradually ML had been implemented in malware classification. Basically, the classification algorithms realized on: Static, Dynamic and Hyper analysis to identify malwares via malware signature or behavior at their environments. The ML based classification algorithms achieved high performance in comparison with the traditional ones[13].

5. High performance classifiers

Many studies had been proposed that implement different algorithms in building efficient classifiers. Such classifiers had variant level of accuracy. Therefore, the listed classifiers are the higher performance among the recently proposed classifiers [1][3][4]. They are:

a) Gradient Boost (GB)

Boosting technique essentially is a class ensemble learning algorithms in which the weak learners had been combined.They form stronger modelsthat have better accuracy level of their predictions such as “IF ELSE” prediction rules. Boosting techniques are used in developed algorithms by combining weaker and well-known algorithm to build newer algorithms such as is AdaBoost. In same way, Gradient Boosting is a high-performance Boosting algorithm in which many decision trees models have been sequentially built based on residuals of previous models such as prediction errors [14].

b) Support Victor Machine (SVM)

Support Vector Machine is supervised machine learning algorithm. The researchers implement it to build a malware classifier that used to analyze the data to recognize the verity of patterns for purpose of classification. The SVM based classifier generates a hyper plane from the patterns’ series of different class. This SVM based linear classifier is mathematically declared as in (Eq.1).[15]

$$f(x) = WT\ X + b \quad (1)$$

c) Random Forest (RF)

Random Forest is a popular machine learning algorithm. It does not require neither data preparation nor modeling.It usually produced accurate results. RF designis based on the decision trees. A collections ofdecision trees are used to producehigher accuracy ofprediction. Therefore, it is known as ’forest’ because “it is basically a set of decision trees”.The design is bas on growing multiple decision trees of the independentsubsets in a dataset. The classification methods distributed n variables out of the feature set at each node randomly.Then, classification performed depending on the best split of the variables[16].

6. Recent proposed Studies

In this section, the most recently prosed methods would be illustrated these methods implemented the highest and most effective classifiers that are recently proposed. They are:

a) Gradient Boost (GB)

1) In [17], the authors had proposed an enhanced feature extraction by implementing three different features of images. Their method based on converting the Android application to RGB images and convert the malware dataset to RGB images too. Five machine learning and two Deep learning algorithms had been used to classify the malicious software depending on Image analysis. The experimental results showed high accuracy level for gradient boost algorithm compared to other algorithms.
2) In [18], the authors proposed a model that trained the Android data by using three global image features. While the dataset had been trained using four additional local features. The model based on 6 machine learning algorithms. The experimental result showed a good performance of Gradient Boost algorithm compared to other five algorithms.
3) In [19], [20] and [21], the authors proposed a new Gradient boost classifier based on New Feature Engineering(NFE) technique. In which, importance of features had been evaluated to be used to build feature ranking that had been used with feature occurrences to build the model. Gradient Boost algorithm had been implemented with 10 top ranks of features, if the accuracy was less than 90% more features are used to build the model. The gradient Boost and other algorithms had been implemented; Gradient Boost had the highest level of accuracy.
4) In [22], the authors proposed a new model for malware classification that extracts raw bytes, n-grams of byte codes, PE imports, strings features, and PE section names as the input features. This model implements Gradient Boost algorithm to extract and classify malicious software either in training or testing. The experimental results showed high level of accuracy.

b) Support Vector Machine (SVM)

1) In[23], the author proposed an HMM based model to detect malware. Then he had proposed a comparison among CNN and SVM classifiers. Both classifiers had been tested for 2 to 10 families. The SVM showed higher performance while CNN shoed drop

in accuracy while increasing families' number. The result showed 99% of accuracy for SVM.

2) In [24], the authors had implemented Combination of three features to classify malwares in prepared dataset of malware that had been collected from estimated malware libraries of antivirus application. In order to achieve comparison, several algorithms were to determine their accuracies values. The SVM algorithm had obtained high accuracy level according to the proposed extraction method.
3) In[25], the Authors implemented One-against-all SVM as multiclassification algorithm to establishes N decision boundaries for N classifications. In which,each decision boundary determined the one attribution of classification to all other classifications. This property enhanced the standard SVM algorithm via improving its optimization problem. The experimental results showed that SVM needs 16 features to reach stable accuracy.
4) In[26], the authers had implemented Multi class SVM algorithm as machine learning method to be compared with deep learning method based on convolutional neural network (CNN) algorithm. The results showed more accuracy of CNN algorithm but it showed high performance of accuracy for multi class SVM algorithm.

c) Random Forest

1) In[27], the authors had implemented Random forest algorithm as a supervised machine algorithm in combination with Principal Component Analysis (PSA) algorithm as unsupervised feature extraction and selection algorithm. The results of comparison showed higher performance than other common algorithms. It gained 89.67 F score as highest result.
2) In [28], the authors had tested n-gram opcode with hash bit sequence to extract features in a model based on Random forest algorithm. The result showed high performance when 2-grams initially. But the accuracy increased when more hash bit had been used in more n-gram cases.
3) In [29], the authors had presented a comparison between the classical Random Forest and DNN with different number of layers that the network is consisted of. The results showed that classical Random Forest had higher performance than the DNN because of the DNN irrespective of the features inputs.

7. Comparison and Discussion

After the illustration of the recent proposed work in Malware Classification that based on Machine Learning Algorithms. A comparison would been performed to evaluate the performance of the proposed works. This comparison has neem build upon pre-determined criteria that allows researchers and readers to measure the performance of these algorithms.

As shown in Table.1, the reduction of the feature number increases the accuracy of classification algorithm. Most proposed models try to detect the meaningful and important features. The Important of features has many ways to be determined such as NFE, Selection and some other statistical methods. In other hand, the number of instances had great effects on the performance of the algorithms. Image based algorithms and its graphical features had less accuracy levels.

Table 1: Comparison Classification Algorithms.

Method	Study	Features	Enhancement	OS	Dataset	Accuracy %	False Positive Rate %	Note
GB	16	- Scale-Invariant - Transform, - Speeded-Up Robust - Oriented - FAST Rotated BRIEF - KAZE - ColourHistogram, - Hu Moments - Haralick Texture. - 4850 samples:	- Converting data to Gray-scale Images. - Manifest image dataset-based. - DEX.	- Android	- "5"grayscale image datasets - Manifest.xml	91.96	8.04	- SIFT achieved higher accuracy than Manifest image
GB	17	- 4 local features - 3 global features	- Converting data to Gray-scale Images. - Manifest image dataset-based. - DEX.	Android	- "3" grayscale image datasets Manifest.xml	97.5	2.5	- Higher accuracy ontaind by using Kaze Feature and DEX
GB	18	- NFE - "11" Important	- The bag of visual words	Windows	- Malware Training - Sets: A	94	6	- NEF features obtained

		features. -	algorithm - NFE Features		machine learning dataset for everyone			Higher accuracy
GB	19	- NFE - - 30 important features -	- Feature standardization - NFE Features	Windows	- "Malware Training - Sets: A machine learning dataset for everyone - Cardiff University dataset	87	13	- GB had achieved higher accuracy.
XGB	20	- NFE - Feature Selection - 14 Feature - 100,000 samples	- NFE - Feature Selection	Android	- McGW - PgMW	99.71	0.29	- Feature selection may causes more accuracy
GB	21	- Fixed Byte Embedding - N-Grams of Code Bytes - Pe Imports and Pe Section Names - String Patterns	- Byte n-grams - Number of families -	Ubuntu Python	- Special small dataset. - Malshare	99.03	0.97	- Increasing families decreases the accuracies
SVM	22	- 25 higher occurring opcodes. -	- Opcode selection	Windows	- Special Design Large Dataset	99	1	- Better than CNN and HMM specially with increasing number of familes
SVM	23	- gray-scale images, - Opcode n-gram,	3-gram sequences. Import functions	Windows	- ESET NOD32 VX Heavens - Threat Trace Security	81.9	18.1	- SVM has the middle accuracy result

		- Import functions. - 25 Opcodes. - 23 Extracted features	selection					- While n-gram =2 or 3, the accuracy is high.
SVM	24	- Probability scoring - n−grams - 174607 malware samples - 63 malware families	probability scoring n−grams	Windows	- VX Heaven	98.82	1.18	- Dynamic Analysis need to be improved
Multi SVM	25	- DebugSize - latRVA - ExportSize - ImageVersion - ResourceSize - VirtualSize - Number of Sections - 5724 Instances	Converting categorical data to numerical data Using MultiClass SVM	Unknown With Python	- Special Design Dataset that suits the project. With 5724 instances	98.94	1.06	- debugSize - and malware name had higher effect in classification.
RF	26	- Principal Component Analysis PCA - Term frequency-inverse document frequency TF-IDF. - 5184 samples - 55 features	PCA TF-IDF	Windows	- ClaMP Raw-5184	89.83 96	10.17 4	- RF is more promising with TF-IDF
RF	27	- ASM instruction frequency. - Opcode 2-gram. - Control STMT	- Control STMT Shingling. - Hash bit string of 10 bit	Unknown With JAVA GUI	- Microsoft Malware Classification Challenge - “9” Malware Families	99.21	0.79	- Combination of techniques has been merged. - Less N-gram are

		shingling. - DLL Boolean feature. - Hash bit - 10867 Instances						better.
RF	28	- Adaptive Synthetic (ADASYN). - 616 opcodes.	None	- Lunix	- Benchmark Malicia Project. - Partial amount of VirusTotal.	98.93	1.07	- RF proved to obtain better results than DNN

In fact, the different datasets that are used in the proposed classifiers are serious problem that does not refers to standard methods of evaluations. Thus, the proposed method should be performed to the same Datasets to ensure their performances.

8. Conclusions and Recommendations

After deeply study of the recently proposed classifiers and extracting the comparison criteria, this paper concludes that the image-based methods had less accuracy levels. The raw-based algorithms(based on string of occurrences) had higher performances. The comparison showed that increasing number of features (either bits or colors) may had bad effects on the classifiers. Therefore, the designers must maintain the number of features and disbanding unnecessary features. The most effective technique that increase the accuracy of Image based classifiers is NFE, while the N-grams is most effective technique to increase the accuracy of other classifiers. It is important to reference that increased number of families had bad effects on proposed classifiers.

Acknowledgment:

Special thanks to University of Mosul and computer Science and mathematics college besides to computer scinceDept. for their Supporting us andprovide computer laboratories.

References:

[1] Shekhawat, N. S., & Mathew, R. (2021). A Review of Malware Classification Methods using Machine Learning. Proceedings of the 54th Hawaii International Conference on System Sciences2021, Available at SSRN 3769906.

[2] Kale, A. S., Di Troia, F., & Stamp, M. (2021). Malware Classification with Word Embedding Features. arXiv preprint arXiv:2103.02711.

[3]Balram, N., Hsieh, G., & McFall, C. (2019, December). Static Malware Analysis Using Machine Learning Algorithms on APT1 Dataset with String and PE Header Features. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 90-95). IEEE.

[4] Devine, S., & Bastian, N. (2021, January). An Adversarial Training Based Machine Learning Approach to Malware Classification under Adversarial Conditions. In Proceedings of the 54th Hawaii International Conference on System Sciences (p. 827).

[5] Aslan, Ö., Ozkan-Okay, M., & Gupta, D. (2021). A Review of Cloud-Based Malware Detection System: Opportunities, Advances and Challenges. European Journal of Engineering and Technology Research, 6(3), 1-8.
[6] Nathan H. (2017). TheCompleteCyber SecurityCourse, StationX Ltd, Vol1, London.
[7] Saravanakumar, R., & Kathiresan, V. (2018) Review Paper on Risks of PC Virus and its Preventions, International Journal of Computer Science Trends and Technology (IJCST) – Volume 6 Issue 5, Sep-Oct 2018.
[8] Imperva (2021). Mitigating backdoor shell attacks with Imperva. Link: https://www.imperva.com/learn/application-security/backdoor-shell-attack/#:~:text=A%20backdoor%20is%20a%20malware,system%20commands%20and%20update%20malware.
[9] Miura, H., Mimura, M., & Tanaka, H. (2018, November). Discovering new malware families using a linguistic-based macros detection method. In 2018 Sixth International Symposium on Computing and Networking Workshops (CANDARW) (pp. 431-437). IEEE.
[10] Gasparinatos ‹S. (2018). Malware development with the use of known techniques (Master's thesis, University of Piraeus).
[11] Richardson, R., & North, M. M. (2017). Ransomware: Evolution, mitigation and prevention. International Management Review, 13(1), 10.
[12] Bruce R. (2017). Statistical and Machine-Learning Data Mining, CRC Press, 3rd Ed. New York.
[13] Luo, J.-S., & Lo, D. C.-T. (2017). Binary malware image classification using machine learning with local binary pattern. 2017 IEEE International Conference on Big Data (Big Data).
[14] Zhang, C., Zhang, Y., Shi, X., Almpanidis, G., Fan, G., & Shen, X. (2019). On incremental learning for gradient boosting decision trees. Neural Processing Letters, 50(1), 957-987.
[15] Makandar, A., & Patrot, A. (2015). Malware image analysis and classification using support vector machine. International Journal of Trends in Computer Science and Engineering, 4(5), 01-03.
[16] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. Expert Systems with Applications.
[17] Bakour, K., & Ünver, H. M. (2020). VisDroid: Android malware classification based on local and global image features, bag of visual words and machine learning techniques. Neural Computing and Applications.
[18] Ünver, H. M., & Bakour, K. (2020). Android malware detection based on image-based features and machine learning techniques. SN Applied Sciences, 2(7), 1-15.
[19] Masabo, E., Kaawaase, K. S., Sansa-Otim, J., Ngubiri, J., & Hanyurwimfura, D. (2020). Improvement of Malware Classification Using Hybrid Feature Engineering. SN Computer Science, 1(1), 1-14.
[20] Masabo, E. (2019). A Feature Engineering Approach for Classification and Detection of Polymorphic Malware using Machine Learning (Doctoral dissertation, Makerere University).
[21] Suarez-Tangil, G., Dash, S. K., Ahmadi, M., Kinder, J., Giacinto, G., & Cavallaro, L. (2017, March). Droidsieve: Fast and accurate classification of obfuscated android malware. In Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy (pp. 309-320).
[22] Yang, L., & Liu, J. (2020). TuningMalconv: malware detection with not just raw bytes. IEEE Access, 8, 140915-140922.
[23] Sethi, A. (2019). Classification of Malware Models.
[24] Vasan, D., Alazab, M., Wassan, S., Safaei, B., & Zheng, Q. (2020). Image-Based malware classification using ensemble of CNN architectures (IMCEC). Computers & Security, 92, 101748.
[25] Xue, D., Li, J., Lv, T., Wu, W., & Wang, J. (2019). Malware classification using probability scoring and machine learning. IEEE Access, 7, 91641-91656.
[26] Udayakumar, N., Saglani, V. J., Cupta, A. V., & Subbulakshmi, T. (2018). Malware Classification Using Machine Learning Algorithms. 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI).
[27] Morales-Molina, C. D., Santamaria-Guerrero, D., Sanchez-Perez, G., Perez-Meana, H., & Hernandez-Suarez, A. (2018). Methodology for Malware Classification using a Random Forest Classifier. 2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC).
[28] Hassen, M., Carvalho, M. M., & Chan, P. K. (2017). Malware classification using static analysis based features. 2017 IEEE Symposium Series on Computational Intelligence (SSCI).
[29] Sewak, M., Sahay, S. K., & Rathore, H. (2018, June). Comparison of deep learning and the classical machine learning algorithm for the malware detection. In 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD) (pp. 293-296). IEEE.