IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005219|

A New Learning Approach to Malware Classification Using Machine Learning Algorithms and Tools

T. Shalini[1], Tutike Naga Venkata Siva Sai Pratap[2], Unnam Chinna Anjnaeyulu[3], Sanduri Madhu[4], Velaga Bhargav[5]

[1] Assistant Professor, Department of ECE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Andhra Pradesh, India

UG Student, Department of ECE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India[2,3,4,5]

ABSTRACT: As the technology is growing day by day, internet is getting developed with huge number of threats. Modern malware is designed with mutation characteristics, namely polymorphism and metamorphism, which causes an enormous growth in the number of variants of malware samples. Identifying certain types of malware is an important step in successfully removing it. Non-computer viewing is an integral part of malware analysis techniques, where a piece of malware is converted into an image to visualize and classify. Despite the great success, it is still difficult to bring out the functional element of the texture challenging database presentations. Existing methods use critical imagery features of the world in the areas of the related code. In this paper, we introduce a new reading framework for more discrimination and strong descriptions of the features. The proposed method works with local descriptions that exist as location binary patterns and the dynamic element of a fixed scale change, by combining them into blocks and new uses a bag of visual word bag to get stronger features, more flexible than earth features and so on stronger than local features. We are exploring the proposed approach to three non-computer program databases. Testing the results show that the definitions obtained lead to the performance of high-level categories.

KEYWORDS: Key word1: Malware, Key word2: UML, Key word3: GIST, Key word4: Dense SIFT, Key word5: LBP

I. INTRODUCTION

With the arrival of the internet, malware has come one of the most important pitfalls. relating certain types of malware is an important step in successfully removing it. Non-computer viewing is an integral part of malware analysis ways, where a piece of malware is converted into an image to fantasize and classify. Despite the great success, it's still delicate to bring out the functional element of the texture Grueling database donations. Being styles use critical imagery features of the world in the areas of the affiliated law. In this paper, we introduce a new reading frame for further demarcation and strong descriptions of the features. The proposed system works with original descriptions that live as position double patterns and the dynamic element of a fixed scale change, by combining them into blocks and new uses a bag of visual word bag to get stronger features, further flexible than earth features and so on stronger than original features. We're exploring the proposed approach to threenon-computer program databases. Testing the results show that the delineations attained lead to the performance of high- position orders.

Malware (e.g. contagions, worms and Trojan nags) has come one of the most significant pitfalls on the Internet. With the help of generation tools, it becomes easy to induce new malware, performing in a veritably rapid-fire increase in the number of malware. AV test reported that around new malware samples were attained in 2017, a 14% increase compared to the former time. Among all these malware attacks, over 67 targeted Windows systems. It has caused serious trouble. For illustration, the ransomware ‘‘WannaCry ’’ spread over 100 countries in the world and caused damage of 8 billion US bones. likewise, those new variant vicious law lines have analogous Geste as benign lines, making them harder to be detected, which has posed a significant challenge to anti-virus merchandisers. Although colorful analysis ways have been studied to deal with malware variants, they aren't sufficient to address adding avoidance ways applied in malware. New analysis ways are still demanded to ameliorate the analysis effectiveness. Among different ways, malware visualization has lately been proposed as an effective approach. In this paper, we propose a new system that classifies malware families using malware visualization. The system transforms malware double lines to grayscale images.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005219|

To gain discrimination features, we present a new literacy frame which is formulated as a multi-layered model to characterize and dissect malware images using bag- of visual- words (BoVW). Starting from being original descriptors (LBP or thick SIFT), we group them into blocks and make histograms. The uprooted features are more flexible than global features (e.g. GIST) and further robust than original features. We estimate the proposed system on three datasets, which are all from the Windows platform. Experimental results demonstrate that the attained descriptors are robust and discriminational, which lead to state- of the- art bracket performance, outperforming being styles.

MOTIVATION:

The worst computer virus outbreak in history, Mydoom caused estimated damage of $38 billion in 2004, but its inflation-adjusted cost is actually $52.2 billion. At one point, the mydoom virus was responsible for 25% of all email sent.

Since 2013, malware has been spreading exponentially. The initial boom doubled the number of malicious files and programs infecting the web. In the following years, the growth might have slowed down, but it definitely hasn't stopped. Even with built-in antivirus software protecting the newest operating systems, there's more malware online than ever before.

- 560,000 new pieces of malware are detected every day.
- There are now more than 1 billion malware programs out there.
- Every minute, four companies fall victim to ransomware attacks.
- Trojans account for 58% of all computer malware.

II. RELATED WORK

Various malware analysis and classification methods have been proposed, including signature-based detection [2], [3], behavior-based methods [4], [5], instruction frequency-based methods [6]–[10], opcode-sequence based methods [11]– [13], etc. Among them, some techniques help analysts analyze malware with feature visualization. Based on the observation that control flow information could be used to identify malware variants, Cesare and Yang [14] developed a control flow graph-based malware classification method. Trinius et al.

III. PROPOSED MODEL

In this paper, we propose a new method that classifies malware families using discriminative feature extraction. This method works with KNN and Random Forest algorithms. The different features were extracted from the images. SIFT, LBP features. The extracted features are combined together to generate the test features. The extracted features are more flexible than global features (e.g., GIST) and more robust than local features. We evaluate the proposed method on three datasets, which are all from the Windows platform.

IV. METHODOLOGY

Before we develop a code for extracting features of malware, we have to think about the layout to be displayed on screen instead of running code and checking errors. With help of tkinter we created a interface that can easily understandable by user. Here we developed a code that work in background with the help of UML diagrams.
Tkinter is the standard GUI library for Python. Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.

UML diagrams:

A UML diagram is a diagram based on the UML (Unified Modeling Language) with the purpose of visually representing a system along with its main actors, roles, actions, artifacts or classes, in order to better understand, alter, maintain, or document information about the system. This is the process we are going to do work in developing code. Figure 1 shows the flow diagram of UML diagram we are implementing.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005219|

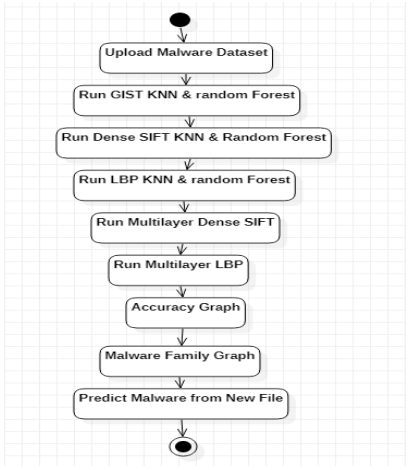


Figure 1: Flow diagram

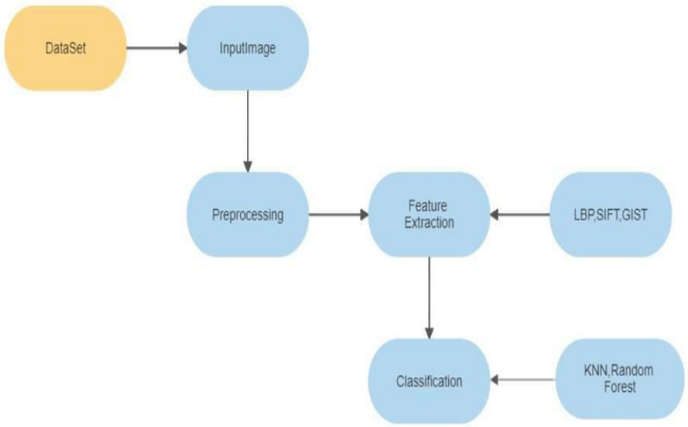


Figure 2: Flow chart of malware classification

Figure 2 describes the process we are following in our implementation.

Multi layered model can be explained in 4 layers.

Layer 1: We apply this layer to either LBP or SIFT to extract features from malware images.

Layer 2: In order to obtain accurate features, image will be split into multiple blocks and then gather relevant features.

Layer 3: KMEANS algorithm will be applied on accurate features to form clusters.

Layer 4: It gathers or extracts all important features from KMEANS and form a Bag of Visual Words vector and this vector will be feed to either KNN or Random Forest to calculate its prediction accuracy.

GIST:

GIST is a data structure and API that can be used to build a variety of disk-based search trees. GiST is a generalization of the B+ tree, providing a concurrent and recoverable height-balanced search tree infrastructure without making any assumptions about the type of data being stored, or the queries being serviced.

DENSE SIFT:

Dense SIFT calculates a SIFT descriptor determined by Lowe's algorithm at every location [22]–[24]. It collects features at each location and scale in an image, which helps increases recognition accuracy. It splits an image into small patches, and each patch is further spilt into smaller bins. The feature is then computed as gradient magnitude histograms in 8 orientations of bins. As the sliding window moves, it computes gradient histograms of each local neighborhood of the image. Finally, it obtains the image feature descriptors using cascaded connection functions

LBP:

LBP (Local Binary Patterns) turns a local center pixel grayscale value into a binary pattern that encodes the relationship of the pixel with its local neighborhood. Each neighboring pixel is set to 1 or 0 according to whether the grayscale value of the pixel is larger than the value of the central pixel.

Local binary patterns (**LBP**) is a type of visual descriptor used for classification in computer vision. It has since been found to be a powerful feature for texture classification; it has further been determined that when LBP is combined with the Histogram of oriented gradients (HOG) descriptor, it improves the detection performance considerably on some datasets.

KNN:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. This algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. This algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

Here, we are working with these algorithms to get more accuracy so that we can easily detect the malware family. After working on the code required for malware detection, while we move for execution process after running the code the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005219|

first page displayed is the layout as shown in figure (a) which can be understand by any others. The image shows the outlook that we are going to do the next steps.

We use confusion matrix to find accuracy.
Confusion Matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

Let's now define the most basic terms, which are whole numbers (not rates):

- **true positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.
- **true negatives (TN):** We predicted no, and they don't have the disease.
- **false positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- **false negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")
- Accuracy: Overall, how often is the classifier correct?
  (TP+TN)/total

At the end of the execution, we came to know about the malware families count and also to which family it belongs with the help of bar graph that is displayed in our output as shown in figure (b) in which Y-axis represent number of malwares of particular family and X-axis represent type of malware family.

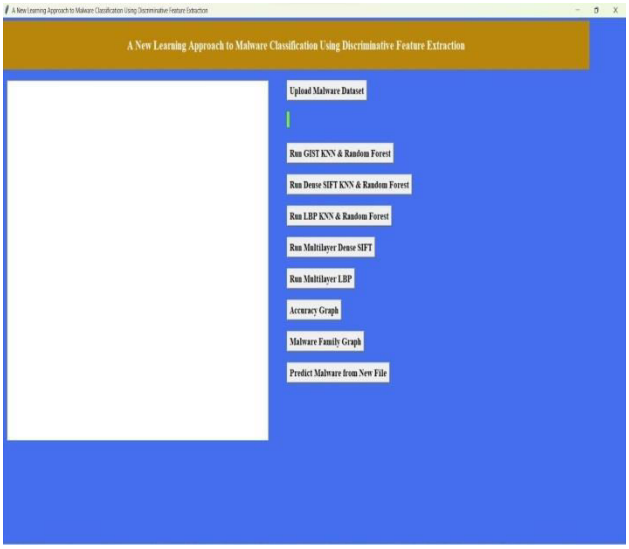


Figure (A)

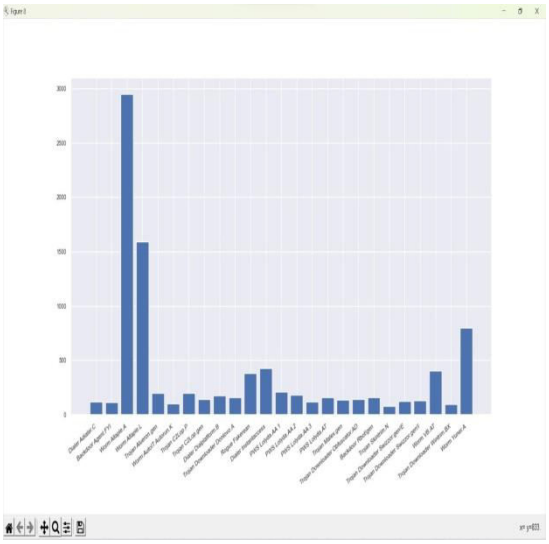


Figure (B)

V. RESULTS & DISCUSSION

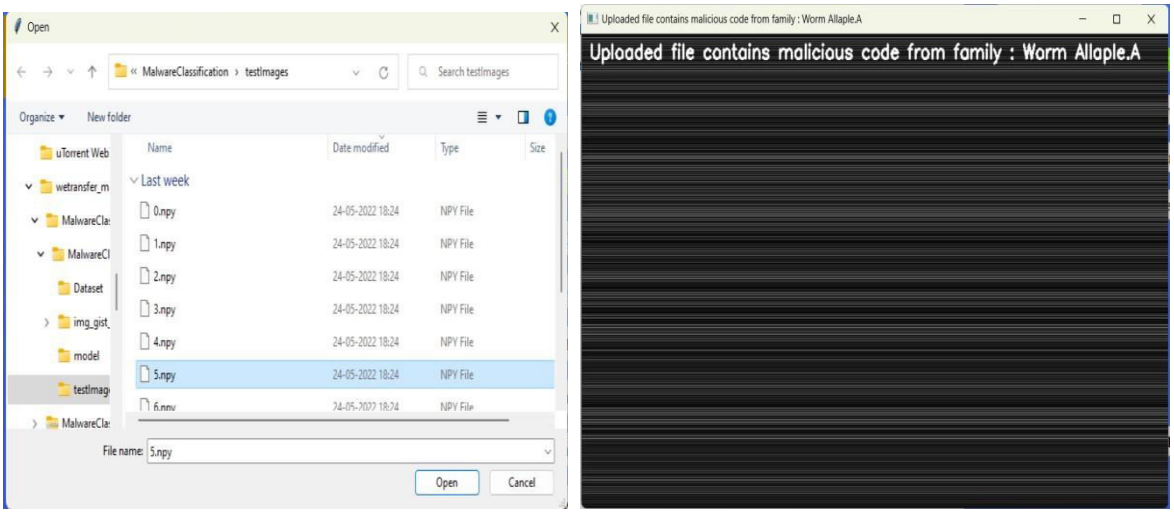
Now, we are checking the result by giving a file as input and it displays the type of family it belongs to. The displayed figure is our final output, which is in the form of image.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005219|



VI. CONCLUSION

In this paper, we propose a multi-layer literacy frame grounded on a bag- of-visual- words (BoVW) model to gain point descriptors of malware images. The model can gain more robust features and achieve better bracket accuracies indeed for further grueling datasets, compared with other Styles. However, if the malware execute lines are packed, multi-subcaste literacy frame will still produce harmonious results.

REFERENCES

1.AV-Test, Apr. 2018.
2.V. S. Sathyanarayan, P. Kohli and B. Bruhadeshwar, "Signature generation and detection of Malware families", Proc. Australas. Conf. Inf. Secur. Privacy, pp. 336-349, Jul. 2008.
3.M. F. B. Abbas and T. Srikanthan, "Low-complexity signature-based Malware detection for IoT devices", Proc. Appl. Techn. Inf. Secur., pp. 181-189, Jun. 2017.
4.A. Mohaisen, O. Alrawi and M. Mohaisen, "AMAL: High-fidelity behavior-based automated Malware analysis and classification", Comput. Secur., vol. 52, pp. 251-266, Jul. 2015.
5.H. S. Galal, Y. B. Mahdy and M. A. Atiea, "Behavior-based features model for Malware detection", *J.* Comput. Virology Hacking Techn., vol. 12, pp. 59-67, May 2016.
6.K. S. Han, B. J. Kang and E. G. Im, "Malware classification using instruction frequencies", *Proc. ACM* Symp. Res. Appl. Comput., pp. 298-300, Jan. 2011.
7.P. Natani and D. Vidyarthi, "Malware detection using API function frequency with ensemble-based classifier", Proc. Int. Symp. Secur. Comput. Commun., vol. 377, pp. 378-388, Nov. 2013.
8.Y. Ye, T. Li, Y. Chen and Q. Jiang, "Automatic Malware categorization using cluster ensemble", Proc. 16th *ACM SIGKDD* Int. Conf. Knowl. Discovery Data Mining, pp. 95-104, Jul. 2010.
9.E. Kirda, C. Kruegel, G. Banks, G. Vigna and R. A. Kemmerer, "Behavior-based spyware detection", Proc. 15th Conf. USENIX Secur. Symp., pp. 273-288, Jul. 2006.
10.S. Cesare and X. Yang, "A fast flowgraph based classification system for packed and polymorphic Malware on the endhost", Proc. 24th IEEE Int. Conf. Adv. Inf. Netw. Appl., pp. 721-728, Apr. 2010.
11. I. Santos et al., ''Idea: Opcode-sequence-based Malware detection,'' in Proc. Int. Symp. Eng. Secure Softw. Syst., vol. 5965, 2010, pp. 35–43.
12. I. Santos, F. Brezo, B. Sanz, C. Laorden, and P. G. Bringas, ''Using opcode sequences in single-class learning to detect unknown Malware,'' IET Inf. Secur., vol. 5, no. 4, pp. 220–227, Dec. 2011.
13. P. O'kane, S. Sezer, and K. Mclaughlin, ''Detecting obfuscated Malware using reduced opcode set and optimised runtime trac,'' Secur. Informat., vol. 5, no. 1, pp. 2–13, May 2016.
14. S. Cesare and X. Yang, ''Classification of Malware using structured control flow,'' in Proc. 8th Australas. Symp. Parallel Distrib. Comput., vol. 107, Jan. 2010, pp. 61–70

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details