IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TO DETECT BEHAVIOUR BASED MALWARE USING INDEPENDENT COMPONENT ANALYSIS

A.Manasa[1]
K.Sireesha[2]
PG Scholar
Madanapalle Institute of Technology & Science, India

Dr.R.Maruthamuthu[3]
Assistant Professor
Department of Computer Applications

Abstract:

Malware has been a concern to enterprises for a long time, and organisations have made little progress in identifying malware in a timely manner [1]. Malware can easily affect the system by executing unwanted services that add to the system's burden and obstruct its smooth operation. There are basically two approaches for identifying malware [2] the traditional method of detecting malware based on signatures and the new method of detecting malware based on behaviour. The operation that the malware conducts when it is active in the machine determines the malware's behaviour [3], such as launching the Operating System services or downloading infected files from the internet. Based on the malware's activity, the suggested method detects it. The proposed model in this paper is a hybrid of Support Vector Machines and Principle Component Analysis. During validation, our suggested model achieved a 97.75 percent accuracy with 97 percent precision, 99 percent recall, and a f1-score of.98 for genuine Malwares.

Keywords : Malware, Malware Detection, Behaviour-based, Principle Component Analysis, Support Vector Machine, Machine Learning

1. INTRODUCTION:

Computers that use the internet to download large amounts of data from the internet may also download viruses [2]. Malware is referred to by a variety of terms, including malicious code, harmful programmes, and malicious executable files. As virus attacks have become increasingly common, [3] computer systems have become more vulnerable to hacking. Malware is "a form of computer programme designed to infect a legitimate user's computer and wreak harm on it in different ways," according to Kaspersky Labs in 2017. [4] With the ever-increasing diversity of malware, anti-virus scanners cannot guarantee the identification of every type of malware based on its signature, resulting in millions of hosts being targeted and causing significant harm to data and other systems. According to Kaspersky Lab (2016), roughly 4,000,000 new forms of malware were detected on 6,563,145 distinct devices. Since a result, securing the network and user machines against malware is a critical cyber security duty for individual users and businesses alike,[5] as even a single attack can result in considerable loss and harm. [6]The goal of this work is to create a malware detection system that can detect malware based on the actions it performs on the machine it's installed on.

2. LITERATURE REVIEW:

[1]M. A. Jerlin and K. Marimuthu, "A New Malware Detection System Using Machine Learning Techniques for API Call Sequences," *J. Appl. Secur. Res.*, vol. 13, no. 1, pp. 45–62, 2018.

Conventional malicious webpage detection methods use blacklists in order to decide whether a webpage is malicious or not.[10] the blacklists are generally maintained by third-party organizations.

The detection and classification of malwares in windows executable is an important and demanding task in the field of data mining [11] The malwares can easily damage the system by creating harm in the user's system, [9] so some of the existing techniques are developed in the traditional works for an accurate malware detection

[2]B. Sanz, I. Santos, C. Laorden, X. Ugarte-Pedrero, P.G. Bringas, and G.Álvarez, "PUMA: Permission usage to detect malware in android," *Adv. Intell. Syst. Comput.*, vol. 189 AISC, pp. 289–298, 2013.

The presence of mobile devices has increased in our lives offering almost the same functionality as a personal computer. [12]Android devices have appeared lately and[14], since then, the number of applications available for this operating system has increased exponentially Google already has its Android Market where applications are offered and,[13] as happens with every popular media, is prone to misuse.

[3]Y. Fan, Y. ye, and L. Chen, "Malicious sequential pattern mining for automatic malware detection," *Expert Syst. Appl.*, vol. 52, pp. 16–25, 2016.

Due to its damage to Internet security, malware [15] (e.g., virus, worm, Trojan) and its detection has caught the attention of both anti-malware industry and researchers for decades.

To protect legitimate users from the attacks, the most significant line of defines against malware is anti-malware software products, [16] which mainly use signature-based method for detection.

[17]A comprehensive experimental study on a real data collection is performed to evaluate our detection framework.

[4]U. Baldangombo, N. Jambaljav, and S.-J. Horng, "A Static Malware Detection System Using Data Mining Methods," 2013

A serious threat today is malicious executables. [6] It is designed to damage computer system and some of them spread over network without the knowledge of the owner using the system. Two approaches have been derived for it [7] i.e. Signature Based Detection and Heuristic Based Detection. These approaches performed well against known malicious programs but cannot catch the new malicious programs

[5]Y. Saint Yen and H. M. Sun, "An Android mutation malware detection based on deep learning using visualization of importance from codes," *Microelectron. Reliab.*, vol. 93, no. October 2018, pp. 109–114, 2019.

Using smartphone especially android platform has already got eighty percent market 10shares, [8] due to aforementioned report, it becomes attacker's primary goal. There is a growing 11number of private data onto smart phones and low safety defense measure, attackers can use 12multiple way to launch and to attack user's smartphones.[9] Nowadays most malware detection methods use only one 18aforementioned feature, and these methods mostly analysis to detect code, but facing the influence 19of malware's code confusion and zero-day attack.

3. PROPOSED METHODLOGY:

In Proposed system we propose an SVM model in Machine Learning this can be improve the accuracy results.[10] Different Malware that were previously detected by various sources were collected and subjected to feature extraction. This section will describe the detailed description of the proposed work done for the detection of malware. [11] Dataset: - Different malware and benign that were previously detected by various sources were collected and subjected to feature extraction. On completion of the feature extraction a total of 77 features were generated which will be used for training the model. The management of people is simple and effective.[2] It is also more cost effective because no specific machines are required, and resources are used much less.[6] It will provide accurate results.

4. IMPLEMENTATION:

This section will go over the research technique process. Figure [1] depicts a general overview of the research approach. A. Data Gathering and Storage A virus data set and a benign instance data set make up the data set.[12] The format of both malware and benign instance data sets is Windows Portable Executable (PE) file binaries. A total of 220 distinct malware samples (including 1 Indonesian malware) were obtained. [13] The benign instance data set samples were taken from system files in the "System32" directory of a fresh Windows XP Professional 32-bit with Service Pack 2 installation. [10] A total of 250 samples of benign software were obtained. The dataset is initially preprocessed for any missing values (as seen in figure 3) by either deleting the missing values or replacing them with the mean.values.After the missing values have been removed, the datatype of With the help of, the complete dataset is synchronized. The pandas library has an as Type ("float64") function. python. Once the dataset was cleaned it was subjected to normalization and scaling of the data. On completion of the preprocessing of the data, it is then subjected to the PCA Analysis for Dimensionality Reduction. The results obtained using the PCA analysis provided with the insight that only 40 features were enough for the training,validation and testing of the model.

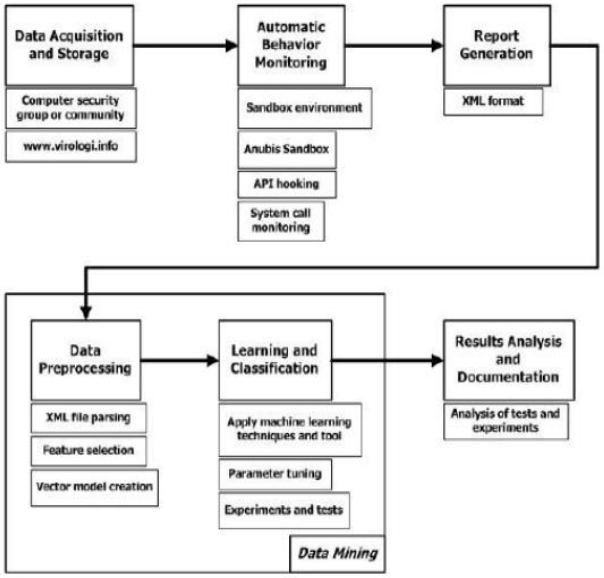

Figure 1. General overview of the research methodology

A. Behaviour Monitoring and Reporting on an Automated Basis: The malware and benign instance data sets will next be subjected to dynamic analysis (behaviour monitoring). [16] This is accomplished by uploading each sample to Anubis [6, a free online automatic dynamic analysis service]. A report file is generated as a result of binary submission and execution of Anubis. [17] All of the generated report files were downloaded in XML format for this study.

B. Data Preparation:

The following step is to perform data preparation. The following are the steps in this study's data pre-processing[17] The most relevant and important attribute values were selected from all

XML report files (feature selection). 2. A term dictionary was generated, containing all of the attribute values that had been processed and selected previously. 3. Each XML [18] report file was compared to the term dictionary by counting the existence (or non-existence) of each term word based on binary weight and term frequency weight in the term dictionary [17]. 4. Sparse vector models and Attribute-Relation File Format (ARFF) files were constructed for each XML report file.

C. Learning and Classification:

The next stage is to use the ARFF files to conduct learning and categorization. [19] For the learning and classification of the ARFF files, machine learning techniques were used. Algorithms utilised include For dimensionality reduction, Principle Component Analysis is performed. PCA aids in the removal of elements that aren't needed.have a lot of weight in categorisation PCA's main objective is to determine the main elements of all provided data features.As a result, by lowering the amount of features,Machines can learn more quickly. [19] SVM stands for Support Vector Machine and is commonly used for Using binary classification based on the hyperplane provides separation. The hyperplane is chosen.so that the distance between them is maximised separation and support vectors, which are the nearest vectors.

The data points that are closest to the support vectors are called support vectors.the plane of hyperplane SVM can be improved by using kernel functions.employed when dealing with non-linear data.

5. RESULTS:

Using SVM and after applying PCA and feeding the findings to train SVM, the model produced the following results. Only employing SVM improves prediction accuracy. The accuracy acquired by the first method was found to be 90.4 percent, whereas the accuracy gained by the second method was found to be 90.4 percent. PCA is used first, followed by SVM with a linear kernel function. The dataset has a 97.75 percent accuracy.

The following is the classification report:

Table 1 : Confusion Matrix

Actual/ Predicted	Benign	Malware
Benign	1415	114
Malware	42	4313

Table 2 : Classification Report

Classification Report	Precision	Recall	f-1 Score	Support
Benign	0.97	0.93	0.95	1592
Malware	0.97	0.99	0.98	4555
Micro Average	0.97	0.97	0.97	5884
Macro Average	0.97	0.96	0.96	5884
Weight Average	0.97	0.97	0.97	5884

6. CONCLUSION:

The suggested approach will be able to identify and differentiate malware from innocuous files and packets with greater ease and in less time. Because the proposed model is so complex, as a result of putting the Principle Component Analysis into practise. Additionally, the computational complexity is reduced. The suggested. The model is also quite accurate and high-performing. It did so because it used SVM with a Gaussian kernel at its heart.

7. REFERENCE:

[1] J. Landage and M. Wankhade, “Malware and Malware Detection Techniques: A Survey,” Int. J. Eng. Res. Technol., vol. 2, no. 12, pp. 61–68, 2013.

[2] Kaspersky, “Machine learning for Cybersecurity.”

[3] P. Kaur and S. Sharma, “Literature Analysis on Malware Detection,” Int. J. Electron. Electr. Eng., vol. 7, no. 7, pp. 717–722, 2014.

[4] I. A. Saeed, A. Selamat, and A. M. A. Abuagoub, “2013-A Survey on Malware and Malware Detection Systems.pdf,” vol. 67, no. 16, pp. 25–31, 2013.

[5] U. Baldangombo, N. Jambaljav, and S.-J. Horng, “A Static Malware Detection System Using Data Mining Methods,” 2013.

[6] Y. Saint Yen and H. M. Sun, “An Android mutation malware detection based on deep learning using visualization of importance from codes,” Microelectron. Reliab., vol. 93, no. October 2018, pp. 109–114, 2019.

[7] S. Sohrabi, O. Udrea, and A. V. Riabov, “Hypothesis Exploration for Malware Detection Using Planning,” Twenty-Seventh AAAI Conf. Artif. Intell., pp. 883– 889, 2013.

[8] J. C. Rosales, “Rehumanización y metáfora religiosa end Luis Rosales,” Insula, vol. 767, pp. 32–34, 2010.

[9] D. Nieuwenhuizen, “A behavioural-based approach to ransomware detection,” Whitepaper. MWR Labs Whitepaper, 2017.

[10] H. S. Galal, Y. Bassyouni, and M. A. Atiea, “Behavior-based features model for malware detection,” J. Comput. Virol. Hacking Tech., no. April, 2018. [11] Y. Fan, Y. Ye, and L. Chen, “Malicious sequential pattern mining for automatic malware detection,” Expert Syst. Appl., vol. 52, pp. 16–25, 2016. [12] C. I. Fan, H. W. Hsiao, C. H. Chou, and Y. F. Tseng, “Malware detection systems based on API log data mining,” Proc. - Int. Comput. Soft. Appl. Conf., vol. 3, pp. 255–260, 2015.

[13] B. Sanz, I. Santos, C. Laorden, X. Ugarte-Pedrero, P. G. Bringas, and G. Álvarez, “PUMA: Permission usage to detect malware in android,” Adv. Intell. Syst. Comput., vol. 189 AISC, pp. 289–298, 2013.

[14] M. A. Jerlin and K. Marimuthu, “A New Malware Detection System Using Machine Learning Techniques for API Call Sequences,” J. Appl. Secur. Res., vol. 13, no. 1, pp. 45–62, 2018.

[15] “General Python FAQ — Python 3.7.3 documentation.” [Online]. Available: https://docs.python.org/3/faq/general.html. [Accessed: 05-Apr-2019].

[16] “Package overview — pandas 0.24.2 documentation.” [Online]. Available: http://pandas.pydata.org/pandasdocs/stable/getting_started/overview.html. [Accessed: 05-Apr-2019].

[17] “An introduction to machine learning with scikitlearn — scikit-learn 0.20.3 documentation.” [Online]. Available: https://scikitlearn.org/stable/tutorial/basic/tutorial.html. [Accessed: 05-Apr-2019].

[18] B. Suite, B. Suite, and B. Suite, “Начнем сначала Getting Started.” [Online]. Available: http://www.pydev.org/manual_101_root.html. [Accessed: 05-Apr-2019].

[19] D. J. Bartholomew, “Principal components analysis,” Int. Encycl. Educ., vol. 2, no. June 2001, pp. 374–377, 2010. [20] D. Srivastava, “Data Classification Using Support Vector,” Journal of Theoretical and Applied Information Technology 12(1):1-7, February 2010.