A Survey on Machine Learning Approach to Detect Malware

Selvarathi C [a], Jeevitha J [b], Kevin Akash M [c], Rajkumar R [d], Vaisaali K [e]

[a] Department of Computer Science and Engineering ,M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India -639113
[b,c,d,e] Department of Computer Science and Engineering , M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India -639113

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Malware is one of the predominant challenges for the Internet users. In recent times, the injection of malwares into machines by anonymous hackers have been increased. This drives us to an urgent need of a system that detects a malware. Our idea is to build a system that learns with the previously collected data related to malwares and detects a malware in the give file, if it is present. We propose a various machine learning algorithm to detect a malware and indicates the user about the danger. In particular we propose to use a algorithm which give a optimal solution to hardware and software oriented malwares.

Keywords: Malware file, feature dataset, extraction, training, testing, classification.

1. Introduction

Malware means “Malicious Software” which can be inject into the computer and get access of the computer programs and files. The malware can do a harmful tasks in the personal computer[6]. The data in the computer has been destroyed by the third force. Some malwares are spyware, viruses, Phishing, fileless malware worms. Malware has been get into a network through the third party and they can do a malicious activity which will get access through social engineering. The cyber attackers can develop a variety of malicious software which can be used to gain a data from the business people. It will cause a huge damage to the user network[11]. Phishing is the one of the malicious activity which can done by a hacker to get a unauthorized data from the user. It will be implemented through the email or social media[5].In this paper, we detect a malware by using various machine learning algorithm[Figure 1]. It contains the previously collected data which was trained to detect the malware in the file[8]. The software tries to learn the data which are related to malicious and which are benign based on databases of both malicious and benign code. The scope of the project is that there is a need to produce a system that efficiently detects a malwares present in the system and indicates the user about the malware danger.

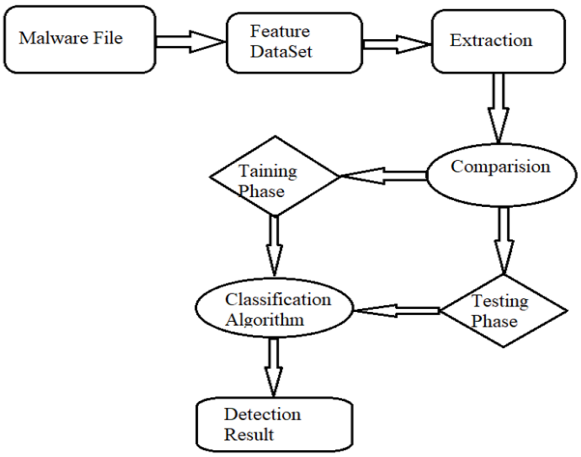

Figure 1: Malware Architecture

There is a tremendous increase of internet access by the humans which leads to the various data corruption[14]. Before that, viruses make attack on a personal computer which can make a copy of itself and insert into the other program or files and it leads to a harmful actions like destroying the data. For example, in Phishing the data has been stealed by the hackers by using the social engineering or mock-up websites. There are many methods have been proposed to detect a Phishing websites[20]. But hackers can evolved their methods to escape from these

detection methods. The most successful methods to detect a malware is machine learning[21]. This is because Phishing has some characteristics which are learned by the machine with the help of previously collected data in the machine learning algorithm. We have compared the results of various machine learning methods to detect a malware. And finally we pick a method which gives a better solution to the system's malware by using its success rate[21]. Our idea is to get a optimal result for the malware detection which are related to both software and hardware .The remaining document is organized as follows: Discuss through several methods of classifying malware and detection of a malware by using machine learning technique[28]. We are trying to bring out a better result in a traditional way, by comparing the results with two or more algorithms. Proposed method of malware detection is algorithm.

2. Machine Learning

The Machine Learning is one among the inspirational technology evolved, which in need to connect the globe inorder to take over the secular tasks in a automated manner. Machine Learning is categorized with the algorithms which allows the application software to explicit with the future predictions without being programmed. The basic operation of machine learning is building algorithms inorder to receive the inputs and output with the help of statistical analysis. Classifications of machine learning can be classified to three learnings. They are1.Supervised 2.Unsupervised 3.Reinforcement.In Supervised learning, the machine holds with the labeled data in which each data tagged with correct label. This is classified into Classification and Regression techniques. In Unsupervised learning, the machine holds with the uncategorized data and in prior the machine will not be trained[25]. This is classified into Clustering and Association. In Reinforcement learning, the machine will not be holds up with any data instead it will interact with the environment which it receives rewards for the correct performance and punishment for the incorrect performance.

2.1Algorithms:

2.1.1 Supervised Learning

These algorithms usually work with labeled data to learn mapping function that turns input into output variable. This helps us to generate the accurate outputs with the help of given new inputs. When output variable is divided into categories then, classification is used to render outcome of a given sample. A classification system also might look to input data for label assignment. When the output variable is related to the real values then, regression is helpful to render the outcome of a provided sample. The examples of supervised learning[3] are Naïve-Bayes ,Linear Regression, CART, Logistic Regression, and K-Nearest Neighbors (KNN).

Another form of supervised learning is combining two models' prediction as the prediction of individual system is not accurate enough. For example, Combining with Random Forests, Boosting with XG Boost are examples of ensemble techniques[3].

2.1.2 Unsupervised Learning Algorithms

These models only work with the input data and not the output data for any given sample. Unlabeled training data is used to model the structure. The objects are similar to one another within the same cluster than to the objects from another cluster. This denotes clustering. Dimensionality reduction is helpful to reduce the number of variables of a data set while confirming that important information is still conveyed [10]. The Feature Extraction methods and Feature Selection methods are used in dimensionality reduction. Feature extraction is nothing but the data transformation from high-dimensional space to low-dimensional space.

2.1.3 Reinforcement learning

The algorithm that allows to decide the futuristic behavior based on its present state which leads to maximize a reward denotes reinforcement learning[33].Through trial and error, the reinforcement learning learns the optimal actions.

3. Machine Learning Algorithms

3.1 Linear Regression

This is a supervised classification algorithm which is helpful to render the probability of a target variable based on the predictions. In logistic regressions the nature of target variable or dependent variable will results only to the two possible classes[35].The dependent variable is binary in nature while processing data and the result of that processed data will be coded as either 1 or 0[35].

3.3 CART

CART - Classification and Regression Trees. These are one implementation of Decision Trees[35].

Classification And Regression Tree(CART), a predictive model, which renders predicted based values and showcases how an outcome variable's values can be predicted. The output of a CART is based on a decision tree where each frame is a split in a predictor based variable and each of the end nodes hold for a prediction for the outcome variable[35].

3.4 Naive Bayes

NaivesBayes, a supervised learning algorithm, based on the Bayes theorem and also used for the major classification problems. This algorithm is one of the most effective and simple Classification algorithms which leads to the quick predictions with the aid of the fast learning models[36].

3.5 KNN

This algorithm is a supervised algorithm and also the simplest algorithm that is helpful in solving both classification and regression problems. It is easier to understand and implement. Usually, the KNN algorithm is used for the recommendation systems. But, this algorithm is cannot be used for the high dimensional data but a efficient algorithm for the baseline systems. It is also known as the instance based learning[36].

3.6 K-means

This clustering algorithm, is one of the simplest and also a popular unsupervised algorithm. In other words, the K-means algorithm denotes k number of centroids, and then assigns every data point to the nearest cluster, while keeping the centroids as tiny as possible in the result. This K-means algorithm is mostly used for the classification process[36].

3.7 Bagging with Random Forests

Random forest algorithm is a supervised machine learning algorithm, which is helpful for both classification and regression problems. Similarly, with the help of the data samples, random forest algorithm creates the decision trees. With the samples from those decision trees the best solution will be predicted[36].

Random forest algorithm is more flexible and also easy to use algorithm that produces result without hyper-parameter tuning, which produces the greatest of all time. It is also one of the most used algorithms, because of its simplicity[36]. Random Forest algorithm can be considered to be one among bagging techniques and not boosting techniques. The random forests trees usually run in parallel. The trees in boosting algorithms will be trained sequentially[36].

4. Case Study

In this project our aim is to build a system that learns with previously collected data related to Malwares and detect a malware in the given file if it is present. We are giving the files and their details as an input. By using the files and their details as an input we are detecting the Malware in the files. We are using various types of algorithms and find the best algorithm. By using this we will detect the Malware in the files.

5. Use Case

The malware detection can be very useful for Business field, IT sectors, Educational field, Healthcare field and Government sectors. Because these fields has very important and confidential data and information. To secure the important data we can use this malware detection.

Some of the malware attacks are:

1) LockerGoga is a malware attack hit in 2019 for the large corporations in the worlds such as Altran Technologies and Hydro. It caused millions of dollars loss for the companies[23].

2) One of the worst attack in history is WannaCryin 2017 through phishing emails. Many of the sectors has been affected by this attack. It nearly causes 4 billion USD of loss[30].

3) CryptoLocker is one of the most worst attack in the year of 2013. This attack has been done through email. It has been said that it has caused 3 million USD loss infecting nearly 200,000 people all over the world[48].

4) NetWalker is one of the latest attack which targeted governmental agencies, healthcare organizations, corporations and remote employees in the year 2020-2021.

5) Tycoon is a recently discovered malware type. Many organizations in the education and software industry has suffered by this malware attack[31].

6) In 2016 Linkedin was attacked and 6.5 million passwords were stolen by the attackers[39].

7) In 2013 Adobe declared that 3 million customer credit card details were stolen by the hackers[41].

8) CovidLock is a malware which encrypts key data on an android device and deny the access for the user[44].

These are some of the popular attacks of malware which had caused lot of loss in many sectors. To prevent these attacks we want to detect these malware earlier by using this malware detection. By using these method we can prevent many data.

6. Advantages of using Malware Detection:

- Safeguard from viruses and its transmission

The main role of this is to stand against viruses and other form of malwares. The viruses will not only damage the data it will also decrease the performance of the system. It will detect malware before it happens[30].

- Defence against Data thieves and Hackers

Malware detection will give protection against hackers and data thieves. It will detect them before they access or hack the data.

- Spyware protection

Spyware is a type of malware that spies on our system Stealing the confidential information. The malware detection has the capability to prevent these type of spyware attacks[19].

- Secure your Data and Files

The reason of this malware detection is to keep our data and files in a secure manner. By using this we can protect our data.

- Control the access of websites to build up the Web protection

While browsing in the internet users can come across different forms of threat. This can be overcome by using this malware detection. User can protect their information using this.

7. Front end:

7.1 React

React is a JavaScript library which is open source and is used for front end development. It was developed by Facebook. It also allows us to build user interfaces especially for single-page application. It also supports mobile

application development. In the modern days, React has become so popular because of its extra simplicity and flexibility. While other popular frameworks were also in the competition at the initial stages of React, programmers were forced to code in most of the occurrences irrespective of the change being minor or major. This prevailed as a problem until the development of React. As mentioned earlier, React is flexible and it can easily adapt to changes. It is not a wonder that many of the top corporations such as Facebook, Uber, PayPal, Airbnb and Instagram make use of React. This also amounted for the huge popularity. This credibility has drawn more people to the framework[Figure 2].

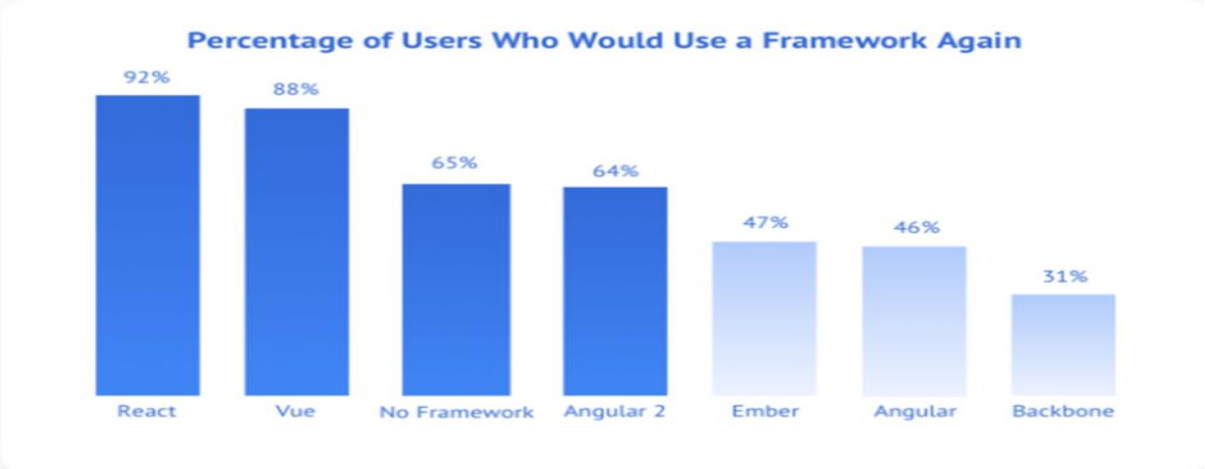

Figure2: Percentage of Users would use a framework again

As it is evident from the picture, use of React keeps increasing in comparison with other peer competitors.

8. Necessity in this project

The front end of the project is single-paged and no other better user interface designer is identified yet. Though all the actions happen in the back end, the aim is not fulfilled without a good and pleasing front end[Figure 3].

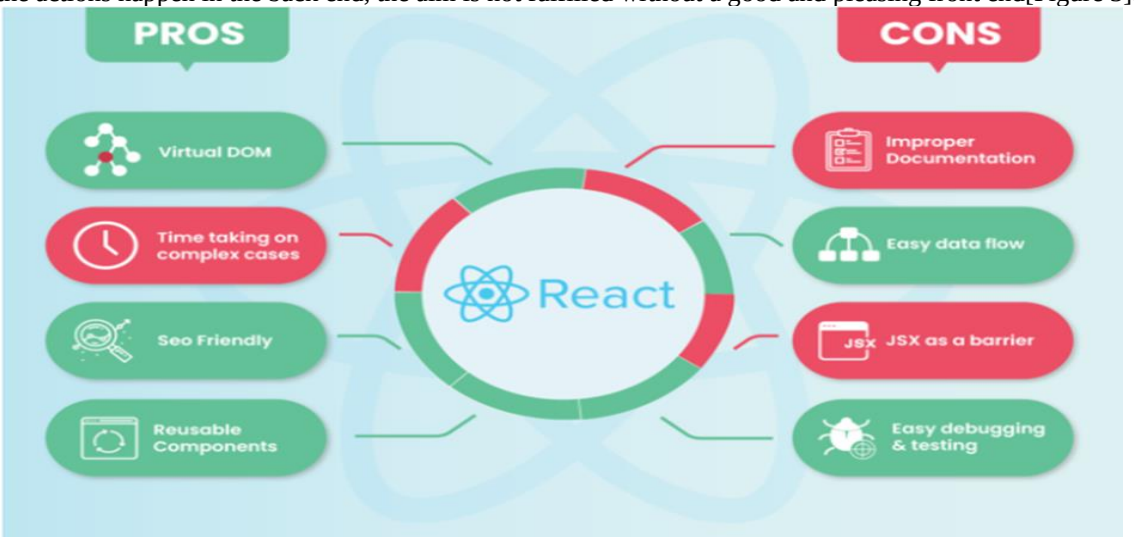

Figure 3: Pros and Cons of react

9. Conclusion

To summarize, although there are quite a few methods of detecting a malware, none of them are highly reliable. Thus, the method we follow will be of good help to not only the Internet giants, but also for the common people who unknowingly fall prey to the illegal malware community. Growth in technology also means growth in risk. The risk management and prevention should also be advanced and this approach takes us one step closer to what is needed.

References

A.Apvrille, D. Gordon, S. Hallyn, M. Pourzandi and V. Roy, "Digsig: Runtime authentication of binaries at kernel level", 18th USENIX Conference on System Administration, 2004.

F. Bellard, "Qemu a fast and portable dynamic translator", USENIX Annual Technical Conference FREENIX Track, pp. 41-46, 2005.

M. Charney, Xed2 user guide, 2011, [online] Available: http://software.intel.com/sites/landingpage/pintool/docs/56759/Xed/htmljmain.html.

L. Davi, M. Hanreich, D. Paul, A.-R. Sadeghi, P. Koeberl, D. Sullivan, et al., "Hafix: hardware-assisted flow integrity extension", Proceedings of the 52nd Annual Design Automation Conference, pp. 74, 2015.

D. R. Ellis, J. G. Aiken, K. S. Attwood and S. D. Tenaglia, "A behavioral approach to worm detection", Proceedings of the 2004 ACM Workshop on Rapid Malcode (2004) WORM '04.

Murugesan, M., Thilagamani, S. ," Efficient anomaly detection in surveillance videos based on multi layer perception recurrent neural network", Journal of Microprocessors and Microsystems, Volume 79, Issue November 2020, https://doi.org/10.1016/j.micpro.2020.103303

I. Santos, Y. K. Penya, J. Devesa and P. G. Garcia, "N-grams-based file signatures for malware detection", 2009.

K. Rieck, T. Holz, C. Willems, P. Düssel and P. Laskov, "Learning and classification of malware behavior", DIMVA '08: Proceedings of the 5th international conference on Detection of Intrusions and Malware and Vulnerability Assessment, pp. 108-125, 2008.

Thilagamani, S., Nandhakumar, C. ." Implementing green revolution for organic plant forming using KNN-classification technique", International Journal of Advanced Science and Technology, Volume 29 , Isuue 7S, pp. 1707–1712

M. Chandrasekaran, V. Vidyaraman and S. J. Upadhyaya, "Spycon: Emulating user activities to detect evasive spyware", IPCCC, pp. 502-509, 2007.

Li, Bo, et al. "Large-scale identification of malicious singleton files." Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy. ACM, 2017.

Thilagamani, S., Shanti, N.," Gaussian and gabor filter approach for object segmentation", Journal of Computing and Information Science in Engineering, 2014, 14(2), 021006, https://doi.org/10.1115/1.4026458

Tang, MingJian, MamounAlazab, and Yuxiu Luo. "Big data for cybersecurity: Vulnerability disclosure trends and dependencies." IEEE Transactions on Big Data (2017). to be published.

M. Alazab, S. Venkatraman, P. Watters, M. Alazab, and A. Alazab, ''Cyber-crime: The case of obfuscated malware,'' in Global Security, Safety and Sustainability & e-Democracy (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), vol. 99, C. K. Georgiadis, H. Jahankhani, E. Pimenidis, R. Bashroush, and A. Al-Nemrat, Eds. Berlin, Germany: Springer, 2012.

Rhagini, A., Thilagamani, S. ,"Women defence system for detecting interpersonal crimes",International Journal of Advanced Science and Technology, 2020, Volume 29,Issue7S, pp. 1669–1675

Su, Jiawei, et al. "Lightweight classification of IoT malware based on image recognition." 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC). Vol. 2. IEEE, 2018.

K.Deepa, S.Thilagamani, "Segmentation Techniques for Overlapped Latent Fingerprint Matching", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-12, October 2019. DOI: 10.35940/ijitee.L2863.1081219

Naeem, Hamad, Bing Guo, and Muhammad Rashid Naeem. "A light-weight malware static visual analysis for IoT infrastructure." 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD). IEEE, 2018

Deepa. K , LekhaSree. R , Renuga Devi. B , Sadhana. V , Virgin Jenifer. S ,"Cervical Cancer Classification", International Journal of Emerging Trends in Engineering Research, 2020, 8(3), pp. 804–807 https://doi.org/10.30534/ijeter/2020/32832020

Nataraj, Lakshmanan, et al. "Sarvam: Search and retrieval of malware." Proceedings of the Annual Computer Security Conference (ACSAC) Worshop on Next Generation Malware Attacks and Defense (NGMAD). 2013.

Santhi, P., Priyanka, T.,Smart India agricultural information reterival system, International Journal of Advanced Science and Technology, 2020, 29(7 Special Issue), pp. 1169–1175.

Anderson, Hyrum S., et al. "Evading machine learning malware detection." Black Hat (2017).

Pascanu, Razvan, et al. "Malware classification with recurrent networks." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.

Raff, Edward, et al. "An investigation of byte n-gram features for malware classification." Journal of Computer Virology and Hacking Techniques 14.1 (2018).

Santhi, P., Lavanya, S., Prediction of diabetes using neural networks, International Journal of Advanced Science and Technology, 2020, 29(7 Special Issue), pp. 1160–1168

Alam, Mohammed S., and Son T. Vuong. "Random forest classification for detecting android malware." 2013 IEEE international conference on green computing and communications and IEEE Internet of Things and IEEE cyber, physical and social computing. IEEE, 2013.

Vijayakumar, P, Pandiaraja, P, Balamurugan, B & Karuppiah, M 2019, 'A Novel Performance enhancing Task Scheduling Algorithm for Cloud based E-Health Environment', International Journal of E-Health and Medical Communications (IJEHMC), Vol 10,Issue 2,pp 102-117.

E. S. Solutions and Q. Heal, "Quick Heal Quarterly Threat Report | Q1 2017," 2017 url:http://www.quickheal.co.in/resources/threat-reports

P. Pandiaraja, N Deepa 2019 ," A Novel Data Privacy-Preserving Protocol for Multi-data Users by using genetic algorithm" , Journal of Soft Computing , Springer , Volume 23 ,Issue 18, Pages 8539-8553.

M. G. Schultz, E. Eskin, and S. J. Stolfo, "Data Mining Methods for Detection of New Malicious Executables," 2001.

D. Bilar, "Opcodes As Predictor for Malware," International Journal of Electronic Security and Digital Forensics, vol. 1, no. 2, pp. 156–168, 2007.

N Deepa , P. Pandiaraja, 2020 ," Hybrid Context Aware Recommendation System for E-Health Care by merkle hash tree from cloud using evolutionary algorithm" , Journal of Soft Computing , Springer , Volume 24 ,Issue 10, Pages 7149–7161. .

R. Moskovitch, D. Stopel, C. Feher, N. Nissim, N. Japkowicz, and Y. Elovici, "Unknown malcode detection and the imbalance problem," Journal in Computer Virology, vol. 5, no. 4, pp. 295–308, 2009.

N Deepa , P. Pandiaraja, 2020 , " E health care data privacy preserving efficient file retrieval from the cloud service provider using attribute based file encryption ", Journal of Ambient Intelligence and Humanized Computing , Springer , https://doi.org/10.1007/s12652-020-01911-5

I. Santos, J. Nieves, and P. G. Bringas, "Semi-supervised learning for unknown malware detection," International Symposium on Distributed Computing and Artificial Intelligence. Springer Berlin Heidelberg, vol. 91, pp. 415–422, 2011.

K Sumathi, P Pandiaraja 2019," Dynamic alternate buffer switching and congestion control in wireless multimedia sensor networks" , Journal of Peer-to-Peer Networking and Applications , Springer , Volume 13,Issue 6,Pages 2001-2010

P. K. Chan and R. Lippmann, "Machine learning for computer security,"Journal of Machine Learning Research, vol. 6, pp. 2669–2672, 2006

Shankar, A., Pandiaraja, P., Sumathi, K., Stephan, T., Sharma, P. ," Privacy preserving E-voting cloud system based on ID based encryption " Journal of Peer-to-Peer Networking and Applications , Springer , https://doi.org/10.1007/s12083-020-00977-4.

M. Chandrasekaran, V. Vidyaraman, and S. J. Upadhyaya, "Spycon: Emulating user activities to detect evasive spyware," in IPCCC. IEEE Computer Society, 2007, pp. 502–509.

Bayer, U., Kruegel, C. and Kirda, E. (2006) TTAnalyze: A Tool for Analyzing Malware. Proceedings of the 15th European Institute for Computer Antivirus Research Annual Conference

Dinaburg, A., Royal, P., Sharif, M. and Lee, W. (2008) Ether: Malware Analysis via Hardware Virtualization Extensions. Proceedings of the 15th ACM Conference on Computer and Communications Security, CCS'08, Alexandria, 27-31 October 2008, 51-62.

Gupta, D., & Rani, R. (2018). Big Data Framework for ZeroDay Malware Detection. Cybernetics and Systems, 49(2), 103-121

Ray, A., & Nath, A. (2016). Introduction to Malware and Malware Analysis: A brief overview. International Journal, 4(10).

AlAhmadi, B. A., & Martinovic, I. (2018, May). MalClassifier: Malware family classification using network flow sequence behaviour. In APWG Symposium on Electronic Crime Research (eCrime), 2018 (pp. 1-13). IEEE.

Khan, M. H., & Khan, I. R. (2017). Malware Detection and Analysis. International Journal of Advanced Research in Computer Science, 8(5).

Wang, C., Ding, J., Guo, T., & Cui, B. (2017, November). A Malware Detection Method Based on Sandbox, Binary Instrumentation and Multidimensional Feature Extraction. In International Conference on Broadband and Wireless Computing, Communication and Applications (pp. 427-438). Springer, Cham.

Gupta, S., Sharma, H., & Kaur, S. (2016, December). Malware Characterization Using Windows API Call Sequences. In International Conference on Security, Privacy, and Applied Cryptography Engineering (pp. 271-280). Springer, Cham.

Schultz, M. G., Eskin, E., Zadok, F., &Stolfo, S. J. (2001). Data mining methods for detection of new malicious executables. In Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on (pp. 38-49). IEEE.

Kosmidis, K., &Kalloniatis, C. (2017, September). Machine Learning and Images for Malware Detection and Classification. In Proceedings of the 21st Pan-Hellenic Conference on Informatics (p. 5). ACM

Gandotra, E., Bansal, D., &Sofat, S. (2014, September). Integrated framework for classification of malwares. In Proceedings of the 7th International Conference on Security of Information and Networks (p. 417). ACM.