STUDY OF MALWARE DETECTION USING MACHINE LEARNING

Research · July 2021

DOI: 10.13140/RG.2.2.11478.16963

2 authors, including:

Raj Sinha
Jayoti Vidyapeeth Women's University
15 PUBLICATIONS **0** CITATIONS

SEE PROFILE

STUDY OF MALWARE DETECTION USING MACHINE LEARNING

Raj Sinha
Research Scholar, Department Of Science and Technology, Jayoti Vidyapeeth Women's University, Jaipur, Rajasthan, India, rajsinha@jvwu.ac.in

Dr. Shobha Lal
Professor of Mathematics and Computing, Department of Science and Technology, Jayoti Vidyapeeth Women's University, Jaipur, Rajasthan, India, dean.fet@jvwu.ac.in

Abstract: Malware are considered as the malicious software program which infects the normal working of the PCs or Mobile. The concept of machine learning is utilized in the detection of the malwares, various sub-approaches like the ANN, CNN etc. are used by various researchers for the detection purpose. In this paper, we explore the various different types of the models which are used by various researchers in the malware detection thereby highlighting the accuracy of these models.
Keywords: Malware Detection, Machine Learning, Deep Learning

1 Introduction

Malware, short for malicious software, is a sweeping term for viruses, worms, trojans and other harmful software programs which can either create harm to data or access some important data illegally. As Microsoft puts it, "[malware] is a trick all term to allude to any software intended to make harm a solitary PC, worker, or PC organization." as such, software is recognized as malware dependent on its expected use, instead of a specific procedure or innovation used to assemble it. [1]

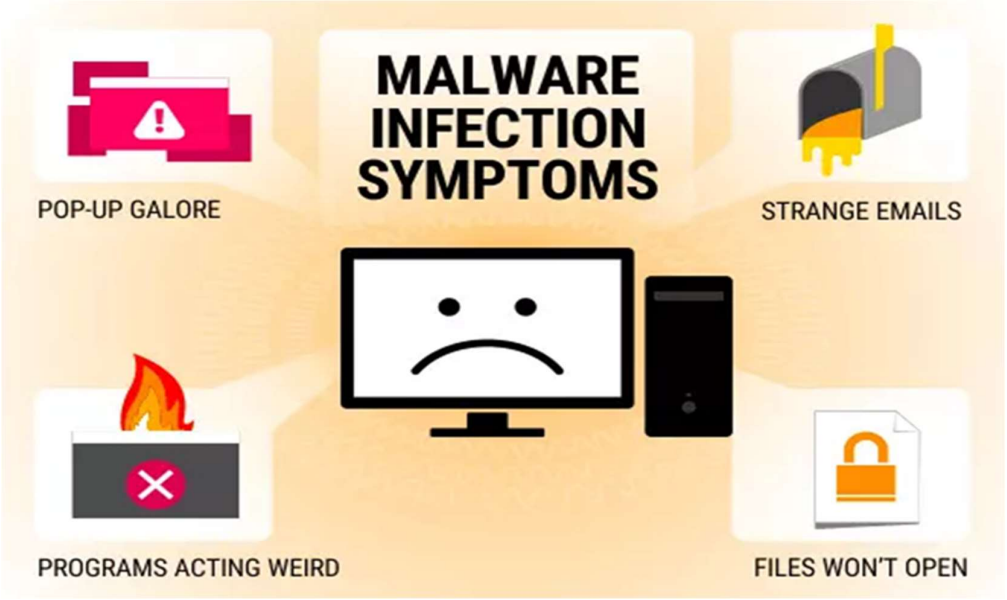

Fig 1 Malware Symptoms

There are various methods of arranging malware; the first is by how the malicious software spreads. You've most likely heard the words virus, trojan, and worm utilized conversely, however as Symantec clarifies, they portray three inconspicuously various ways malware can taint target PCs: [1]

- A worm is an independent piece of malicious software that duplicates itself and spreads from one PC to another. [1]

- A virus is a piece of PC code that embeds itself inside the code of another independent program, trigger them to make a malicious move and spread itself.

- A trojan is a program that can't repeat itself yet takes on the appearance of something the client needs and fools them into initiating it so it can do its harm and spread. [1]

Malware can likewise be introduced on a PC "physically" by the actual attackers, either by acquiring actual admittance to the PC or utilizing advantage heightening to acquire distant director access.

Another approach is to order malware is by what it does once it has effectively contaminated PCs data. There is a wide scope of potential attack procedures utilized by malware: [2]

- Spyware is characterized by Webroot Cybersecurity as "malware utilized with the end goal of covertly assembling information on a clueless client." basically, it keeps an eye on your conduct as you utilize your PC, and on the information, you send and get, ordinarily to send that data to an outsider. A keylogger is a particular sort of spyware that records every one of the keystrokes a client makes—incredible for taking passwords. [2]
- A rootkit is, as portrayed by TechTarget, "a program or, all the more frequently, an assortment of software devices that gives a danger entertainer distant admittance to and authority over a PC or other framework." It gets its name since it's a pack of devices that (for the most part unlawfully) acquire root access (head level control, in Unix terms) over the objective framework, and utilize that ability to conceal their quality. [3]

- Adware is malware that powers your program to divert to web commercials, which frequently try themselves to download further, more malicious software. As The New York Times notes, adware frequently piggybacks onto enticing "free" programs like games or program augmentations. [3]

- Ransomware is a kind of malware that scrambles your hard drive's records and requests an installment, ordinarily in Bitcoin, in return for the decryption key. A few prominent malware flare-ups of the most recent couple of years, like Petya, are ransomware. Without the decryption key, it's numerically incomprehensible for casualties to recapture admittance to their records. Alleged scareware is such a shadow form of ransomware; it professes to have assumed responsibility for your PC and requests a ransom, yet really is simply utilizing stunts like program divert circles to cause it to appear as though it's accomplished more harm than it truly has, and dissimilar to ransomware can be moderately handily debilitated. [3]

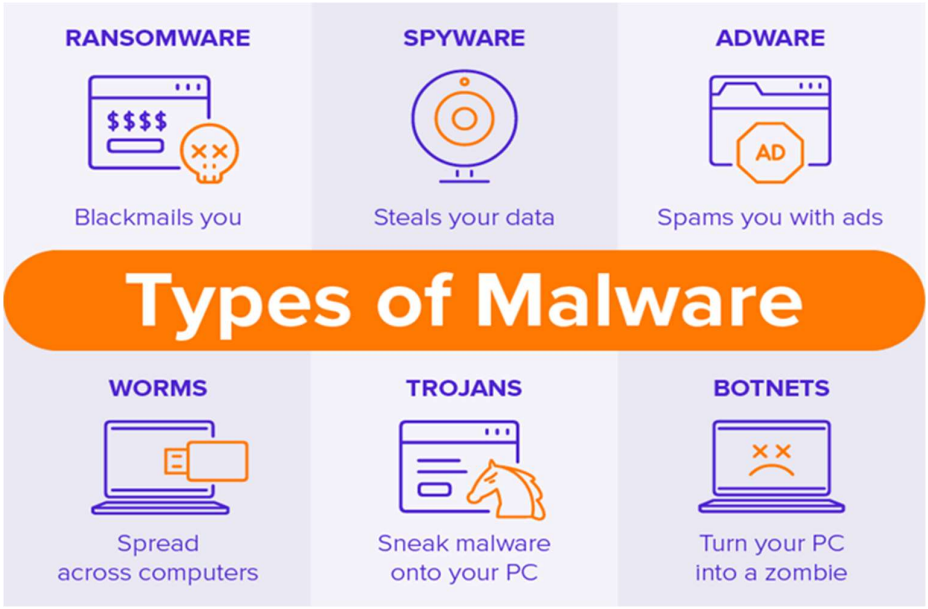

Fig 2 Malware Types

- Cryptojacking is another way attacker can constrain you to supply them with Bitcoin—just it works without you essentially knowing. The crypto mining malware contaminates your PC and utilizations your CPU cycles to dig Bitcoin for your attacker's benefit. The mining software may run behind the scenes on your working framework or even as JavaScript in a program window.

Malvertising is the utilization of genuine advertisements or promotion organizations to secretly convey malware to clueless clients' PCs. For instance, a cybercriminal may pay to put a promotion on an authentic site. At the point when a client taps on the promotion, code in the advertisement either diverts them to a malicious site or introduces malware on their PC. After that , the malware inserted in an advertisement may execute naturally with no activity from the client, a strategy alluded to as a "drive-by download." [4]

A particular piece of malware has both the methods for disease and a social classification. Along these lines, for example, WannaCry is a ransomware worm. What's more, a specific piece of malware may have various structures with various attack vectors: for example, the Emotet banking malware has been seen in the wild as both a trojan and a worm. [4]

2 Machine Learning

Machine learning is a strategy for information examination that computerizes analytical model structure. It is a part of artificial intelligence dependent on the possibility that frameworks can gain from information, distinguish examples and settle on choices with insignificant human intervention. Because of new registering innovations, machine learning today isn't care for machine learning of the past. It was brought into the world from design acknowledgment and the hypothesis that computers can learn without being programmed to perform explicit errands; scientists keen on artificial intelligence needed to check whether computers could gain from information. The iterative part of machine learning is

significant in light of the fact that as models are presented to new information, they can freely adjust. They gain from past calculations to deliver solid, repeatable choices and results. It's a science that is not new – but rather one that has acquired new momentum. [5]

Fig 3 Machine Learning

A computer is supposed to gain from Experiences regarding some class of Tasks, if its presentation in a given Task improves with the Experience.

A computer program is said to gain for a fact E regarding some class of undertakings T and execution measure P, if its exhibition at errands in T, as estimated by P, improves with experience E.

3 Malware Detection using Machine Learning

Machine learning is the broader fields and it is subdivided into the various aspects like supervised learning, unsupervised learning, the domain of the study which we select for the study of malware detection using machine learning are the techniques like Support Vector Machine, Naïve Bayes and Neural Networks.

3.1 Malware Detection using Support Vector Machine

In order to proceed for the malware detection using SVM, first we have a glimpse over the concept of the support vector machine uses. Support vector machine is profoundly liked by numerous individuals as it produces huge precision with less calculation power.

Support Vector Machine, truncated as SVM can be utilized for both regression and classification undertakings. "Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be utilized for either classification or regression challenges. Notwithstanding, it is generally utilized in classification issues. In the SVM algorithm, we plot every information thing as a point in n-dimensional space (where n is number of highlights you have) with the estimation of each element being the estimation of a specific organize. At that point, we perform classification by tracking down the hyper-plane that separates the two classes well indeed (take a gander at the beneath depiction). [5]

BaigaltugsSanjaa and ErdenebatChuluun , (2013) authors have made use of the "data mining" approach for the purpose of the malicious software detection and has also performed some of the experimental investigation on the malware detection using the concept of the linear SVM algorithm. The SVM has detected malware with the probability of 74 - 83%.

EvgenyBurnaev and Dmitry Smolyakov, (2016) presents the new approach to one-class classification based on SVM. Authors formulate the new problem statement and also the correspondingalgorithm, one that allow taking into account the privileged information during the process of the training phase. Authors have evaluated the performance of the proposed approach using the concept of synthetic datasets, as well as the publicly available Microsoft MalwareClassification Challenge dataset. On the classes category of Malware best accuracy attended is 97%.

Junmei Sun et. al (2017) Feature extraction concept using SVM know as extraction method of Android malware feature based on KCD, is proposed. The method makes use of the Keywords Correlation Distance in order to compute the correlation in between the key codes such as the API calls, the Android permissions, the common parameters, and the common key words in Android malware source code. The experiments also show that the method is much more efficient and also effective in the process of detecting malwares on Android platform. The accuracy achieved is around 87-88%.

Amr I. Elkhawasand NashwaAbdelbaki, (2018) authors considered text mining approach for the purpose of detection method in polymorphism and metamorphism. The instruction sequence for the critical code in the malware on the assembly-based level is basically the same across the malware families. Authors used opcode trigram sequences as the main feature for their machine learning algorithm. They have used Support Vector Machine (SVM) as classification algorithm. The 79% of balanced accuracy is achieved.

PinnuMadhukiranet. al (2020) highlights that the various applications access location in order to show the nearby information like restaurants, schools etc. and gain access to the private and sensitive data. Authors make use of the SVM to implement the permission access control over the applications.

Table 1. Malware Detection using SVM

Authors	**Paper Title**	**Model Used**	**Accuracy Achieved**
BaigaltugsSanjaa and ErdenebatChuluun (2013)	Malware detection using linear SVM	Linear SVMalgorithm	74-83%
EvgenyBurnaev and Dmitry Smolyakov, (2016)	One-Class SVM with Privileged Information and Its Application to Malware Detection	One-class classification based on SVM	Best accuracy in malware category is 97%
Junmei Sun et. al (2017)	Malware detection on android smartphones using keywords vector and SVM	Extraction method of Android malware feature based on KCD	The accuracy achieved is around 87-88%.

Amr I. Elkhawas and NashwaAbdelbaki, (2018)	Malware Detection using Opcode Trigram Sequence with SVM	Opcode trigram sequences with Support Vector Machine (SVM)	79% of balanced accuracy is achieved.
PinnuMadhukiran et. al (2020)	Malware Detection in Smartphone Using SVM	SVM to implement the permission access control over the applications.	Improved Security Level

3.2 Malware Detection using Naïve Bayes

It is a classification technique dependent on Bayes' Theorem with a suspicion of freedom among indicators. In straightforward terms, a Naive Bayes classifier expects that the presence of a specific element in a class is inconsequential to the presence of some other feature. A naive Bayes classifier accepts that the presence (or nonappearance) of a specific element of a class is disconnected to the presence (or nonattendance) of some other component, given the class variable. Fundamentally, it's "naive" in light of the fact that it makes suppositions that could conceivably end up being right.

LuizaSayfullina , et. al. (2015)In this paper, authors present malware classification based on the features extracted from Android application package (APK) files. Authors proposed Normalized Bernoulli Naive Bayes classifier which results in improved class separation and higher accuracy. Results achieved 0.1% false positive rate with overall accuracy of 91%.

Cangzhou Yuan, et. al (2016) In this paper, authors present classification approach for Android applications based on Bayesian classification. Based on the permission extraction by applications, they experiment with 13005 applications that are composed of 18 categories with Naive Bayes. They gained 94% accuracy.

Min Tan et.al (2017) for the classification of the android application author proposed the concept of combining feature correlation and Bayes classification model. Experiment results suggest that the improved classification method is more effective than the Naive Bayes classification model in detecting Android malware. The 90.665 accuracy is achieved.

RidhoAlifUtama, et.al (2018)this paper classifies the applications as dangerous on the basis of the permissions and vulnerabilities by making use of Naive Bayes (NB) algorithm. The accuracy which is obtained from this research is 97.2%.

Jiaqi Pang and JialiBian , (2019) Inthis paper, authors proposed the Android malware static detection method base on Naive Bayes. Authors requested the permissions, the system API calls, and the proportion of Activity among the four Android major components through Android packages. They used the three types of information as the features to characterize each of the applications, and then performs the classification model training and malware detection through Naive Bayes classifier. For datasets they tested 6120 Android malwares and 6032 benign applications. The accuracy level which is achieved is 87.18%.

Table 2. Malware Detection using Naïve Bayes

Authors	Paper Title	Model Used	Accuracy Achieved
LuizaSayfullina , et. al . (2015)	Efficient Detection of Zero-day Android Malware Using Normalized Bernoulli Naive Bayes	Normalized Bernoulli Naive Bayes classifier	91%
Cangzhou Yuan, et. al (2016)	Android Applications Categorization Using Bayesian Classification	classification approach for Android applications based on Bayesian classification	94%
Min Tan et.al (2017)	Android malware detection combining feature correlation and Bayes classification model	feature correlation and Bayes classification	90.665%
RidhoAlifUtama, et.al (2018)	Analysis and Classification of Danger Level in Android Applications Using Naive Bayes Algorithm	Dangerous on the basis of the permissions and vulnerabilities by making use of Naive Bayes (NB) algorithm.	97.2%
Jiaqi Pang and JialiBian , (2019)	Android Malware Detection Based on Naive Bayes	the Android malware static detection method base on Naive Bayes	87.18%

3.3 Malware Detection using Neural Networks

Neural Networks are essentially a piece of Deep Learning, which thus is a subset of Machine Learning. So, Neural Networks are only an exceptionally progressed use of Machine Learning that is currently discovering applications in numerous fields of interest. Neural networks are a class of machine learning algorithms which is used to demonstrate complex patterns in datasets using different secret layers and non-straight actuation functions. Every neuron's coefficients (weights) are then adjusted comparative with the amount they added to the complete error. Scholarly articles for neural networks machine learning. A neural network is a series of algorithms that endeavors to perceive basic relationships in a set of data through a process that mimics the manner in which the human mind operates. In this sense, neural networks allude to systems of neurons, either natural or artificial in nature.

Y. Jin et. al (2018) Author proposed integrated the single machine learning classifier in CNN with the Adaptive Selection of Classifiers (ASC) in order to improve the performance of the malware

classification. Tests are performed over the 1746 apk samples with 1000 malware, the results indicated that the 4.27% better accuracy achieved as compared to the simple CNN.

C. Hasegawa and H. Iyatomi, (2018) authors' proposed light-weighted approach of Android malware detection method. According to their proposed method , the concept treats the very limited part of raw APK (Android application package) file of the target as the short string and analyzes it with the one-dimensional convolutional neural network (1-D CNN). They test on the two different datasets each consisting of 5,000 malwares and 2,000 goodwares. The results shows the accuracy of 95.40-97.04%.

D. Li, et. al (2018) used the deep-learning-based method in order to detect the Android malware and implement an automatic detection engine in order to detect the families of malicious applications. The results of the evaluation show that the engine can detect 97% of the malware at 0.1% false positive rate (FPR) when detecting the fine-grained malware families.

M. Masum and H. Shahriar (2019) ,In this paper, authors proposed the deep learning framework, called Droid-NNet, for the purpose of the malware classification. The proposed method Droid-NNet is the deep learner one that outperforms existing cutting-edge machine learning methods. Authors performed experiments on two datasets (Malgenome-215 & Drebin-215) of Android apps to evaluate Droid-NNet. The experimental result shows the robustness and effectiveness of Droid-NNet. Achieved 98% accuracy.

Table 3. Malware Detection using Neural Networks

Authors	Paper Title	Model Used	Accuracy Achieved
Y. Jin et. al (2018)	Android Malware Detector Exploiting Convolutional Neural Network and Adaptive Classifier Selection	Integrated the single machine learning classifier in CNN with the Adaptive Selection of Classifiers (ASC)	4.27% better accuracy than CNN
C. Hasegawa and H. Iyatomi, (2018)	One-dimensional convolutional neural networks for Android malware detection	light-weighted approach of Android malware detection	95.40-97.04%.
D. Li, et. al (2018)	Fine-grained Android Malware Detection based on Deep Learning	the deep-learning-based method in order to detect the Android malware and implement an automatic detection engine	97%
M. Masum and H. Shahriar (2019)	Droid-NNet: Deep Learning Neural Network for	deep learning framework, called Droid-NNet	98 %

	Android Malware Detection		

6 Conclusion

In this paper, we explore the various different types of the modals which are used by various researchers in the malware detection and highlight the accuracy of these models. And as per the accuracy of the results we found that the Malware detection based on the neural networks are more effective and accurate as compared to the other approaches.

References(APA)

1. Mohsen Kakavand Mohammad Dabbagh and Ali. Dehghantanha Application of Machine Learning Algorithms for Android Malware Detection pp. 32-36 2018.
2. M. Kalash M. Rochan N. Mohammed N. D. Bruce Y. Wang and F. Iqbal "Malware classification with deep convolutional neural networks" 2018 9th IFIP International Conference on New Technologies Mobility and Security (NTMS) pp. 1-5 2018 February.
3. A. Mujumdar G. Masiwal and D. B. Meshram "Analysis of signature-based and behavior-based anti-malware approaches" International Journal of Advanced Research in Computer Engineering and Technology (IJARCET) vol. 2 no. 6 2013.
4. I. Burguera U. Zurutuza and S. Nadjm-Tehrani "Crowdroid: behavior-based malware detection system for Android" Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices pp. 15-26 2011 October.
5. D. Gavrilut M. Cimpoesu D. Anton and L. Ciortuz Malware Detection Using Machine Learning Proceedings of the International Multiconference on Computer Science and Information Technology pp. 735-741 2009.
6. B. Sanjaa and E. Chuluun, "Malware detection using linear SVM," Ifost, Ulaanbaatar, Mongolia, 2013, pp. 136-138.
7. P. M. Kiran, P. N. Reddy and L. SujiHelen, "Malware Detection in Smartphone Using SVM," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India, 2020, pp. 344-347.
8. A. I. Elkhawas and N. Abdelbaki, "Malware Detection using Opcode Trigram Sequence with SVM," 2018 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split, Croatia, 2018, pp. 1-6.
9. E. Burnaev and D. Smolyakov, "One-Class SVM with Privileged Information and Its Application to Malware Detection," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 2016, pp. 273-280.
10. J. Sun, K. Yan, X. Liu, C. Yang and Y. Fu, "Malware detection on android smartphones using keywords vector and SVM," 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, China, 2017, pp. 833-838.
11. J. Pang and J. Bian, "Android Malware Detection Based on Naive Bayes," 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2019, pp. 483-486.
12. M. Tan, M. Yu, Y. Wang, S. Li and C. Liu, "Android malware detection combining feature correlation and Bayes classification model," 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN), Guangzhou, 2017, pp. 664-668.
13. R. A. Utama, P. Sukarno and E. M. Jadied, "Analysis and Classification of Danger Level in

Android Applications Using Naive Bayes Algorithm," 2018 6th International Conference on Information and Communication Technology (ICoICT), Bandung, 2018, pp. 281-285

14. L. Sayfullinaet al., "Efficient Detection of Zero-day Android Malware Using Normalized Bernoulli Naive Bayes," 2015 IEEE Trustcom/BigDataSE/ISPA, Helsinki, Finland, 2015, pp. 198-205.
15. C. Yuan, S. Wei, Y. Wang, Y. You and S. G. ZiLiang, "Android Applications Categorization Using Bayesian Classification," 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Chengdu, China, 2016.
16. Y. Jin, T. Liu, A. He, Y. Qu and J. Chi, "Android Malware Detector Exploiting Convolutional Neural Network and Adaptive Classifier Selection," 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, Japan, 2018, pp. 833-834.
17. C. Hasegawa and H. Iyatomi, "One-dimensional convolutional neural networks for Android malware detection," 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA), Penang, Malaysia, 2018, pp. 99-102.
18. D. Li, Z. Wang and Y. Xue, "Fine-grained Android Malware Detection based on Deep Learning," 2018 IEEE Conference on Communications and Network Security (CNS), Beijing, China, 2018, pp. 1-2.
19. M. Masum and H. Shahriar, "Droid-NNet: Deep Learning Neural Network for Android Malware Detection," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 5789-5793.