A Critique on a paper Malware Class Recognition Using Image Processing Techniques
Aziz Makandar and Anita Patrot

Sri Ravali Baru
Department of Computer Science
The University of Texas at Dallas
Dallas, Texas, USA.
SriRavali.Baru@UTDallas.edu

Abstract — Malware analysts in the current era need to enhance their procedures and methodologies because of the mutations happening to the existing traditional malware protection leading to malware attacks. The identification of malware and classification to the specific family is also important to protect the system from damage. The rapid growth and complexity of the malware instructions led to the invention of techniques such as SVM, image processing, random Forest, KNN, deep learning other than the traditional way, leading to a more efficient way of handling malware. In those techniques, image processing techniques are widely used, which in turn are detected by the texture of the image. Texture helps to identify which family of malware does it belong to. Example whether it is virus, worm, trojan horse, rogue class, etc. I am to survey the paper 'Malware Class Recognition Using Image Processing Techniques', which was presented at the 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) by author Aziz Makandar and Anita Patrot. I will provide a comprehensive view of the Support Vector Machine (SVM) from the perspective of image processing. Two data sets were used in the paper 1) Malheur dataset and 2) Malimg dataset, at the end a confusion matrix for SVM and KNN algorithms are created to check the accuracy of the used implementation of those algorithms. In addition to surveying the paper, I also cover keywords such as Gabor Wavelet; Discrete Wavelet Transform; Support Vector Machine briefly, such that it gives an easy way for the reader to understand the approach easily. Finally, this paper concludes with a discussion on advantages and challenges faced using this method, and preliminary approaches addressing some of these challenges.

I. INTRODUCTION

This is the survey of the paper 'Malware Class Recognition Using Image Processing Techniques'. The software which is intentionally designed to cause damage to a computer is called malware. It acts against user interest and may also cause unauthorized access. Different malware types exist, including computer viruses, worms, Trojan horses, ransomware, spyware, adware, rogue software, wiper, and scareware [3]. Malware attacks are increasing in recent years these were reported in [1]. The traditional way to identify them and classify malware families [2] in real time based on malicious byte sequences of API call and string pattern matching. The majority of antivirus uses character strings with different patterns to detect signatures of malware [3].

Classifying them on the global structure is difficult. The paper mentioned identifies and classifies variants based on global features. Dataset was taken from the Vision Lab, with 3131 malware samples, which are malware images.

By image processing techniques these are classified into corresponding malware families (worms, trojan horses, ransomware..). The similar texture pattern obtained is considered as one family. From the input samples, a feature vector is generated, then the classification is done on malware based on machine learning techniques. We got results in the form of True Positive Rate and False Positive Rate, which indeed gives accurate predictions and results of the methodology.

Gabor wavelet and Filter: It is a two-dimensional Gabor function that consists of a sinusoidal plane of a certain frequency and orientation. It is modulated by a Gaussian envelope. Its frequency and orientation can be selected based on requirements. By varying the frequencies and orientations, we obtain a new one every time. An image is passed through this to obtain several filtered images from which texture-based features are extracted.

Discrete Wavelet Transform: This is like Fourier transform it returns a data vector of the same length as the input given. A signal passed to it and decomposed same or lower number of the wavelet coefficient spectrum as is the number of signal data points.

Support Vector Machine: It is a supervised machine learning algorithm used for both classification and regression challenges. Each data corresponds to a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then a hyper-plane that differentiates the two classes is found.

Organization of the paper — The organization of the paper is as follows section I gives a brief introduction, section II gives the approach taken for malware classification, section III gives the advantages and disadvantages of the taken methodology, section IV summary and directions for extending this paper, finally references utilized while creating this paper.

II. WORK RELATED

Firstly, this paper gives a brief idea of visualization technique usage. It says “visualization could help to understand the malicious data. The visualization classification is used to compare the malware analysis system and its categories this is based on system call and image-based. Also, the features are extracted by Hidden Markov Models and Simple Substitution Distance then by using SVMs sometimes fail because of the metamorphic nature of malware [4]. The better approach is done by a novel method, which uses global features of malware visualization and texture patterns depending on binary texture analysis for malware classification. This technique uses only static analysis and works on packed or unpacked files and for large malware datasets. After analyzing only that specific part is treated as a grayscale image is considered as malicious data, which gives more internal information related to the existing malware and its family.[5].”

Then it explains about resources used and malware analysis “Two standard datasets of malware samples which are grayscale images are used having various texture patterns of the malware families. Namely Malheur dataset and Malimg dataset. The Malheur dataset contains of 3,131 malware samples from 24 different malware families and the Malimg dataset contains 9339 samples from 25 different malware families. The malware binaries

were labeled such that they are compatible with majorities of six different Antivirus products used in the experimental results. These samples are already in pixel format with a .png extension. Before going to do the preprocessing stage, the dataset is modified according to the malware family variants by creating a root folder, which in turn had subfolders for each malware family. Then the input data set is divided randomly for training and testing phases, 25 samples from each family are taken for training and the whole data set is applied for testing." The following fig.1 shows the similarities of malware samples and different texture patterns corresponding to the malware of each family.

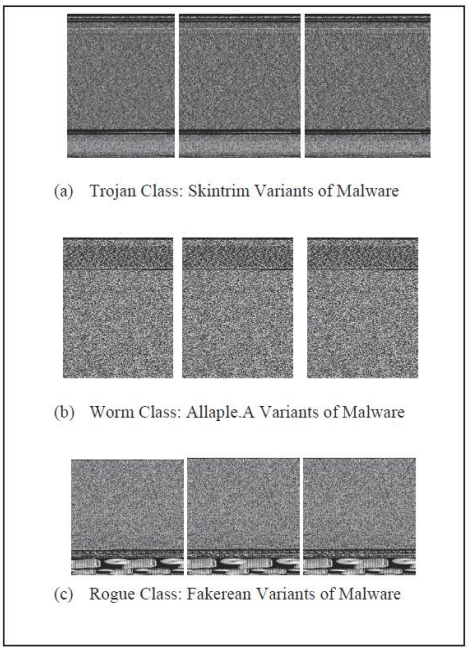

Fig.1 Texture Similarities found in malware variants from different malware family.

The Proposed algorithm is explained clearly, which is the main step for the classification of malware. Steps in the algorithm are as follows:
Step-1: Collect and prepare a dataset.
Step-2: Preprocessing of data is done to get uniformity by normalizing data.
Step-3: Feature Extraction by passing through discrete wavelet transforms.
Step-4: Feature Selection by applying mean and with PCA.
Step-5: Train the obtained data.
Step-6: Test the data and classify it accordingly.

Normalization of 64X64 malware samples and applying a filter for malware is done in the first and second steps. The third step involves passing the normalized sample into discrete wavelet transform with three-level decomposition using the db4 wavelet family which gives effective energy coefficients. In the fourth step, four-direction energy coefficients are extracted then the mean of the feature vector is applied. The final step has the training and testing phase of this feature vector, which in turn gives clear classification, where each test sample belongs to. The following fig.2 shows the proposed methodology.

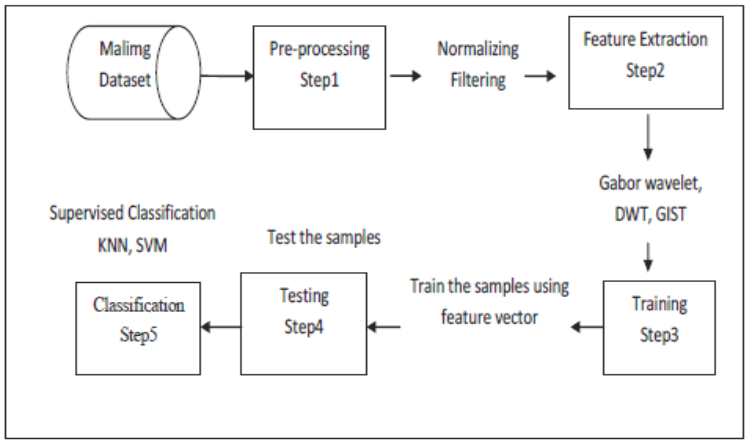

Fig 2 Flow of the algorithm mentioned.

Experimental result analysis and conclusion are presented at the end. For the result analysis, the malimg dataset on DWT level three decomposition with db4 wavelet family were generated and SVM is applied on them. The results were efficient with an accuracy of 98.88 % to identify malware class.

This feature vector is also passed to the k-nearest neighbor classifier value of k is taken as 3 suppose, then the binary classification is done on the dataset with calculation of distance metrics. Result of the experiment is Euclidian distance metric gives better classification result compared to other distance methods with accuracy of 98.84%.

Malware Family	Samples	False Positive Rate(FPR)	True Positive Rate(TPR)	Accuracy
C1	219	14	219	100%
C2	295	0	295	100%
C3	335	0	335	100%
C4	111	0	111	100%
C5	80	0	77	96%
C6	97	12	85	88%
C7	88	4	84	95%
C8	485	0	485	100%
Average Accuracy				98.88%

TABLE I. MALWARE CLASS RECOGNITION by SVM

KNN (K=3)	Euclidian (Acc= %)	City block (Acc= %)	Cosine (Acc= %)	Correlation (Acc= %)
C1	92.78%	92.30%	86.53%	87.50%
C2	99.29%	99.29%	98.94%	98.94%
C3	100%	100%	99.76%	100%
C4	100%	100%	99%	100%
C5	100%	100%	100%	100%
C6	100%	100%	98.80%	99.89%
C7	98.68%	98.68%	100%	100%
C8	100%	100%	100%	100%
Average Accuracy	**98.84%**	**98.78%**	**97.87%**	**98.29%**

TABLE II. MALWARE CLASS RECOGNITION By KNN

On comparison of two classifiers on the dataset, feature vector in that SVM provides better classification accuracy compared to KNN. This is observed by the confusion matrix results of SVM and KNN.

Algorithm	Training Samples	Testing Samples	Feature Vector	Accuracy (%)
KNN k=3	1610	1710	56	98.84%
SVM	1610	1710	56	98.88%

TABLE III. MALWARE CLASS RECOGNITION by SVM and KNN

The conclusion given by authors is as follows “They compared the results with existing work of [6] which uses 320 feature vector for classification of malware variants but in this we used 56 feature vector, which can reduce the complexity by using wavelet transforms. The investigational results (Table I, II, III) showed that the proposed algorithm gives better results for classifying the malware families. This feature extraction method gives accurate malware class with less classification error.”

III. ADVANTAGES AND DISADVANTAGES

Advantages

- Efficient classification results could be obtained using this procedure compared to traditional methodologies.
- This method classifies accurately the malware classification using image processing techniques.
- Classification error obtained is less compared to other methods like artificial neural networks, Gray level matrix, system call.
- Visualization is an effective way for antivirus software to detect malware compared to others.
- This method for both packed and unpacked data.

Disadvantages

- Does not represent actual malware behavior.
- Uses only static analysis of malware detection.
- More time is consumed converting the binaries to a greyscale images for large data set.
- Classifying malware samples on global structure is a challenging task.
- Analysis of the large malware samples using image processing techniques is also a challenging task.

IV. SUMMARY AND DIRECTIONS

Detecting and classifying malware is of utmost important aspect in the present era. Depending on the malware detected the antivirus software companies can upgrade their software and analyze the future trends in malware. The approach taken by this paper is an efficient one compared to other traditional approaches, which only work only on small data sets. The feature vector extracted from the dataset is passed via two algorithms KNN and SVM and final accuracy obtained that is 98.84% and 98.88% respectively.

Enhancements

- The method can also be extended in future to work on dynamic analysis instead of only static analysis.
- 56 feature vector’s complexity could be reduced more by choosing proper wavelet transforms, which in turn saves time while executing program.
- Proper techniques could be invented to classify data samples globally with ease.
- I also believe that using various computer vision techniques along with this opens the path to a broader spectrum of novel ways to analyze malware.

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] 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.

[5] 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.

[6] Nataraj L., Karthikeyan S., Jacob G., Manjunath B. S.,"Malware images: Visualization and automatic classification," InProc. 8th Int. Symp. Visualization for Cyber Security, VizSec(2011), ACM, pp. 4-7.

[7] Aziz Makandar and Anita Patrot," Malware Analysis and Classification using Artificial Neural Network," IEEE Explorer International Conference on Automation, Communication and Computing Technologies (ITACT 2015), December 22 and 23, Bangalore.

[8] Lee, T. and Mody, J.J. 2006. Behavioral classification. EICAR 2006.

[9]. Tuceryan, M. and Jain, A.K. 1998. Texture Analysis, In The Handbook of Pattern Recognition and Computer Vision (2nd Edition),pp. 207-248.