Malware Analysis and Classification using Artificial Neural Network

Aziz Makandar
Professor
Department of Computer Science
Karnataka State Women's University, Vijayapur
azizkswu@gmail.com

Anita Patrot
Research Scholar
Department of Computer Science
Karnataka State Women's University, Vijayapur
patrotanita@gmail.com

Abstract— **Today major and serious threat on internet is malicious software or data which damage the system. Malware variants identification and classification is the one of the most important research problem in digital forensics. Malware binaries are set of instructions which may affect your system without your authority. Many researchers worked in this area mainly relied on specific API calls, sequences of bytes, statistic and dynamic analysis is used for detection and classification of malware. The proposed method malware is represented as 2Dimensional gray scale image is observed malware images of all the available variants and their texture similarity, which motivate to classify malware based on texture features. The texture plays a very significant role in identify and classify malware. The objective of this paper is to identify a behavior of malicious data based on global features using Gabor wavelet transform and GIST. The experiment done on Mahenhur dataset which includes 3131 binaries samples comprising 24 unique malware families. The algorithm has been implemented using feed forward Artificial Neural Networks (ANN) it gives their overview uniqueness. The experimental results are promising to effectively detecting and classifying malware with good accuracy 96.35 %.**

Keywords— ANN; Gabor wavelet; GIST; Image Processing; Malware; Texture Analysis

I. INTRODUCTION

Internet plays a very important role which also motivates the unauthorized access. Today development of the internet and their uses is growing day by day which motivates the number of malware distributes more, especially for economic profits. According to the report of symantec every day a millions of malware variants are observed an exigent task to say zero day attack. Malware is a term used to refer a variety of forms of unsympathetic or intrusive software including computer viruses, worms and other malicious programs. It can take form of executables code and script content and other software [1]. Malware analysis includes two type static analysis and dynamic analysis. Static analysis which includes the signatures of malware identified. Malware is a term used for malicious data that get installed on your machine and performs unwanted tasks such as stealing passwords and data. Malware visualization is a field of knowledge that focuses on representing malware in the form of visual features. That could possibly be used to deliver more information about a particular malware. Graphical visualization helps to gain more information about malware. Its ever increasing new malware produced by every day is a challenging task [2]. The exponential increase in the number of new signatures released every year [3] Symantec reported corpus over 286 million in 2010, to 2,895,802 new signatures in 2009, to 169,323 in 2008.

The boarder level all malicious data stored in drives can be represented as a binary string made up of number of zeros and ones. This represents the binary string which is reshaped in to a matrix and represented as grayscale image. That's why the description of all malicious data is converted into gray scale image. The description of an image has been well studied in the field of computer vision. GIST descriptors specially used on scene classification based on texture and object identification as well as classification. The descriptions are forwarded into classification algorithm for training and testing of malware image. The proposed work for automatic identifies a malware behavior using machine learning techniques such as ANN. The feed forward method is used to classify malware binaries. The training has done on malware images and testing total 3131 images of Mahenhur dataset. The texture features are extracted by applying Gabor wavelet with 8 orientations and 4 scales. Total 320 features are select to train the malware by using neural network tool.

II. MALWARE VISUALIZATION

A. Malware Analysis

The malware analysis is a study of malware behavior to detect its different components; the components may be different variants which may affect system without any influence. The malware analysis has two type static analysis and dynamic analysis. The static analysis works on only signatures of malware variants by globally. Dynamic analysis works on run time checking of op code on API calls whether it is a malicious data or not. The proposed method malware is visualized as grayscale image by taking binaries to 8 bit vector. The fig.1 illustrates the visualization technique of

978-1-4673-6667-0/15/$31.00©2015 IEEE

malware. The grayscale image is having 256 gray levels, 0-255 range values, where 0 represents white and 255 represents black color. Every 8 bits is considered as a one pixel in grayscale image. To identify behavior of malware based on global features of samples.

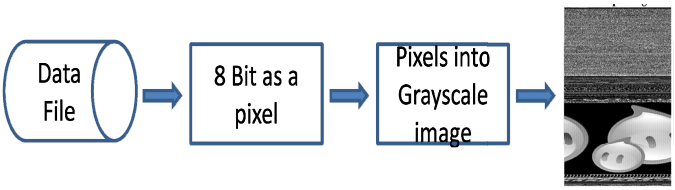

Fig. 1. Malware Image Visualization

B. Mahenhur Dataset

The database consists of around 3131 samples shown in Table I. This includes total 24 different unique families and their number of samples present in each family. It allows identifying malware classes with similar texture behavior and assign to unknown to known malware. The capable of processing the texture feature based on behavior of thousand malware binaries on daily basis. Each malware family consist of maximum 300 file samples which contains variants of malware according to the standard antivirus software they have assigned a 24 different family names based on variants belongs to that category.

TABLE I. Malware Dataset

Sl. No	Malware Dataset	
	Malware Family	*No. of Samples*
1	ADULTBROWSER	262
2	ALLAPLE	300
3	BANCOS	48
4	CASINO	140
5	DORFDO	65
6	EJIK	168
7	FLYSTUDIO	33
8	LDPINCH	43
9	LOOPER	209
10	MAGICCASINO	174
11	PONDNUHA	300
12	POISON	26
13	PORNDIALER	97
14	RBOT	101
15	ROTATOR	300
16	SALITY	84
17	SPYGAMES	139
18	SWIZZOR	78
19	VAPSUP	45
20	VIKING_DLL	158
21	VIKING_DZ	68
22	VIRUT	202
23	WOIKOINER	50
24	ZHELATIN	41
Total		**3131**

III. Related Work

The classification of malware based on behavior of different malware is done by using machine learning techniques and data mining techniques. At present there are few researches on malware visualization technique. Kyoung Soo Han.et.al [5] proposed a algorithm that a new malware family classification by converting binaries into 2 dimensional gray scale images and computing entropy of individual families and similarities also calculated and stored in feature vector in database and entropy graph effectively distinguish a malware families. The state of an art of malware is still ongoing research of many antivirus companies [6]. Natraj. L.et.al [7-8] the author introduces malware binaries into 8 bit grayscale image and extracted global features of malware, this author introduces a malware visualization novel method to classify a malware based on binary texture analysis. The author proposed a file fragment into grayscale image and classification done by using machine learning techniques.

Kong.et.al [9] proposed a malware samples as malware classification is done on distance learning based for identification of malware structural information. Tantan Xu.et.al [10] they classified a fragment in terms of two models file unbiased and type unbiased. G.Conti.et.al [11] the fragments are treated as a gray scale image, classification done on fragments they provides preliminary solution for automatic classification. Shui.yu.et.al [12] the malware detection and classification is done using data mining concept. Traditional way to detect and classify malware is through scanning technology based on the code as well as PE faces a rigorous challenging task [12]. In order to overcome defects in data mining and machine learning techniques introduce a field Antivirus [13]. There are various methods of detection of malware and classification including graph based detection of malware [14-16], instruction sequence based classification [17-18], API based classification [19-20].

The malware analysis and detection [21] recently various researchers uses visualization technique to understand malware visually that can help antivirus software to detect malware

effectively. The GIST feature extraction method is also used for scene classification and human identification [22]. The malware analysis and identification recently various researchers uses visualization technique to understand malware visually that can help antivirus software to detect malware efficiently. The GIST method is used for texture feature extraction which is also used for scene classification and human identification [23-25]. The visualized the malware behavior into tree maps and thread graph by using API calls [26-32]. The following table 2 illustrates the malware visualization techniques, types of analysis, feature types and limitations of visualization techniques. Syed Zainudeen Mohd Shaid et.al [4][29] proposed visualizing malware behavior and different types of analysis as well as features used for classification of malware with high accuracy.

IV. Proposed Work

Proposed work analyzing the malware by understanding behavior and applying global texture features. The Gabor and GIST is used to identify a malware behavior and similarities among samples based on texture feature similarities are observed and extracted 320 features. The algorithm has three steps as shown in figure. 5

STEP 1: Read the binaries into 8 bit for each pixel in grayscale image. The bitmap image conversion is done.

STEP 2: Resize the grayscale image, the width of the image is fixed the height of the image is varied based on file size. After resizing image into 64*64 apply sub band filtering with 8 orientations and 4 scales by using Gabor Wavelet and GIST. Local Representation of an image is

$$I^L(x) = \{ i_1(x), i_2(x),\ldots..,i_j(x),\ldots,i_N(x)\} \quad \text{--------- (1)}$$

Where, 'I' is an image of M X N matrix of number of rows and columns, the set of sub block of gray scale image is extracted and calculating average of each sub blocks after sub band filtering.

The averaging window is represented as in equation.2. The GIST descriptor gives by default 512 feature vector in that we are taking high features to train 666 malware images in neural network tool with 50 iterations for each epoch. M is the total average of sub block which is calculated by using mean, where ∑ is a sum of average sub blocks.

$$m(x) = \sum I^L(x')\, w(x' - x) \quad \text{---------- (2)}$$

STEP 3: The feature vector of M X N pixels that is default values in GIST there are 512 feature vectors among these feature set 320 features are selected as shown in figure.2. It is used for further classification using feed forward back propagation method we can effectively classify malware according to which sample belongs to particular family.

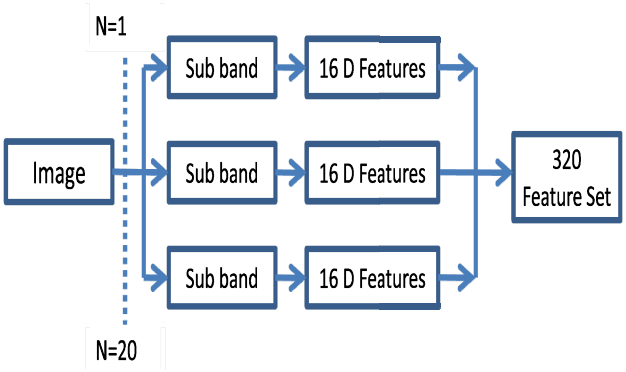

Fig. 2. Compute Texture Feature on Imag

A. Texture Feature

The texture is a repeated pattern or blocks in an image which may contain different texture information to identify high level intensity values based on that classification is done. The texture is used in medical image analysis example bone calcification and tumor identification, etc. The examples of various malware variants are shown in figure.3. (1) Adult browser (2) Allapale (3) Bancos (4) Dorfdo (5) Casino and so on as listed in Table I. Texture plays an important role in any research in image processing area. The Texel is look like any of the elements of art that can enhance and support the research work. In image processing area textures are one way that can be used to help in image segmentation and classification of images. The analysis of an image texture there are two ways. The statistical approach for texture image is quantitative measure of intensities.

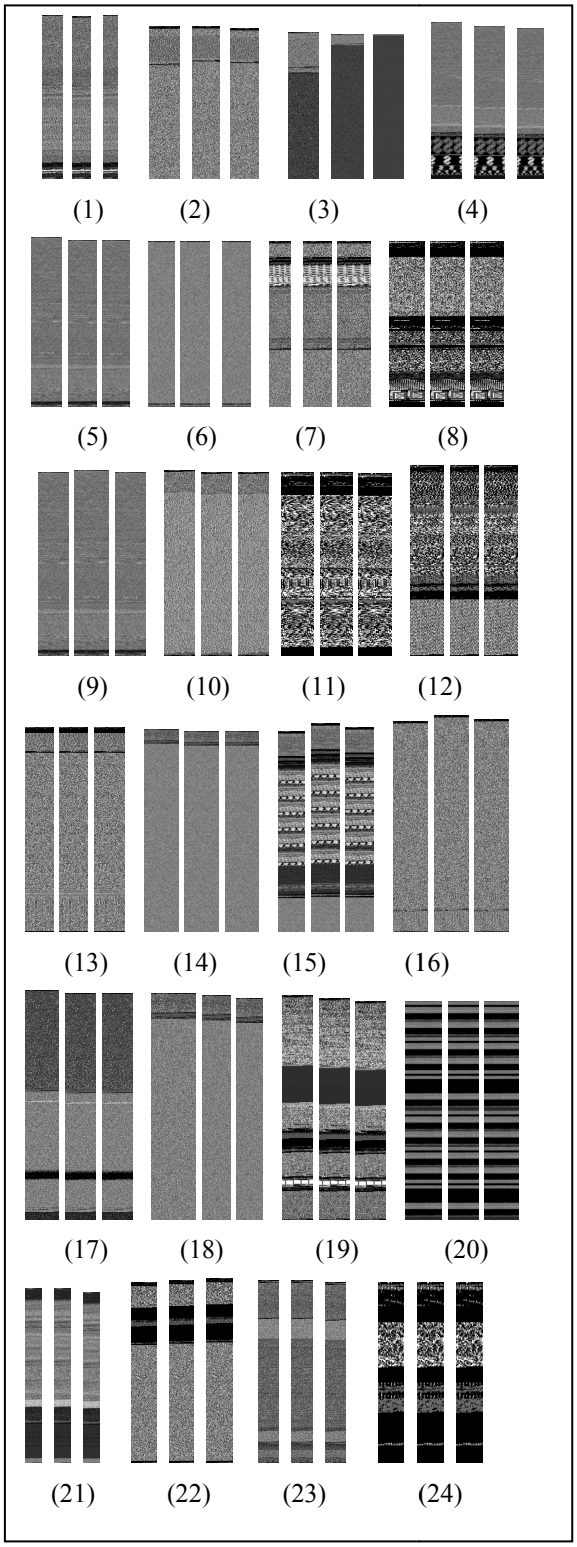

Fig. 3. Variants of Malware

B. Feed Forward Back Propagation Method

Neural networks used for remarkable ability to derive the meaning for complicated data can used to extract patterns that are too complicated to understand for human and computer techniques in this trained neural network tool in that information is given for train the data and to analyze.

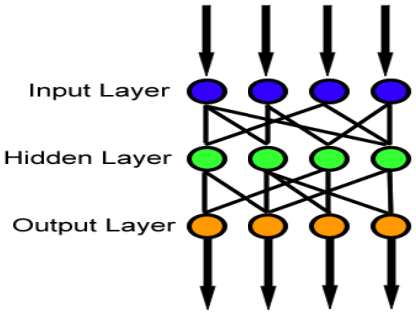

Fig. 4. Feed Forward Approach of ANN

The block diagram of feed forward back propagation method is shown in figure. 4. This allows traveling pixels in one way through input to output there is no feedback loops. That illustrates the output of one layer is does not affect same layer these methods are more used in pattern recognition. The applications of neural networks are pattern recognition, medical image analysis, and biomedical, content based image retrieval so on. The advantage of neural networks are easy to implement and flexible. It can be handle more complex iterations for training dataset. The dataset is easily trained using some decision on feature vector and learning rules of neural networks.

The feed forward network is a static mapping between its input and output spaces. In feed forward method the effective features are extracted by using Gabor wavelet and GIST descriptor total 512 features are extracted and formed feature vector. Among 512 feature set we are considering only 320 features as an input vector and corresponding target vectors are used to train the neural network to classify input feature vector in an appropriate way. The standard database is used for this research work which consists of 3131 samples of all malicious affected executable files and it contains 24 different families. For training the malware image in feed forward method of neural network. From dataset each family of malware we are considering 20 samples i.e. 20 (samples from each family) X 24=480 samples of malware are trained by using neural network and further these trained feature vectors are compared with other samples in database these trained feature vectors are used for classification.

V. EXPERIMENTAL RESULT

The proposed algorithm of malware identification of malware behavior and classify malware effectively by using feed forward back propagation method. The proposed feature extraction method is used to get 320 features set for classification of malware. Neural network used to train 666 malware images and for testing total 3131 images are used. Results are compared with previous techniques used to detect and classify new malware this method classifies malware more effectively with high accuracy 96.35%. The confusion matrix is drawn based on the variant which does not belong to any category that is confusion among variants is shown in figure. 7 The computing texture features are shown in figure.6 which is scattered the features of malware.

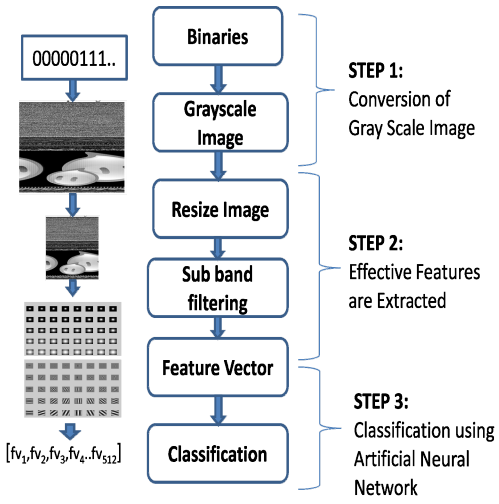

Fig. 5. Flow of Algorithm

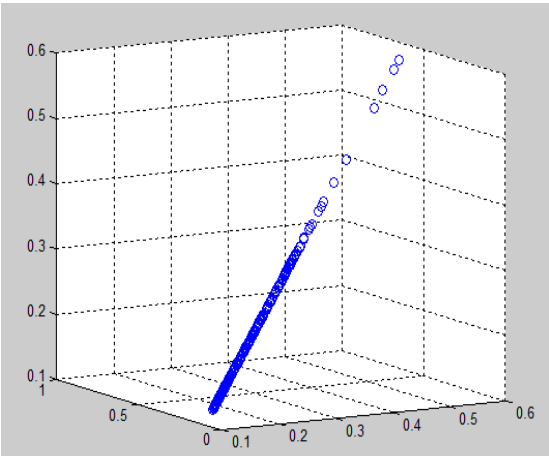

Fig. 6. Computing Texture Feature Plot

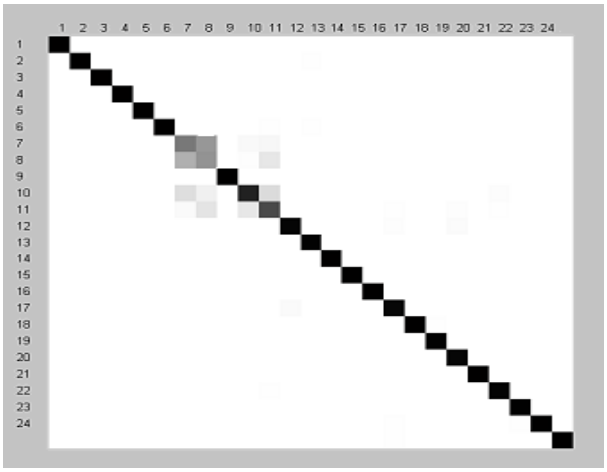

Fig. 7. Confusion Matrix of Classification

TABLE III. Experimental Results of malware classification

Classification Method	Malware Dataset			
	Samples	*TPR*	*TFR*	*Accuracy*
ANN	3131	3017	114	96.35%

VI. CONCLUSION AND FUTURE WORK

Classifying malware is an important and research problem in digital forensics. The proposed work is reopresented as 2Dimentional array, which is visualized as grayscale image that is used to deal with the problem of malware classification. The grayscale image is obtained and GIST is used. The experimental resuts of classification using feed forward back propagation neural network gives 96.35% accuracy. The database which we are using is mahuer database which is consists of 3131 malware samples with 24 different malware families.

This paper proposed a malware variants of different families classification method using visualzed images and Gabor+Gist decriptor is used to extract effective texture features. The contributions of the paper are the fallowing:

- We proposed a new method for efective texture features from malware grayscale image.
- We also proposed a visualization of malware as a gray scale image.
- Experimental results are showed that our proposed method can classify malware variants with true positive rate and true false positive rate.

In future work we concentrate on more acuurate feature set that can be used to classify malware variants more accurately. The machine learning techniques, clustering techniques and Wavelet tranform can be used to reduce the dimentions of feature vector. Also we concentrate on packed and unpacked malicious data which is used to harm the system.

ACKNOWLEDGMENT

This research work is founded by UGC under Rajiv Gandhi National Fellowship (RGNF) UGC Letter No: F1-17.1/2014-15/RGNF-2014-15-SC-KAR-69608, February, 2015.

REFERENCES

[1] Malware- Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/Malware

[2] M. Labs. McAfee threats report: Second quarter 2013. Technical report, McAfee, 2013

[3] Symantec Global Internet Security Threat Report, April 2010.

[4] Syed Zainudeen Mohd Shaid, Mohd Aizaini Maarof, "Malware Behavior Image for Malware Variant Identification," IEEE, International Symposium on Biometric and Security Technologies (ISBAST), 2014.

[5] Kyoung Soo Han, Jae Hyun Lim, Boojoong Kang, and Eul Gyu Im, "Malware Analysis Using Entropy Graphs," Springer-Verlag Berlin Heidelberg, International Journal of Information Security. 14:1-14, DOI: 10.1007/s10207-014-0242-0, 2015.

[6] Infographic: The State of Malware, 2013

[7] Natraj. L, Yegneswaran.V, Porras.P and Zhang. J, "A Comparative Assessment of Malware Classification Using Binary Texture Analysis," Procedings of the 4th ACM Workshop on Security and Artificial Intelligence, pp.21-30, 2011.

[8] Nataraj, L. " SigMal: A Static Signal Processing Based Malware Triage,". 2013

[9] Kong, D. and Yan, G. Discriminant. "Malware Distance Learning on Structural Information for Automated Malware Classification," Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems, pp. 347-348, 2013.

[10] Tantan Xu, 2014,"A file fragment classification method based on grayscale image, "Journal of computers, vol. 9, No. 8, August 2014.

[11] G.Conti, S. Bratus, A. Shubina, A.Lichtenberg, R. Ragsdale, R.Perez-Alemany, B.Sangster, and A.M.Supan, " A Visual Study of Primitive Binary Fragment Types," in Black Hat USA, 2010.

[12] Shui Yu, Guofei Gu, Barnawi, Song Guo, Stojmenovic, "Malware Propagation in Large Scale Network,"Knowledge and Data Engineering, IEEE Transaction on, 2014, 27(1):170-179.

[13] Acar Tamersoy, Kevin Roundy, Duen Horng Chau, Guilt by Association: Large Scale Malware Detection by Mining File-relation Graphs. In Proceedings of KDD'14, August 24-27, 2014, New York, NY,USA, Pages:1524-1533.

[14] Cesare, S.,Xiang, Y."A fast flowgrah based classification based classification system for packed and polymorphic malware on the end host," Advanced Information Networking and Applications (AINA), 2010, 24th IEEE International Conference on 2010, pp.721-728. IEEE.

[15] Chowdhury, G. "Introduction Modern Information Retrieval Facet publishing 2010.

[16] Shang,S.,Zheng,N.,Xu,J.,Xu,M.,Zhang, H."Detecting malware variants via function-call graph similarity,",Malicious and Unwanted Software (MALWARE), 2010, 5th International Conference on 2010,pp.113-120.IEEE

[17] Abou-Assaleh,T.,Cercone,N.,Keselj, V.,Sweidan, R."Detection of new malicious code using n-grams signatures," Proceedings of Second Annual Conference on Privacy, Security and Trust, 2004,pp.193-196.

[18] Santos,I.,Brezo,F.Nieyes,J.,Penya,Y.K,Sanz,B.Laorden,C.,Bringas, P.G."Opcode-sequence-based malware detection,".Engineering Secure Software and Systems,pp.35-43.Springer Berlin (2010)

[19] Egele,M.,Kruegel, C.,Kirda, E.,Yin,H.,Song, D."Dynamic spyware analysis," Usenix Annual Technical Conference 2007

[20] Miao, Q.-G., Wang, Y., Cao, Y., Zhang, X.-G., Liu, Z.-L."API Capture a tool for monitoring the behavior of malware," Advanced Computer Theory and Engineering (ICACTE),2010, 3rd International conference on 2010,pp.V4-390-V394-394. IEEE.

[21] Aziz Makandar and Anita Patrot," Overview of Malware Analysis and Detection," International Journal of Computer Applications (0975-8887) National Conference on Knowledge, Innovation in Technology and Engineering (NCKITE 2015).

[22] A. Torralba, K. Murphy, W. Freeman, and M. Rubin.Context-based vision systems for place and object recognition. In Proceedings of ICCV, 2003.

[23] A. Olivia and A. Torralba. "Modeling the shape of a scene: a holistic representation of the spatial envelope,". Intl. Journal of Computer Vision, 42(3):145–175, 2001.

[24] GISTCode. http://people.csail.mit.edu/torralba/code/spatialenvelope.

[25] Z. Wen, Y.Hu and W.Zhu, "Research on Feature Extraction of Halftone Image," Journal of Software, vol. 10, pp.2575-2580, 2013.

[26] Y. Lan, Y.Zhang and H.Ren, "A Combinational K-View Based Algorithm for Texture Classification," Jjornal of Software, vol. 8,pp.218-227, 2013.

[27] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International Journal of Computer Vision, vol.42,pp.145-175,2001.

[28] Tantan Xu. "A File Fragment Classification Method Based on Grayscale Image," Journal of Computers, Vol.9, No.8,August 2014.

[29] Syed Zainuden Mohd Shaid, Mohd Aizaini Maarof. "Malware Behaviour Visualization," Journal Tknologi, Eissn 2180-3722,pp.25-33, 2014.

[30] Kyoung Soo Han. Jae Hyun Lim. Boojoong Kang. Eul Gyu Im. , "Malware analysis using visualized images and entropy graphs," International Journal of Information Security, DOI:10.1007/s10207-014-0242-0, 2015.

[31] Jackson Akpojaro, Princewill Aigbe, Ugochukwu Onwudebelu. "Unsupervised Machine Learning Techniques for Detecting Malware Applications in Wireless Devices," TMLAI Transaction on learning and artificial intelligence. Socity for science and eductaion United Kingdom, vol.2, Issue.3,ISSN:2054-7390, DOI:10.14738/tmlai.23.206, July 2014.

[32] Ekta Gandotra, Divya Bansal, Sanjeev Sofat,"Malware Analysis and Classification:A Survey," Journal of Information Security,2014,Vol.5,pp.56-64,April 2014.

[33] Aziz Makandar and Anita Patrot," Malware Image Analysis and Classification using Support Vctor Machine," International Journal of Advanced Trends in Computer Science and Engineering impact factor (IJATCSE 2015), October 27. Vol.04, No.05, pp.01-03(2015) Special issue of ICETEM 2015 on 27 October.