ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
June 2017,
Vol. 10, No. (2):
Pgs. 400-406

Wavelet Statistical Feature Based Malware Class Recognition and Classification using Supervised Learning Classifier

AZIZ MAKANDAR and ANITA PATROT*

Department of Computer Science, Akkamahadevi Women's University, Vijayapura, India.
*Corresponding author E-mail: patrotanita@gmail.com

http://dx.doi.org/10.13005/ojcst/10.02.20

(Received: April 16, 2017; Accepted: May 04, 2017)

ABSTRACT

Malware is a malicious instructions which may harm to the unauthorized private access through internet. The types of malware are incresing day to day life, it is a challenging task for the antivius vendors to predict and caught on access time. This paper aims to design an automated analysis system for malware classes based on the features extracted by Discrete Wavelet Transformation (DWT) and then by applying four level decomposition of malware. The proposed system works in three stages, pre-processing, feature extraction and classification. In preprocessing, input image is normalized in to 256x256 by applying wavelet we are denoising the image which helps to enhance the image. In feature extraction, DWT is used to decompose image into four level. For classification the support vector machine (SVM) classifiers are used to discriminate the malware classes with statistical features extracted from level 4 decomposition of DWT such as Daubechies (db4), Coiflet (coif5) and Bi-orthogonal (bior 2.8). Among these wavelet features the db4 features effectively classify the malware class type with high accuracy 91.05% and 92.53% respectively on both dataset. The analysis of proposed method conducted on two dataset and the results are promising.

Keywords: Classification, Discrete Wavelet Transform, Feature Extraction, Malware Class, Texture and Pattern**.**

INTRODUCTION

The analysis of texons played a major role in classification the pattern classification techniques and applications in the areas of image processing are growing increasingly. The image processing and pattern classification represents the state of art developments in the field. Texture pattern recognition is the task of classify input feature vector data in to classes based on the selected features from the vector. There are two types of classification supervised classification and unsupervised classification. The pattern recognition has applications in computer vision, SAR image classification, and speech classification and texture classification. The texture classification plays a major role in many applications Such as medical image analysis, pattern classification and so on. Supervised classification methods are used for face recognition, OCR, object detection and

classification. Unsupervised classification methods are used in finding hidden structures, segmentation and clustering. Wavelet transforms have become one of the most important and powerful tool of signal processing and representation. Now a day, it has been used in image processing, data compression and signal processing in different applications different wavelets are used. In this paper we present the overview of wavelets transformations in image processing. The objective of this paper is to give comparison results of the wavelet transform with their family.

Malware[1] is software that performs unwanted features like Virus, Worm and Trojan horse. The functionalities of a malware such as execution and infection, self replication that infect another host, privilege escalation, manipulation that damages the host and concealment that hides from detection. The visualization of malware is an image is read as binary vector of 8 bit unsigned integers that are to be organized into a 2D array. This can be visualized as a gray scale image in the range [0, 255] the width of an image is fixed and height is allowed to vary depending on the file size. 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 is. 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[2]. 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[4] 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[5, 6] 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 using SVM[7]. The file fragment used as a grayscale image[8] identification of malware. The behavior of malware[9] is analyzed the entropy based[10] effective features are used for classification with entropy graph. The distance learning techniques are used with structural information for classification done on automatic

Related Work

Texture plays a very important role in many research areas including image processing, pattern recognition, and medical image analysis also in computer vision. Texture analysis aims to finding a distinctive way of representing the primary characteristics of textures and represent them in some simpler but unique form, so that they can be used for robust, accurate classification and segmentation of objects. Through the texture statistical features plays a significant role in image analysis. Only a few architectures implement on-board textural feature extraction. Statistical texture features are formulated by using gray level of malware image. The motivation of this work is that textures of a malware images are extracted effective features that considers the spatial relationship of pixels in a level co-occurrence matrix this matrix also called as gray level spatial dependence matrix a number of texture features are extracted namely contrast, correlation, energy, mean , standard deviation, entropy, RMS and homogeneity are computed.

MATERIALS

The proposed work analyzed by using standard databases mahenhuer and malimng dataset. The datasets are consists of 24 malware family with 3131 variants of it and another dataset consists of 25 malware types. The details are listed in table 1. There are 3131 malware images and 1245 malware images of different malware families listed below.

METHODOLOGY

The proposed methodology we are applying wavelet low pass and high pass filters on malware image and extracted the effective features for classification. The classification consist of training phase and testing phase, where we are considering effective features selection for training images from the database. The following Fig1 illustrates the methodology of detection of malware variants.

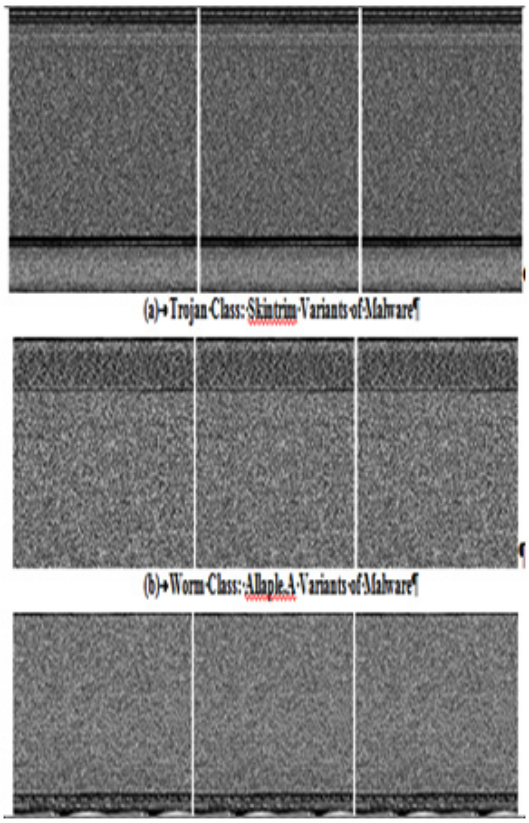

Fig. 1: Texture Similarities of Malware Classes (a) Trojan Class (b) Worm Class with different variants of Malware

Pre-processing

The first we need to prepare the dataset for testing and training data from the dataset. In this stage we are trained the dataset using dataset images, where we are collected randomly images from individual malware family samples are varied from 20 to 25 images and train the samples using the extracted feature vector to individual malware family samples total 666 images are trained from 3131 dataset. In this testing stage we are testing the complete dataset of each sample of the malware family from the dataset for SVM multiple class classifier. The pre-processing stage we are loading the image and applying common operations such as normalization, filter and sub block average. The resultant filtered image is send to the next stage.

Table 1: Details of Malware datasets

Malimng Dataset	Type of Malware Family	No. of Samples
Allaple.A	worm	2949
Yuner.A	worm	800
Lolyda.AA1	PWS	213
Lolyda.AA2	PWS	184
Lolyda.AA3	PWS	123
C2Lop.P	Trojan	146
C2Lop.gen!G	Trojan	200
Instantaccess	Dialer	431
Swizzor.gen!I	Trojan Downloader	132
Swizzor.genE	Trojan Downloader	128
VB.AT	worm	408
Fakerean	Rouge	381
Alueron.gen!J	Trojan	198
Malex.gen!J	Trojan	136
Lolyda.AT	PWS	159
Adialer.C	Dialer	125
Wintrim.BX	Trojan Downloader	97
DialplatformB	Dialer	177
Dontovo.A	Trojan Downloader	162
ObfuscatorAD	TrojanDownloader	142
Agent.FYI	Backdoor	116
Autorun.K	Worm	106
Rbot!gen	Trojan	158
Skintrim.N	Trojan	80

Statistical Feature Extraction (SFE)

The scale and translation parameters are given by,S=2-m and T=n2-m where m ,n are the subset of all integers. Thus, the family of wavelet is defined in equation 1.

$$\psi_{m,n}(t)=2^{m/2}\psi(2^m t-n) \quad ...(1)$$

The wavelet transform decomposes a signal x(t) into a family of waveletsas given in equation 2 and

$$x(t)=\sum_m\sum_n c_{m,n}\psi_{m,n}(t) \quad ...(2)$$

Where

$$c_{m,n}=\langle x(t),\psi_{m,n}(t)\rangle \quad ...(3)$$

For a discrete time signal x[n], the decomposition is given by:

$$x[n]=\sum_{i=1to d}\sum_{k\in Z}c_{i,k}g[n-2^i k]+\sum_{k\in Z}d_{I,k}h_I[n-2^I k] \quad ...(4)$$

In case of images, the DWT is applied to each dimensionality separately. The resulting image X is decomposed in first level is xA, xH,xV and xD as approximation, horizontal, Vertical and diagonal respectively. The xA component contains low frequency components and remaining contains high frequency components[29]. Hence, X= xA+{xH+xV+xD}. Then DWT applied to xA for second level , third level and fourth level decomposition. Hence the wavelet provides hierarchical framework to interpret the image information. The basis of wavelet transform that is localized on mother wavelet. The statistical feature extraction (SFE) stage we are applying wavelet filters such as Discrete Wavelet Transform then the extracted 11 statistical features are constructed a feature vector and to get normalized features for classification. The SFE features such as contrast, correlation, energy, homogeneity, mean, standard deviation, entropy, RMS, variance, smoothness, kurtosis, and skewness.

Energy: $\sum_i\sum_j p[i,j]^2$ ...(5)

Contrast: $\sum_{n=0}^{Ng-1} n^2\left\{\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} p[i,j]\right\}$ ...(6)

Correlation: $\frac{\sum_{i=1}^{Ng}\sum_{j=1}^{Ng}(i,j)p[i,j]-\mu_x\mu_y}{\sigma_x\sigma_y}$ ...(7)

Entropy: $\sum_i\sum_j p[i,j]\log(p[i,j])$ ...(8)

Supervised Classification

SVM is a supervised learning classifier that seeks an optimal hyper-plane to separate two or more classes of samples from the dataset. The mapping the input data into a higher dimensional space is done by using Kernel functions with the aim of obtaining a better distribution of the data in the form of three kernels rbf, linear and distributed. Then, an optimal separating hyper-plane will be drawn in the high-dimensional feature space can be easily found in the diagram shown below. In classification stage we are measuring the TPR (True Positive Rate) and FPR (False Positive Rate) with

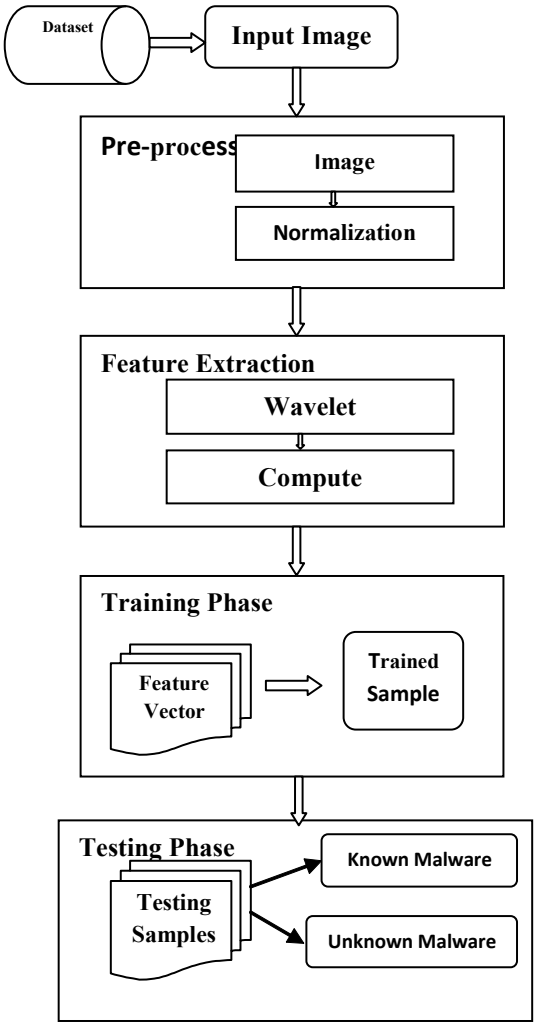

Fig. 2: Proposed Methodology of Malware Recognition

the detection rate of the malware samples. The TPR means correctly classified images and FPR means misclassified images. The accuracy of the classifier is calculated by using formula.

Accuracy= Correctly Classified Images/ Total Images ...(1)

Result Analysis Of Malware Recognition

The experimental analysis is done on the both malware dataset which consists of the 24 malware class and 9 Trojan classes. The results are analysed through the wavelets based statistical feature for malware classification and recognition. The wavelets family applied on discrete wavelet transform (DWT).

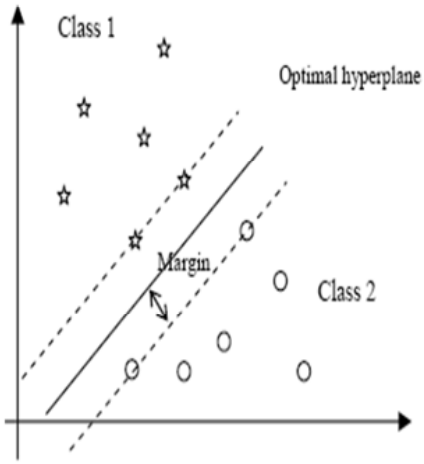

Fig. 3: Support Vector Machine Classifier

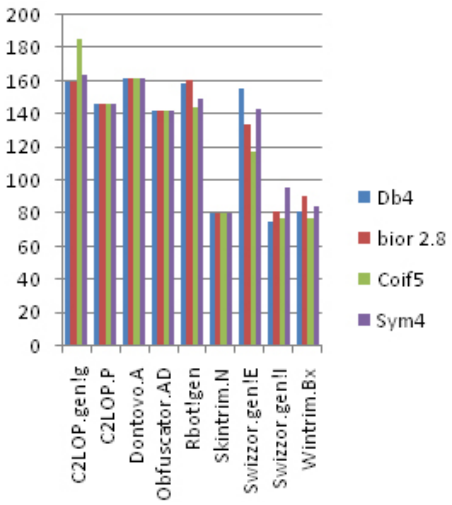

Fig. 4: Comparison of Trojan Malware Classifcation using Wavelets db4, bior 2.8,coif5,sym4

Table 2: Trojan Malware TPR for wavelet family

Trojan Malware	Db4	bior 2.8	Coif5	Sym4
C2LOP.gen!g	160	160	186	164
C2LOP.P	146	146	146	146
Dontovo.A	162	162	162	162
Obfuscator.AD	142	142	142	142
Rbot!gen	159	161	144	149
Skintrim.N	80	80	80	80
Swizzor.gen!E	156	134	118	143
Swizzor.gen!I	75	81	77	96
Wintrim.Bx	81	91	77	85

Table 3: Trojan Malware TPR for wavelet family

Malware Class	Sym4	coif5	bior 2.8	db4
ADULTBROWSER	262	262	262	262
ALLAPLE	300	299	298	300
BANCOS	48	47	48	47
CASINO	140	139	140	139
DORFDO	65	65	65	65
EJIK	167	167	168	167
FLYSTUDIO	33	17	30	33
LDPINCH	44	42	42	43
LOOPER	209	190	209	209
MAGICCASINO	177	174	174	174
PONDNUHA	300	300	300	300
POISON	26	24	22	28
PORNDIALER	103	96	97	97
RBOT	98	99	85	101
ROTATOR	300	298	299	300
SALITY	7	40	34	63
SPYGAMES	115	48	27	40
SWIZZOR	44	76	73	64
VAPSUP	0	0	12	44
VIKING_DLL	132	106	131	126
VIKING_DZ	66	64	64	64
VIRUT	135	95	123	97
WOIKOINER	50	50	50	50
ZHELATIN	41	41	38	38
Total	2293	2739	2791	2851
Accuracy	73.23%	87.48%	89.14%	91.05%

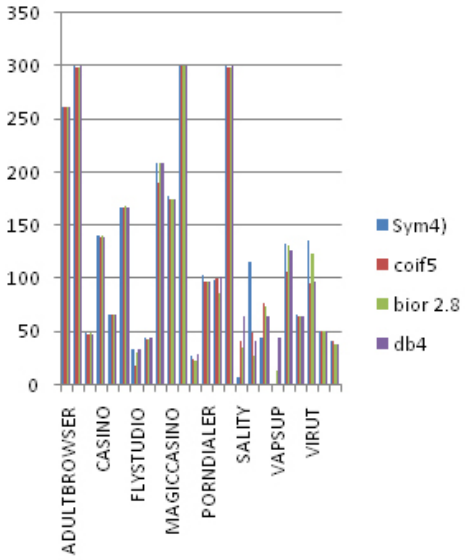

Fig. 5: Comparison of Malware on Mahenhuer dataset using Wavelets db4, bior 2.8,coif5,sym4

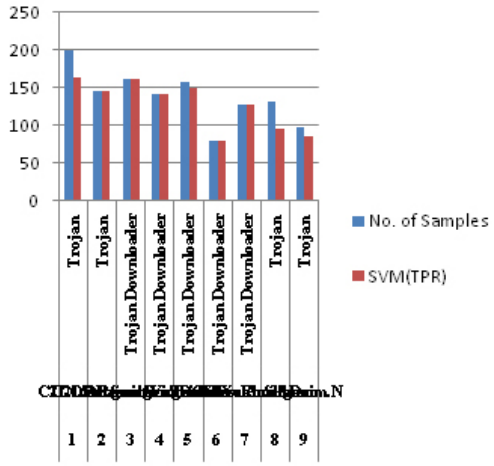

Fig. 6: Comparison of Trojan Malware Classifcation using SVM

Table 4: Comparison of TPR for malware dataset with wavelet family

Dataset	Training Data	Testing data	Method	TPR	Accuracy
Mahenhuer	666	3131	WSF+SVM	2851	91.05%
Malimng	642	1245	WSF+SVM	1152	92.53%

CONCLUSION

We proposed an efficient malware class recognition technique based on wavelet based statistical feature extraction method for texture of malware variants. In this paper we proposed an efficient Trojan malware class recognition model using image processing techniques, based on various textures of Trojan malware variants. In our work 11 statistical features are used as a feature vector to training dataset and testing dataset that can reduce the complexity by using wavelet transforms with supervised learning classifiers.

In this work we presented our result analysis in experimental shows that the proposed algorithm gives better classification results on Trojan Malware Class Recognition and malware class classification. This feature extraction method gives accurate malware class by using image processing techniques. The SVM classifier gives 92.52% accuracy. The classification error rate is very less compare to existing work on classification of malware. The contributions of this paper are as fallows. Wavelet Transform with DWT is used to extract effective wavelet based statistical features by applying wavelet transforms with wavelet family like db4, bior2.8, sym4 and coen5. Further our future work we develop model where we can classify and detect the particular Trojan malware family more accurately with genetic algorithm and adaboost techniques for classification of further research work.

ACKNOWLEDGEMENT

This research work is funded and supported 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," Euro graphics 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. 4-7.
6. Tanuvir Singh, Fabio Di Troia, Visaggio Aaron Corrado, Thomas H. Austin. Mark Stamp1 2015, "Support vector machines and malware detection," *Journal Computer Virol Hack Tech*, 2015.
7. Tantan Xu," A file fragment classification method based on gray scale image," *Journal of computers*, **9**(8), 2014.
8. 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.
9. Said Zainudeen Mohd Shaid, Mohd Aizaini Maarof, "Malware Behavior Image for Malware Variant Identification," IEEE, International Symposium on Biometric and Security Technologies (ISBAST), 2014.
10. 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 Modelling of Computer Systems, 2013, pp. 347-348.
11. 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.
12. Aziz Makandar and Anita Patrot, "Computation Pre-Processing Techniques for Image Restoration," *International Journal of Computer Applications* (0975-8887), **113**(4), 2015.
13. Z. Wen, Y.Hu and W.Zhu. (2013)," Research on Feature Extraction of Halftone Image," *Journal of Software*, **10**, pp.2575-2580.
14. Y. Lan, Y.Zhang and H.Ren.(2013), "A Combinational K-View Based Algorithm for Texture Classification," *Journal of Software*, **8**, pp.218-227.
15. Smita Navali, Vijay Laxmi, Manoj Singh Gaur and Vinod P," An efficient block-discriminate identification of packed malware," *Sadhana*. **40**(5), pp. 1435–1456, 2015.
16. Stavros D. Nikolopoulos Iosif Polenakis,"A graph-based model for malware detection and classification using system-call groups," *Journal Computer Virol Hack Tech*.
17. Z. Wen, Y.Hu and W.Zhu. (2013)," Research on Feature Extraction of Halftone Image," *Journal of Software*, **10**, pp.2575-2580.
18. Y. Lan, Y.Zhang and H.Ren. (2013), "A Combinational K-View Based Algorithm for Texture Classification," *Journal of Software*, **8**, pp.218-227.
19. Acar Tamer soy, 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.
20. Aziz Makandar and Anita Patrot,"Malware Image Analysis and Classification using Support Vector Machine," *International Journal of Trends in Computer Science and Engineering*,**4**(5), pp.01-03, 2015.http://www.warse.org/IJATCSE/static/pdf/Issue/icetem2015sp01.pdf
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), pp.35-40.
22. Aziz Makandar and Anita Patrot, "Color Image Analysis and Contrast Stretching using Histogram Equalization," *International*

Journal of Advanced Information Science and Technology (IJAIST) ISSN 2319:2682, **27**(27), ,pp.119-125, 2014.

23. Aziz Makandar and Anita Patrot," Malware Image Analysis and Classification using Support Vector Machine," *International Conference on Emerging Trends in Engineering and Management* (ICETEM 2015).
24. 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, India (December 3rd to 5th, 2015), Published by McGraHill publication.
25. Aziz Makandar and Anita Patrot," Malware Analysis and Classification using Artificial Neural Network," **IEEE Xplorer,** International Conference on Automation, Communication and Computing Technologies (ITACT 2015), December 22 and 23, Bangalore, IEEE Xplorer.
26. Aziz Makandar and Anita Patrot, "An approach to analysis of malware using Supervised Learning Classification". International Conference on Recent Trends in Engineering, Science & Technology ICRTEST 2016. 25th–27th October 2016, IET Inspec**.**
27. Aziz Makandar and Anita Patrot, "Trojan Malware Image Pattern Classification," International Conference on Cognition and Recognition, ICCR 2016,30-31,Mysore, December,2016, *Springer***.**
28. Aziz Makandar and Anita Patrot," Malware Class Recognition using Image Processing Techniques", ICDMAI 2017, 24th to 26th Feb 2017, IEEE Xplorer, Puna.