Texture Based Malware Pattern Identification and Classification

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

Anita Patrot
Department of Computer Science
Karnataka State Women's University
Vijayapura, Karnataka, India
anitapatrot@gmail.com

Abstract**—**Malware texture pattern plays an essential role in defense against malicious instructions which were analyzed by malware analyst. It is identified as a security threat. Classifying malware samples based on static analysis which is a challenging task. This paper introduces an approach to classify malware variants as a gray scale image based on texture features such as different patterns of malware samples. Malicious samples are classified through the machine learning techniques. The proposed method experimented on malware dataset which is consisting of large number of malware samples. The similarities are calculated by texture analysis methods with Euclidian distance for various variants of malware families. The available samples are named by the Antivirus companies which can analyze through supervised learning techniques. The experimental results show that the effective identification of malware texture pattern through the image processing which gives better accuracy results compared to existing work.

Keywords*-Malware analysis; Classification; Gabor wavelet; Malware; Machine learning techniques;*

*****

I. Introduction

Malware different texture patterns motivates for classification. Malware texture pattern classification is most significant and complicated difficulty in digital forensics. In which number of unique malware variants released every year. The malware is a major security threat on internet. Malicious instructions are divided into various types of malware, which includes Virus, Worms and Trojan these together called malware. Malware attack is high incidence in network today because of increased significantly in the recent years [1][2]. The traditional approach to identify real time malware detection has relied on using signatures of specific byte sequences of API call and string pattern matching.

It affects computer system without any authority. Number of new malware variants on the internet has been continuously increasing with the help of various tools. The most of antivirus programs use character strings and patterns to detect signatures of malware [3]. It is estimated that more than million unique variants of malware are released per day. Analyzing more number of malware variants every day is an exigent task for malware analyst. Manually identifying and classifying malware samples is something which is inevitable due to growing number of malware variants.

Malware variant identification is done by using machine learning techniques. The analysis of malware is classified as static and dynamic analysis. The static analysis is done on global features of malware image. The dynamic analysis is done on sequence bytes methods, instruction frequency based techniques, and API calls are used for feature extraction. The similarity between malware variants and global behavior based methods has been proposed to detect and classify malwares. Recently, several visualization techniques have been proposed to compensate malware analysis.

II. Related Work

There are four major visualization techniques which are used for malware behavior identification, detection and classification. Such as Malware Tree Map, Malware Thread Graph, and Malware Image. The systematic brief introduction and categorization of malware visualization systems and they are identified and evaluated data providers and commercial tools used meaning full data for review malware visualization system [4] the visualization helps to understand the malicious data which are currently under represented this allows new research opportunities in the visualization system.

The objective of visualization system to compare the malware analysis system and its categories this is based on the two criteria one is feature based and image based approach this helps to understand the difference in characteristics of both approach. The novel method introduced by [5] using global features of malware visualization and texture features for malware classification based on binary texture analysis [6] to extract effective texture features from the 2D gray scale malware images to use for classification.

The advantages of this visualization technique are based on image approach. This technique can apply any file whether it is packed or unpacked that can be computed efficiently which is important for large malware dataset. This technique uses only static analysis that's why it is limited because it does not use dynamic analysis technique. The combined the features characteristics which is extracted by Hidden Markov Models and Simple Substitution Distance then by using SVMs (Support Vector Machine) they analyzed by employing morphing strategies that causes to fail, because of that combining scores are used using support vector machine yield more significantly robust [7].

The author work on the file fragment of affected part in the file which is represented in gray scale image with different extensions is used for classification [8]. The technique involved entirely on dead code insertion still its challenge presented by metamorphic malware [10]. The features extracted from the content and structure of malicious websites and web pages, which could be used by web security threat. The features are builds based on predictor and five machine learning techniques which are applied to classify known and unknown web pages and applications, these features are able to classify malicious websites.

They classified a fragment in terms of two models file unbiased and type unbiased. The affected fragments by malicious data only that part is treated as a gray scale image which gives more information related to malware [11][12]. The classification done on fragments they provide preliminary solution for automatic classification. The malware variants identification and classification is done using several data mining concept and machine learning techniques in various researches in different fields. Traditional way to detect malicious data is a long process for identifying malicious and non malicious [13]. The overview of malware analysis and detection is described in the [14].

GIST method is used for global descriptor of effective feature which is also used for scene classification as well as iris identification and handwritten OCR [15]. The most visualized techniques are used for malware global behaviors in [16]. The pre-processing of images can be done by using computation process of restoration [17]. Malware pattern analysis and differentiate in texture of individual malware family is analyzed by global features of image using image processing techniques and classification done by Support vector machine which gives better detection rate [18].

Many data mining techniques are used for most effective classification. In data mining and machine learning techniques are introduced a field Antivirus and digital forensics [19]. There are several methods of detection of malware and classification in this recently including graph based detection of malware [20].The instruction sequence based classification of malicious fragments [21]. Application Programming Interface calls are used for sequence based classification [22]. The analyzing and identify uniqueness in malware contents.

III. Proposed Work

The proposed work is the static analysis of gray scale image of malware as shown in Fig.1 is resized then extracted texture feature descriptor by using Gabor wavelet with GIST is computed on with 4 scales and 8 orientations that produces 32 features of same size of image as shown in the Fig.2. The feature space is divided into 16 regions by 4X4 grid then the average values within each region. The 16 average values of 32 feature resultant 512 as expressed as in (1), (i.e. 16x32=512) GIST descriptor. The texture descriptors summarize the gradient information for different parts of an image by using sub block average m(x) as in (2) taken from the previous paper [23][24] in that the neural network is used. Then feature vector is used for classification using machine learning technique.

Let I is an image and L is a length of the image in width and height represented as I^L = I (x,y) ,where x and y are the number of rows and number of columns of the image.

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

W is a window where decomposing image by applying wavelet filters then reconstructing image. The resultant image is dividing into n number of blocks later taking averaging of each block built a feature vector. X^1 is a sub block average within the window w and m is a maximum average is retrieved from the image.

$$\mathrm{Max}(x) = \sum I^L(x')\, w(x' - x) \qquad (2)$$

A. *Malware Image*

The affected executable files are converted into binary files then these binaries are treated as 8 bit as a pixel the range of gray scale image is 0-255. The image size is depends on file size, although the file size changes. The overall structure is visible from the images. The structure of malware gray scale image consist of various components of executable file such as code, zero padding, ASCII text, uninitialized data and initialized data. Compare byte sequence of API calls to identify malware behaviour the visualized gray scale image gives more information for analyzing malicious behaviour by visualizing malware as a gray scale image. The file size range will be between less than 10KB to more than1000KB based on this file size range the conversion of gray scale image size will be differ in image length and width from 32 pixels to 1024 pixels. The proposed work consists of global features are effectively extracted from a malware gray scale image, which is used for texture feature. In which it gives gradient information of different part of the malware image.

B. *Data Sources / Dataset*

The malicious instructions are identified by the antivirus vendors based on dynamic analysis of malware which includes system calls and API calls. It is difficult to access real malware images for experiment due to large in size. The dataset reference samples are collected from Microsoft and based on the reference of naming system the different patterns are identified and gives the malware family name.

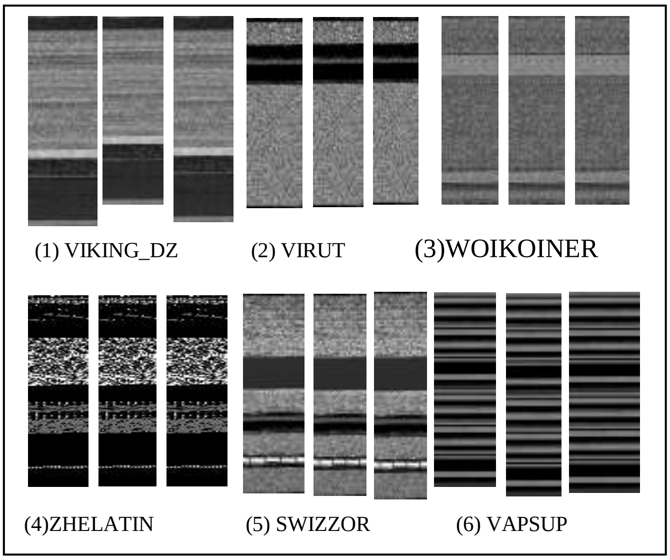
It consists of 3131 gray scale images which belong to the total 24 different malware family. These samples are categorized based on the textural behaviour of images. All gray scale images are digitized at a resolution of 1024X1024 pixels and 8 bit gray level. The proposed work is built based on wavelet analysis on Mahler reference dataset and by applying machine learning techniques. Wavelet analysis is probably the most recent solution to overcome the shortcomings of the Fourier transform, orthogonal wavelets and Bi-orthogonal wavelets, Two functions ‘f’ and ‘g’ are said to be orthogonal to each other if their inner product is zero as shown in eq.(1). The symbol * mean a convolution operation. Bi-orthogonal wavelet is a wavelet where the associated wavelet transform is invertible but not necessarily orthogonal

$$< f (t), g (t) > = f (t)\, g * (t)\, dt = 0 \qquad (3)$$

The two dimensions wavelets transform consist of two scaling functions, the horizontal measures variations along columns (horizontal edges), vertical responds to variations along rows (like vertical edges) and variations along diagonals. F (m, n) is a approximation of image, sub band filtering is applied row wise and column wise which gives the vertical and horizontal pixel intensity information, the diagonal information. Gabor wavelets are most used for texture based feature extraction in image analysis and image processing. Wavelet transforms are used tremendously in many image processing applications. The Gabor wavelet is used with 4 scales and 8 orientations. The degrees of directions are applied on the image and getting magnitude and orientations from various degrees.

IV. METHODOLOGY

The proposed system consist of three stages such as pre-processing, feature extraction and classification of malware. The pre-processing technique is used to enhance the original images and it is difficult to interpret large size malware grayscale images for computation. This technique is more reliable before feature extraction.


(1) VIKING_DZ (2) VIRUT (3)WOIKOINER

(4)ZHELATIN (5) SWIZZOR (6) VAPSUP

Figure 1. Malware Gray Scale Images

A. Pre-processing

The pre-processing stage original image is digitized and normalization done then histogram is applied to get information about the pixel intensities and gray level. This technique is first stage of research methodology because it is helpful for further feature extraction technique.

B. Feature Extraction

The feature extraction stage we are applying wavelet for more accurate features are extracted for analyzing and extraction of effective texture features for classification. Features are extracted from wavelet decomposition is done on original images, in which we get four decomposition horizontal, vertical, and diagonal information from the image based on scale N. The coefficients vectors are normalized after feature extracted. The energy is computed by squaring each element from the coefficient vector. The energy of each vector is considered for feature vector for classification.

C. Classification

The researchers need a quick and easy analysis of malware variants especially on behavioral aspects of the malware. The proposed method introduce a malware behavior in the form of malware gray scale image analysis by using visualization technique and gradient features extracted by GIST, which is already used for scene classification of natural images. This is used in malware image classification, and the features are extracted by using Gabor Wavelet. We can also choose ‘N’ number of scales and orientations to extract gradient features.

The comparison of machine learning techniques are shown in experimental results in the form of True Positive and True Negative Rates can effectively identify and classify malware variants. Malware Visualization is a technique is used to represent malware binary samples into particular pixels in static analysis. The image processing area several classification techniques are available to classify images based on their effective features. These global features are used for further classification to analyze and identify different patterns of variants of malicious data, for detection and classification. The increasing use of machine learning techniques for various applications such as OCR (Optical character recognition), Iris identification for security ,medical image analysis, human identification, face recognition, optical character recognition, and malware detection and classification. The machine learning techniques are Support Vector Machine (SVM) classifier and Artificial Neural Network (ANN) Classifier.

The texture features are extracted by applying Gabor wavelet which gives gradient features of texture of different parts of the malware image. Feature vector is formed with 320 dimensional vectors from 3131 malware dataset which contains the different malware variants in 24 malware families of dataset. The experimental results are shown in Fig.3, which illustrates the comparison results of three classifications of machine learning techniques. The efficiency of an algorithm is calculated using accuracy and Error. The classification of malware samples where accuracy performance is calculated by using the true positive rate and false negative rate. Table.1 Illustrate the total accuracy rate and detection rate in Fig.5 of samples individual family.

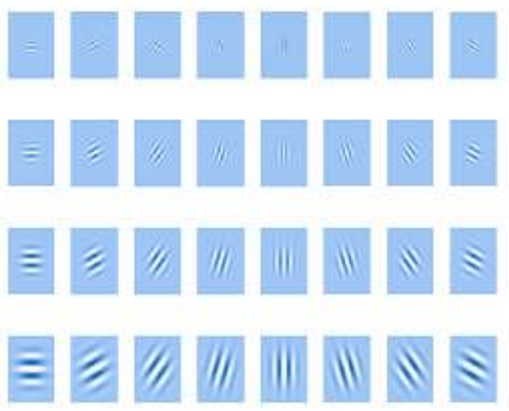


Figure 2. Gabor Wavelet with 8 Orientation and 4 Scales

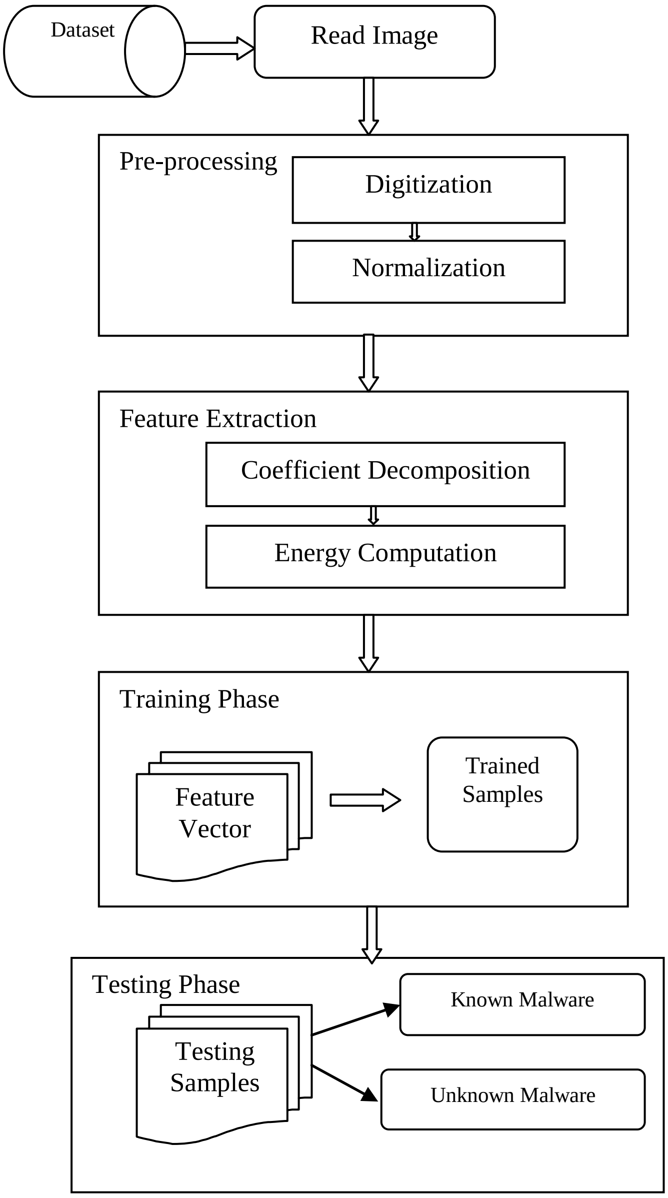


Figure 3. Block Diagram of Classification

V. Experimental Results

Let X_(n)ϵ R^l represent a malware sample, where l is a length of the malware, we assume that in order to represent malware as a digital grayscale image f(x,y) of dimensions, where x is a number of columns and y is a number of rows l_x X l_y , where x and y also represent width and height of the image. Image is resized to a standard size and that can be parameterized as $R_f = (R_{fx},R_{fy})$, where R_{fx},R_{fy} are the resizing factors in image horizontal and vertical direction. Let N be the number of filters used to filter the image such as Wavelets and Gabor Filters. The True positive rate of the correctly classified malware samples are illustrated in Fig.6 and detection rate in Fig.7

TABLE I. MALWARE VARIANTS WITH FAMILY

Malware Dataset			
Malware Family	***TPR (True Positive rate)***	***FPR (False Positive rate)***	***Accuracy of detection***
ADULTBROWSER	262	0	1.00
ALLAPLE	282	4	0.96
BANCOS	35	0	1.00
CASINO	140	0	1.00
DORFDO	65	0	1.00
EJIK	168	2	0.98
FLYSTUDIO	32	0	1.00
LDPINCH	43	0	1.00
LOOPER	190	0	1.00
MAGICCASINO	174	0	1.00
PODNUHA	299	16	0.86
POISON	26	0	1.00
PORNDIALER	97	0	1.00
RBOT	92	0	1.00
ROTATOR	286	0	1.00
SALITY	42	0	1.00
SPYGAMES	121	3	0.97
SWIZZOR	31	1	0.99
VAPSUP	45	4	0.96
VIKING_DLL	72	0	1.00
VIKKING_DZ	28	0	1.00
VIRUT	202	0	1.00
WOIKOINER	50	0	1.00
ZHELATIN	26	0	1.00

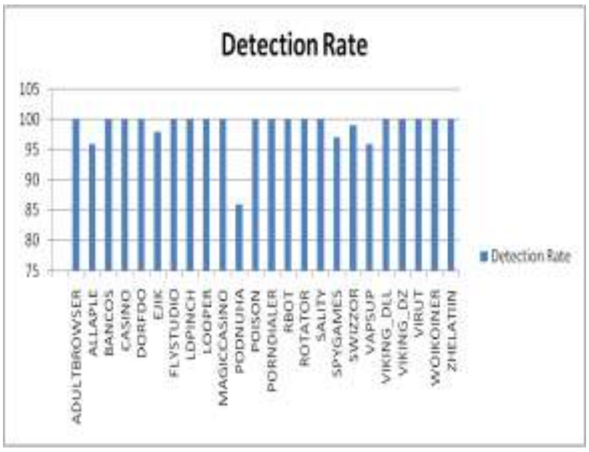


Figure 4. Detection rate of individual malware family

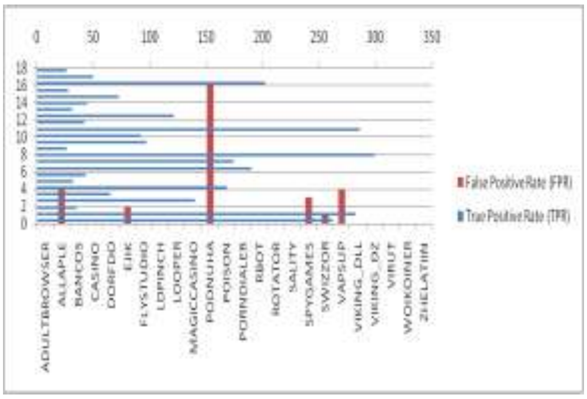


Figure 5. Identified correct malware samples and their family.

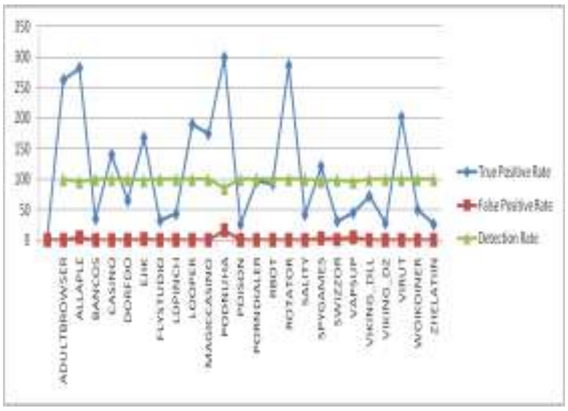


Figure 6. Results of malware samples similarities and classification with Cross Validation of TPR and FPR with detection Rate.

VI. CONCLUSION

This paper proposed a malware texture pattern classification using machine learning techniques. The contributions of the paper are analyzing large number of malware samples by using image processing techniques. To calculate similar features from the malware gray scale image by applying wavelet transform with db4 wavelet decomposition method. After construction of feature vector, the classification is done on malware based on machine learning techniques such as Artificial Neural Network and Support Vector Machine classifiers. The contributions of the paper are as fallows

- ✓ Analyzing large dataset of malware and visualized malware as a gray scale image.
- ✓ The malware samples are analyses through image processing techniques based on the existing system we are getting better results for classification of malware samples.
- ✓ Where we are getting results in the form of True Positive Rate and False Positive Rate.
- ✓ The experimental results show the better accuracy compared with existing work.

The existing techniques for classification either require disassembly or execution whereas this method does not require disassembly but still show significant improvement in accuracy. This proposed technique should be very valuable for anti-virus companies and security.

ACKNOWLEDGMENT

This research work is funded 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] M. Labs. McAfee threats report: Second quarter (2015). Technical report, McAfee.

[2] Symantec Global Internet Security Threat Report, 2015.

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

[4] M. Wagner, F. Fischer, R. Luh, A. Haberson, A. Rind, D. A. Keim, and W. Aigner, "A Survey of Visualization Systems for Malware Analysis," Eurographics Conference on Visualization (EuroVis) (2015),Springer.

[5] Nataraj L., Karthikeyan S., Jacob G., Manjunath B. S, "Malware images: Visualization and automatic classification," In Proc. 8th Int. Symp. Visualization for Cyber Security, VizSec (2011), ACM, pp. 4:1–4:7. doi:10.1145/ 2016904.2016908.

[6] Nataraj L., Yegneswaran V., Porras P.,Zhang J., " A comparative assessment of malware classification using binary texture analysis and dynamic analysis," In Proc. 4th ACM Workshop on Security and Artificial Intelligence, AISec (2011), pp. 21–30. doi:10.1145/2046684.2046689.

[7] Tanuvir Singh, Fabio Di Troia ,Visaggio Aaron Corrado, Thomas H. Austin.Mark Stamp1 2015, "Support vector machines and malware detection," J Comput Virol Hack Tech,DOI 10.1007/s11416-015-0252-0,2015.

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

[9] 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. 2015, 14:1-14, DOI: 10.1007/s10207-014-0242-0.

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

[11] 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, 2013, pp. 347-348.

[12] 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, 2013, pp. 347-348.

[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, New York, NY, USA, 2014, pp: 1524-1533.

[14] Aziz Makandar and Anita Patrot. Article: Overview of Malware Analysis and Detection. IJCA Proceedings on National Conference on Knowledge, Innovation in Technology and Engineering (NCKITE 2015) NCKITE 2015(1):35-40, July 2015.

[15] Z. Wen, Y.Hu and W.Zhu. (2013), “Research on Feature Extraction of Halftone Image,” Journal of Software, vol. 10, pp.2575-2580.

[16] Y. Lan, Y.Zhang and H.Ren.(2013),” A Combinational K-View Based Algorithm for Texture Classification.,”Jornal of Software, vol. 8, pp.218-227.

[17] Aziz Makandar and Anita Patrot. Article: Computation Pre-processing Techniques for Image Restoration. International Journal of Computer Applications 113(4):11-17, March 2015.

[18] Aziz Makandar and Anita 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.ISBN 978-93-5212-748-1.

http://www.warse.org/IJATCSE/static/pdf/Issue/icetem2015sp01.pdf

[19] Smita Navalli, Vijaylaxmi, Manoj Singh and Vinod.P, “An efficient block-discriminant identification of packed malware,” Sa.dhana.Vol. 40, Part 5, August 2015, pp. 1435–1456.

[20] Stavros D. Nikolopoulos Iosif Polenakis,” A graph-based model for malware detection and classification using system-call groups,” J Comput Virol Hack Tech DOI 10.1007/s11416-016-0267-1.

[21] Aziz Makandar, Anita Patrot and Bhagirathi Halalli, “Color Image Analysis and Contrast Stretching using Histogram Equalization,” International Journal of Advanced Information Science and Technology (IJAIST) ISSN 2319:2682, Vol.27, No.27, July 2014,pp.119-125.

[22] Aziz Makandar and Anita Patrot, “Texture Feature Extraction of Malware Gray scale image by using M-band Wavelet,” International Conference on Communication Networks and Signal Processing (ICCNSP 2015),Bangalore, Published by McGraHill publication ISBN(13) 978-93-85880-73-5.

[23] Aziz Makandar and Anita Patrot, “Malware Analysis and Classification using Artificial Neural Network,” IEEE International Conference on Automation, Communication and Computing Technologies (ITACT 2015), 21 &22 December 2015, Bangalore.pp.46-51.

[24] Aziz Makandar, N.B. Mokashi, D. Jahagiradar and A.A. Patrot, “Performance and Analysis of Intelligence and Neural Networks in IMT,” International Journal of Research in Computer Science and Information Technology (IJRCSIT), 2014, Vol.2, Issue.2,pp. 169-176.