Malware Family Classification Model Using User Defined Features and Representation Learning

T. Gayathri, M. S. Vijaya

MALWARE FAMILY CLASSIFICATION MODEL USING USER DEFINED FEATURES AND REPRESENTATION LEARNING

Gayathri T[1][0000-0003-1638-6668] and vijaya M S[2]

[1] PSGR Krishnammal College for Women, Coimbatore
Gayathrithangamuthu73@gmail.com
[2] PSGR Krishnammal College for Women, Coimbatore
msvijaya@psgrkc.ac.in

Abstract. Malware is very dangerous for system and network user. Malware identification is essential tasks in effective detecting and preventing the computer system from being infected, protecting it from potential information loss and system compromise. Commonly, there are 25 malware families exists. Traditional malware detection and anti-virus systems fail to classify the new variants of unknown malware into their corresponding families. With development of malicious code engineering, it is possible to understand the malware variants and their features for new malware samples which carry variability and polymorphism. The detection methods can hardly detect such variants but it is significant in the cyber security field to analyze and detect large-scale malware samples more efficiently. Hence it is proposed to develop an accurate malware family classification model contemporary deep learning technique. In this paper, malware family recognition is formulated as multi classification task and appropriate solution is obtained using representation learning based on binary array of malware executable files. Six families of malware have been considered here for building the models. The feature dataset with 690 instances is applied to deep neural network to build the classifier. The experimental results, based on a dataset of 6 classes of malware families and 690 malware files trained model provides an accuracy of over 86.8% in discriminating from malware families. The techniques provide better results for classifying malware into families.

Keywords: Malware classification, machine learning, representation learning, deep neural network.

1 INTRODUCTION

Malware is a malicious code, and it is harmful when executed on a computing device or system. Software that is specifically designed to disrupt and damage, or gain

unauthorized access to a computer system. These are proposed to gain access to computer systems and network resources, disturb computer operations, and gather personal information without taking the consent of system's owner, thus creating a menace to the availability of the internet, integrity of its hosts, and the privacy of its users.

The threats posed by malware are perceived as too numerous and agile to be managed by humans in a meaningful way. Generic detection of malware comes at the cost of precision; leading to information that is often limited to the fact that malware has been detected. This is useful for much of the network defense community in incident response and remediation activities often need correct and concise identification of malware-related threats. Such identification provides incident responders with the information necessary to understand what threats they are actually facing and to allocate resources accordingly.

Web usage mining is discovery of meaningful pattern from data generated by client server transaction on one or more web localities. Several web transactions automatically make the data which gets gathered in server access logs, refers logs, agent logs, client side's cookies, user profile, metadata, page attribute, page content and site structure. Search engines began to understand their unintended contribution in malware distribution. Web services enable the detection of malware with a huge partner's data. But even a massive malware database does not guarantee detection of recent ones. Most of anti-malware software products, such as Kaspersky, Symantec, and MacAfee normally use the signature based method to recognize threats. This malicious software can perform heterogeneity of functions such as encrypting and destroy data, hijacking core computing functions and accessing user system activity without their permission.

The damage affected by a virus that corrupts a computer or a corporate network can be different from an irrelevant increase in outgoing traffic to the complete network breakdown or the loss of hypercritical data. The scale of the damage depends on the purpose of the virus, and sometimes the results of its activity are undetectable for the users of a compromised machine. A virus on a commercial network can be considered a force mature and the damage affected by it as being equal to the loss associated with the network downtime essential for disinfection.

Malwares come in wide range of variations like Virus, Worm, Trojan horse, Ransomware, Backdoor, Zeus, Key loggers, Adware and six malware families such as allaple, cryptolocker, agent, Trojan generic, wannacry and zbot are taken into account for implementation.

2 LITERATURE SURVEY

Several research works have been carried out currently using machine learning. Various features such as static and dynamic have been used to build models. In few cases signature based methods, image processing methods have been used. Based on the study of various literatures available on malware family classification, a brief report is presented in this section about the developments in the respective area in the last several years.

Schultz et al. [1] were the first to introduce the concept of data mining for detecting malwares. They applied three different static features for malware classification: Portable Executable (PE), strings and byte sequences. They used a data set consisted of 4266 files including 3265 malicious and 1001 benign programs. The Naive Bayes algorithm, taking strings as input data, gave the highest classification accuracy of 97.11%. The authors claimed that the rate of detection of malwares using data mining method was twice as compared to signature based method.

Nari et al. [2] presented a framework for automated malware classification into their respective families based on network behavior. Network traces were taken as input to the framework in the form of pcap files, from which the network flows were extracted. From these behavior graphs, the features like graph size, root out-degree, average out-degree, maximum out-degree, number of specific nodes were extracted. These features were used to classify malwares using classification algorithms available in WEKA library and it was concluded that J48 decision tree performs better than other classifiers.

Nataraj et al. [3] were the first to explore the use of byte plot visualization for automatic malware classification. They converted all the malware samples to grayscale byte plot representations and extracted texture-based features from the malware image. They used an abstract representation technique, GIST (global image descriptors), for computing texture features from images. The dataset consisted of 9,458 malware samples belonging to 25 different classes, collected from the Anubis system. They used the global image-based features to train a K-Nearest Neighbour model, with Euclidean distance as distance measure, to classify malware samples into their respective classes and an accuracy of 97.18% was obtained.

Rieck et al. [4] suggested a new method for automated identification of new classes of malware with similar type of behavior using clustering and classifying previously unseen malware to these discovered classes by classification using machine learning. They used more than 10,000 malware samples, belonging to 14 different families, in their experiment. These malware samples were collected using honeypots and spamtraps. They reported an accuracy of 88% on family classification using simple SVM based classifier.

Tian et al. [5] used function length frequency to classify Trojans. Function length was measured by the number of bytes in the code. Their results specify that the function length along with its frequency was important in classifying malware family and can be grouped with another feature for quick and scalable malware families' classification. They applied machine learning algorithms available in the WEKA library for classifying malware. They used a data set of 1368 malware to demonstrate their work and achieved an accuracy of over 97%.

In the existing work, the classifications were performed by training malware executable files or images or PE (portable executable) files. The features based on texture were extracted from executable files that assisted in recognizing the malwares or in classification of malware. In a few cases, malware image datasets were used and the corresponding binary values were trained to build classifiers. This motivated to carry out research work for building the malware family classification model by deriving an

array of binary features and training the feature dataset through deep learning classifiers.

3 MALWARE FAMILY CLASSIFICATION MODEL

The problem of malware family identification is formulated as multiclass classification task and solved using machine learning technique. The methodology of proposed malware family classification includes four different stages. In the first stage, malware family corpus development is performed wherein the virus executable files corresponding to six families of malware are collected. The second stage is malware family dataset creation. In this stage, the executable files are converted into images which are then converted into binary arrays to form the feature vectors of malware family dataset. In the third stage, the training dataset is used to develop the malware family classification models by implementing supervised learning and deep learning algorithms. Finally, the classification models are evaluated in terms of precision, recall, F-measure, and accuracy. The methodology of the malware family classification model is shown in Fig.1

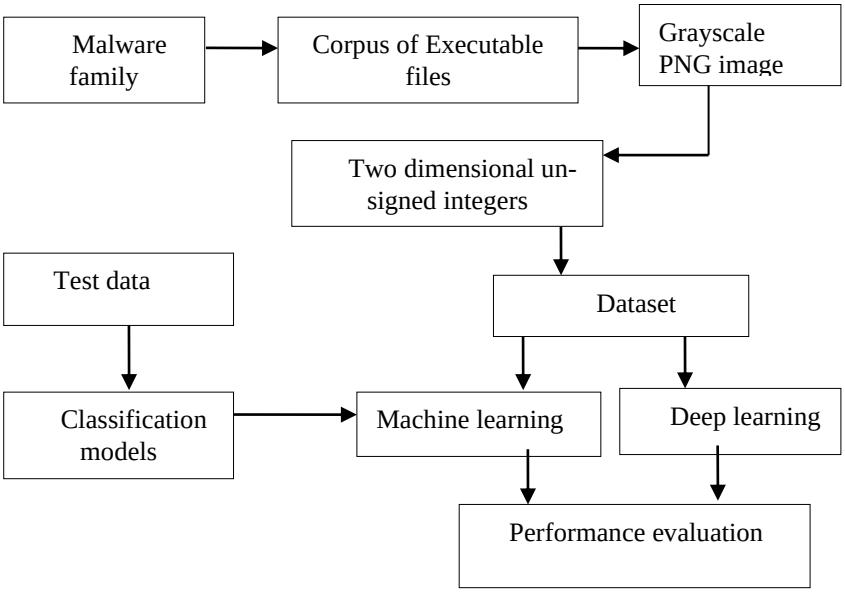

Fig.1. Methodology of Malware Family Classification Model

3.1 Corpus Development and Dataset Creation

Initially a corpus consisting of 690 malware executable files related to six types of malware family is developed. Here allaple, cryptolocker, agent, Trojan generic, wannacry and zbot are six different types of malware family considered. These viruses corresponding to six malware families are collected from the das malwark, virustotal and a malwr.com malware samples. These portals have to collect malware using honeypots and also users around the world submit files over them for analysis and sharing malware samples. A set of 100 files of allaple malware, 150 exe files of cryptolocker, 100 files of agent family, 150 files of Trojan generic, 90 files of wannacry and 100 files of zbot malware family has been collected and a malware family corpus of size 690 is developed. Malware executable file in hexadecimal view is shown in Fig.2

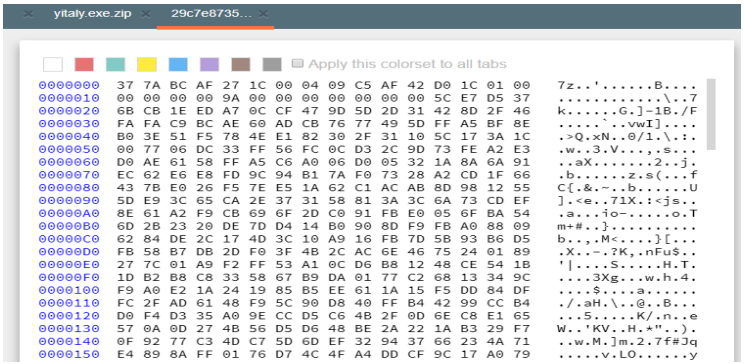

Fig.2. Malware Executable Files in Hexadecimal View

Malware coders change small parts of the original source code to produce a new variant. Images can capture small changes yet retain the global structure. Images give more information about the structure of the malware. The malware files in the corpus are converted into a grayscale PNG (Portable Network Graphics) image of size 64 x 64 using python code. Conversion of executable file to image results are shown in Fig.3

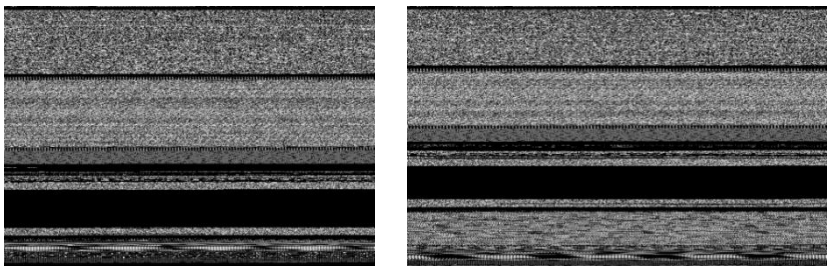

Fig.3. Image of Zbot Exe file

A sequence of pixel values corresponding to each byte is stored in the binary file as a sequence of ones and zeros. Width of the image is set according to the file size. Height of the image varies depending on the file size. Each image is resized to 28 x 28 and represented in binary form. The binary files are difficult to read as its dimension is high. Hence, it is mapped into a one dimensional array of integers between 0 and 255. It is a vector of 8-bit unsigned integers. This vector is then reshaped into a two dimensional array of unsigned integers. The resizing is done based on the height and width of the file size. The height of the file is the total length of the one dimensional array and the width of the array is nothing but the file size. Finally, the two dimensional feature matrix of size 28x28 amounting to 784 feature values is derived.

The rows of 2D matrix is organized as a feature vector of size 784 and the dimensionality reduction technique namely PCA (Principle Component Analysis) is used to reduce the size of the feature vectors. This dimensionality reduction method efficiently represents interesting parts of an array as a compact feature vector of size 480. Data transformation and dimensionality reduction are performed to achieve maximum prediction accuracy.

In this manner all the 690 virus exe files are converted into feature vectors which are then assigned class labels 1 to 6 as 1 for alleple, 2 for cryptolocker, 3 for agent, 4 for Trojan generic, 5 for wannacry and 6 for zbot. Finally a malware family dataset with 690 instances of dimension 480 is created and stored in a csv file.

3.2 Model Building

Malware family classification model is built with the above dataset as input and a deep learning architecture namely deep neural network is employed to build the malware family classification model. Deep neural network performs the representation learning from the features and self-extracts hidden patterns during training. Various hyper parameters such as learning rate, epochs, loss function, activation function and optimizers are defined to improve the efficiency of learning and to build the accurate malware family classification model. The learning rate is used to adjust the weights and back propagate the weights to make correct predictions. Epochs defines the number of iterations, the learning algorithm work through the entire training dataset for leaning the patterns efficiently. The learning algorithms train the patterns through the layers such that the error rate of the model is sufficiently minimized. Loss function helps in optimizing the parameters of the neural networks. This is minimizing the loss for a neural network by optimizing its parameters. The loss is calculated using loss function by matching the target value and predicted value by a neural network. The optimization is used to produce slightly better and faster results by updating the model parameters such as weights and bias values. The performance of the model is evaluated using various metrics such as precision, recall, F measure and accuracy.

4 EXPERIMENTS AND RESULTS

Six types of malware families are taken into account for developing the classifiers and hence malware recognition problem turn out to be multi classification. In these experiments, the sequential model is developed to build DNN classifier. The hyper parameters used to build model are learning rate is 0.001 and dropouts is 0.2, 0.3. The input layers have been given 480 attributes and similarly the hidden layer has been given as a 240 in sequential model. The output layers have been specified with class labels 1 to 6 for recognizing six families of malwares. The activation functions namely softmax and relu with adam optimizers are used in this work. A softmax function is used in output layer and relu functions used for hidden layer. Adam optimizer is used with value of epochs is 200 and batch size is 10 to increase the prediction accuracy. Deep neural network is prepared to assign the layers and input of one layer is passed as output to the other layer. The functioning of DNN based malware family classification method is determined based on classification metrics. Prediction accuracy of 86.8% is achieved by DNN classifier and evaluated with respect to six class labels i.e. are allaple, cryptolocker, agent, Trojan generic, wannacry and zbot. The precision is high for allaple, agent and zbot with the value of 1.00 and recall value of 1.00 is maximum for class allaple and zbot. The F-measure with the value of 1.00 is excessive for class allaple and zbot. The average values of DNN methods with precision of 0.82, recall of 0.80 and F-measure of 0.85 is obtained for six malware family types. Class-wise performance of DNN model is shown in Table 1 and results obtained for various dropouts and epochs are illustrated in Table 2.

Table 1. Performance Results of Deep Neural Network Classifier by Class

Class labels	**Precision**	**Recall**	**F-measure**
Allaple	1.00	1.00	1.00
Cryptolocker	0.97	0.95	0.96
Agent	1.00	0.96	0.98
Trojan generic	0.78	0.64	0.68
Wannacry	0.48	0.58	0.53
Zbot	1.00	1.00	1.00
Accuracy	0.78	0.86	0.81

The malware classification models based on deep neural network is built with performance of the model evaluated using classification metrics such as precision, recall, F-score and accuracy. The performance results of the deep neural network classification based on various metrics are shown Table 2.

Table 2. Performance Results of Deep Learning

Dropout	Epochs	Precision	Recall	F-measure	Accuracy
0.2	50	0.58	0.68	0.60	0.63
	100	0.78	0.77	0.77	0.77
	150	0.80	0.81	0.80	0.79
	200	**0.75**	**0.81**	**0.77**	**0.81**
0.3	50	0.75	0.83	0.78	0.80
	100	0.81	0.80	0.79	0.83
	150	0.84	0.83	0.83	0.85
	200	**0.82**	**0.80**	**0.85**	**0.86**

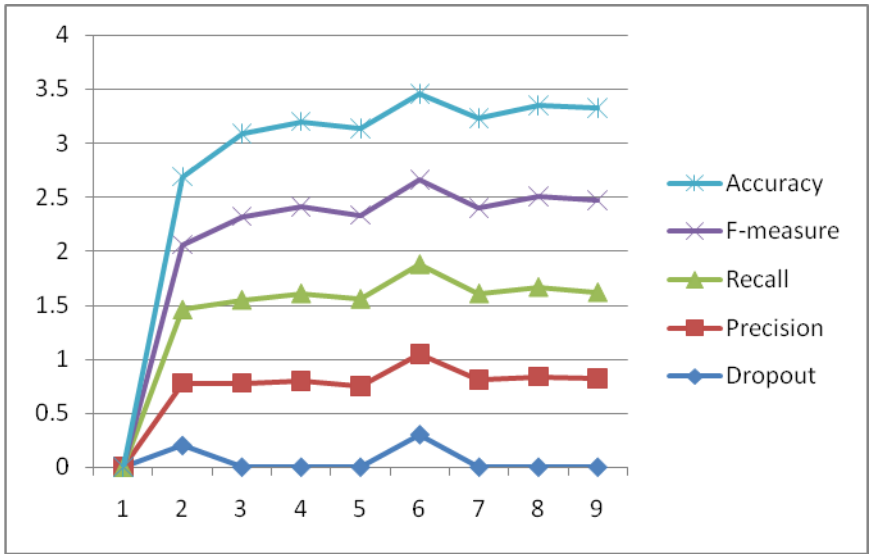

Fig.8. Performance Results of Deep Learning at Various Dropouts

The results of DNN classifier is compared with the implementation results of supervised learning algorithms such as decision tree, neural networks, support vector machine which have been implemented using the same dataset. It is discovered that deep neural network had achieved highest prediction accuracy of 86.8% was acquired by trained supervised models. DNN classifier obtained a precision of 0.82, recall val-

ue of 0.80 with F-measure of 0.85 and accuracy 86.8%. The comparative predictive performances of malware family classifiers are shown in Table 3.

Table 3. Comparative Results of DNN and Supervised Classification Algorithms

Performance Evaluation	Classifiers			
	Neural Network (%)	Decision Tree (%)	Support Vector Machine (%)	Deep Neural Network (%)
Precision	75	80	76	82
Recall	82	86	81	80
F-measure	78	86	78	85
Accuracy	85	82	85	86

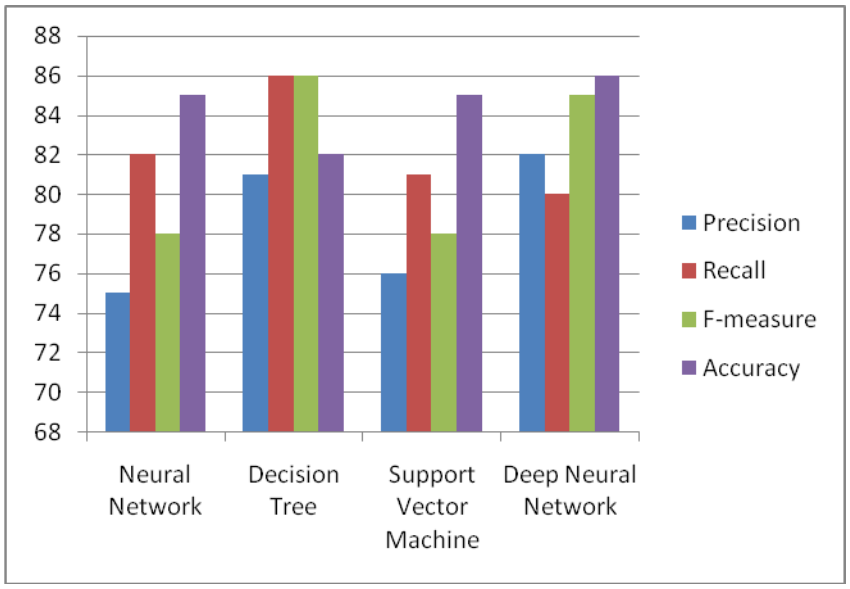

Fig.9. Comparison of Classifiers Based on Various Evaluation Metrics

The deep neural network model have capability to modify the weights in deep neural network so the error rate is minimized which gives accurate prediction. The significant result of deep neural network based classification model is attained by maximizing hidden layers and number of epochs in sequential model. The performance of the

deep neural network classifier is validated using its measures with high accuracy and least error rate for malware family classification. Deep neural network classifier achieves better performance through image features and outperforms with supervised learning classification models.

5 CONCLUSION

This paper demonstrated the implementation of deep neural network based malware family classification model built with the aim to predict the malware family through DNN classifier using malware family dataset. DNN method automatically learns high level features from the user defined features in the malware family dataset. The executable malware files are collected from various malware resources and malware family corpus has been developed. The performance of the supervised and deep learning classifiers is evaluated in terms of accuracy, precision, recall, F-measure and the results are compared. The DNN based malware family classification model has shown high classification accuracy compared with supervised classifiers. Finally, it is concluded that deep neural network based classification method performs accurate prediction in classifying the malware families.

References

1. Schultz, M., Eskin, E., Zadok, F. and Stolfo, S., Data Mining Methods for Detection of New Malicious Executables, Proceedings of 2001 IEEE Symposium on Security and Privacy, Oakland, 14-16 May 2001, 38-49 (2001)
2. Nari, S. and Ghorbani, A. Automated Malware Classification Based on Network Behavior. Proceedings of International Conference on Computing, Networking and Communications (ICNC), San Diego, 642-647 (2013)
3. Nataraj, L., Karthikeyan, S., Jacob, G. and Manjunath, B., Malware Images: Visualization and Automatic Classification, Proceedings of the 8th International Symposium on Visualization for Cyber Security, Article No. 4 (2011)
4. Rieck, K., Trinius, P., Willems, C. and Holz, T., Automatic Analysis of Malware Behavior Using Machine Learning. Journal of Computer Security, 19, 639-668 (2011)
5. Tian, R., Batten, L., Islam, R. and Versteeg, S. An Automated Classification System Based on the Strings of Trojan and Virus Families. Proceedings of the 4th International Conference on Malicious and Unwanted Software, Montréal, (2009)
6. Park, Y., Reeves, D., Mulukutla, V. and Sundaravel B., Fast Malware Classification by Automated Behavioral Graph Matching, Proceedings of the 6th Annual Workshop on Cyber Security and Information Intelligence Research, Article No. 45 (2010)
7. Bugra Cakir and Erdogan Dogdu Malware Classification Using Deep Learning Methods on ACM SE '18: ACM SE '18: Southeast Conference, Richmond, KY, USA. ACM, New York, NY, USA, 5 pages (2018)
8. M. Bailey, J. Oberheide, J. Andersen, Z. M. Mao, F. Jahanian, and J.Nazario automated classification and analysis of internet malware. In Proceedings of the 10th Symposium on Recent Advances in Intrusion Detection (2007)

9. Dolly Uppal, Rakhi Sinha, Vishakha Mehra and Vinesh Jain Malware Detection and Classification Based on Extraction of API Sequences International Conference on Advances in Computing,Communications and Informatics (ICACCI) (2014)
10. A. Makandar and A. Patrot Malware image analysis and classification using support vector machine, International Journal of Trends in Computer Science and Engineering, vol. 4, no. 5, pp. 01–03 (2015)
11. Zolkipli, M.F. and Jantan, A. An Approach for Malware Behavior Identification and Classification. Proceeding of 3rd International Conference on Computer Research and Development, Shanghai, 11-13 March 2011, 191-194 (2011)
12. Biley, M., Oberheid, J., Andersen, J., Morley Mao, Z., Jahanian, F. and Nazario, J. Automated Classification and Analysis of Internet Malware. Proceedings of the 10th International Conference on Recent Advances in Intrusion Detection, 4637, 178-197 (2007)
13. Islam, R., Tian, R., Battenb, L. and Versteeg, S. Classification of Malware Based on Integrated Static and Dynamic Features. Journal of Network and Computer Application, 36, 646-556 (2013)
14. Niket Bhodia, Pratikkumar Prajapati, Fabio Di Troia and Mark Stamp, Transfer Learning for Image-Based Malware Classification 3rd International Workshop on Formal Methods for Security Engineering (ForSE 2019), in conjunction with the 5th International Conference on Information Systems Security and Privacy (ICISSP 2019), At Prague, Czech Republic (2019)
15. Ekta Gandotra, Divya Bansal, Sanjeev Sofat, Malware Analysis and Classification: A Survey Journal of Information Security, 5, 56-64 (2014)
16. Fathimath Zuha Maksood Analysis of Data Mining Techniques and its Applications International Journal of Computer Applications (0975 – 8887) Volume 140 – No.3, April (2016)
17. Shaik. Irfan Babu , Dr. M.V.P. Chandra Sekhara Rao, G.Nagi Reddy, Research Methodology on Web Mining for Malware detection International Journal of Computer Trends and Technology (IJCTT) – volume 12 number 4 (2014)
18. Dragos Gavrilut, Mihai Cimpoes, Dan Anton, and Liviu Ciortuz, Malware detection using machine learning Proceedings of the International Multi conference on Computer Science and Information Technology, page 735–741(2009)
19. Joshua Saxe, Konstantin Berlin, deep neural network based malware detection using two dimensional binary program features, 10th International Conference on Malicious and Unwanted Software: "Know Your Enemy" (MALWARE) (2015)
20. Ivan Dychka,Denys Chernyshev, Malware Detection Using Artificial Neural Networks, ICCSEEA ,Advances in Computer Science for Engineering and Education II, volume 938, pp 3-12 (2019)
21. Rafiqul Islam Irfan Altas, A Comparative Study of Malware Family Classification, International Conference on Information and Communications Security ICICS 2012: Information and Communications Security pp 488-496 (2012)
22. J. Saxe and K. Berlin, Deep Neural Network Based Malware Detection Using Two Dimensional Binary Program Features, in Proceedings of the 10th International Conference on Malicious and Unwanted Software (MALWARE), IEEE, pp. 11–20 (2015)
23. Nazario, J., Oberheid, J., Andersen, J., Morley Mao, Z., Jahanian, F. and Biley, M. Automated Classification and Analysis of Internet Malware, Proceedings of the 10th International Conference on Recent Advances in Intrusion Detection, 4637, 178-197 (2007)

24. A. Mohaisen and O. Alrawi., Unveiling Zeus: Automated Classification of Malware Samples, In Proceedings of the 22nd International Conference on World Wide Web, pages 829–832 (2013)
25. Kong, D. and Yan, G., Discriminate Malware Distance Learning on Structural Information for Automated Malware Classification, Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems, 347-348 (2013)
26. Rieck, K., Holz, T., Willems, C., Learning and Classification of Malware Behavior, In International Conference on Detection of Intrusions and Malware pp. 108–125 (2008)
27. Anil Thomas , Hermineh Sanossian, Jack W Stokes, Razvan Pascanu , and Mady Marinescu, Malware classification with recurrent networks, In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on IEEE, 1916–1920 (2015)