Encounter of the Malicious Code Based on Deep Learning

G Wani
B. Tech scholar
Dept. of Computer Science
Ilahia college of engineering and technology
Muvattupuzha, India

Shapna V M
B. Tech scholar
Dept. of Computer Science
Ilahia college of engineering and technology
Muvattupuzha, India

Veena T P
B. Tech scholar
Dept. of Computer Science
Ilahia college of engineering and technology
Muvattupuzha, India

Aleena Anna Jacob
B. Tech scholar
Dept. of Computer science
Ilahia college of engineering and technology
Muvattupuzha, India

Shanavas K A
Asst. Professor
Dept. of Computer science
Ilahia college of engineering and technology
Muvattupuzha, India

***Abstract*— The pandemic series we have been facing has literally made us depend on the online platform. From our jobs to the education system has now been running online. The number of users that use the online sites from each corner of the world has increased. These numerous numbers of users that visit different online sites has end up operating a malicious code to the site or software they are using automatically ,which also malfunction the computer security. In order to solve this criterion, this paper introduces a method that identify and classifies the malicious code based on deep learning, in this paper we study about various kinds of technique which identifies different malwares and the types and their corresponding family and their purpose. In order to implement, in this paper it is used deep learning, with the help of Convolutional Neural Network (CNN), image processing, based on deep learning, deep belief network (DBN), and many other kinds of techniques are used to identify the malwares and their purpose.**

Keywords*—. Malicious code; Deep learning; Coloured image; Convolutional neural network..

I. INTRODUCTION

The mass usage of internet has been a threat to the computer system through the excessive usage of various online sites and other platforms. This has increased rapidly during the pandemic series that happened all around the world. Literally the whole system of life has been transformed to the world of internet, from the jobs to the education system has been running online, causing a major threat to every system that has been handled by numerous users. One of the main threats of internet security is the exponential growth of malicious code. Sometimes computer users download the file or software from internet which is malicious intent. Malicious software inserted in a system for a harmful purpose that can be categorised in to independent and dependent if host need a program. Viruses, logic bombs, and backdoors are examples. Independent malware is a constrained program run by the system.

Such software comes in many forms such as Virus, Trojan, worm etc…which is used to destroy computer operation and private security [1]. Cyber security reports that, millions of account details were stolen from websites, and offered for sale [2]. In recent report google detected around 600 - 800 malware infected site per week [3]. Infect the system to steel or disturb the business through malware is an old technique since 1988. And also, every day by day it is increased. According to the latest survey report there is more than1 billion malware exists [4]. The below figure (1) shows the increasing growth rate of Malicious code.

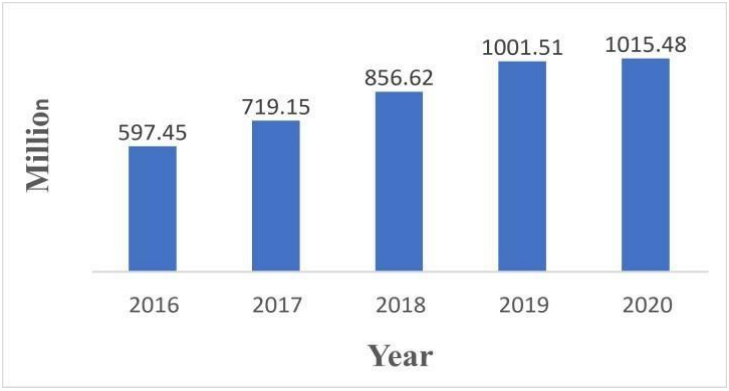

fig (1): Increasing growth rate of malicious code

For security protection malicious codes are analysed. There are two methods to malicious detection. Static detection and dynamic detection both are feature-based detection methods. To identify whether it is malware or not, dynamic detection helps to verify the behaviour and

action of code. Static analysis, collect the information about malicious application without running it [5] both these methods have some disadvantages. Some of the code may ignored because the execution environment does not comply with the rules.

Recently, Zhuhai cui et al. [6] introduced new method which convert malicious code into grey scale images and then extract the features. The approach uses CNN to identify the malware, with the help of deep learning.

Challenges**:** Malicious codes are become more complicated in day by day. So, using grey scale image is not easy to extract the features and also malware families are increased in every month. There is a chance to detect a new malware which is not till discovered.

II. RELATED WORKS

Ankur Singh brist,[7] he describes Malicious codes are Menace to our society. Computer viruses, Trojans, Worms etc. are the examples of malicious code which makes a copy of itself to spread to other computers. Malicious codes are an area of unaltered programming, they carry their own functions from being found and destroyed, for protection, they have to come through tough impenetrable conditions. Identification and detection of malicious codes are effective in preventing infections. There are many ways used for detection and identification of malicious codes, through basic concepts such as Signature based, Datamining, Rule
based, Scan string

- Use string scanning techniques.
- Used various methodologies for virus prevention.

His specification includes, Heuristic technology, Rule based system, Check sum, Scan the strings, which will provide improper output. Related to the above method this paper introduces Data mining techniques in worm detection and Natural virology mapping. The virus also can be detected using deep learning.

Dongzhi Cao, Yongli yang [8] came up by saying that, internet grows malicious code also generates rapidly which becomes great cause of security issue. In order to prevent the above issue, introduce or focus mainly in static and dynamic detection. The existing papers includes varies method for detection of malicious code based on signature behaver based, heuristic based methods, these may cause some problems in order to resolve we used malicious code detection based on Convolution Neutral Network Restricted Boltzmann (RBM) in Deep belief network (DBN). Detection of malicious code can be effectively done by differentiating the code in to normal and malicious data, to bring out the relevant characteristics and to build a better model, that differentiate malicious data and two criteria which leads to better performance of malicious code, data. As mentioned above the Auto encoders and DBN has the following advantages, which ensures time management.

- Proposed an automatic detection.

According to him binary file of the code to be detected is mapped to an un compressed grey scale image. Operations on grey scale images such as resizing, de- averaging. Based on the above this paper involves binary files of malicious code are transferred into coloured images through CNN.

According to Rajesh Kumar, Zhang xiasong [9] In this they have come across many malware detection methods such as signaturebased Datamining etc., which might have some kind of problems or issues during detection we cannot detect with proper accuracy of some malware such as Trojan's horse, Root list, etc., which are very difficult in detections, in order to resolve the problem, here detection based on image processing using deep learning. This image processing technique analyse the malware binaries as grey scale images with accuracy and Taobao is a popular image processing method for object recognition. Their main specification is to detect the unknown or new type of malware using CNN approach and convert into grey scale image, based on the above process malware detection through image processing techniques that convert malware binaries as greyscale and converted to coloured to improve the accuracy.

According to Yuanchengi Li, Ruhai [10] ,This article describes the detection of malicious code based on Auto encode and Deep Belief Network (DBN), it helpful in reducing the dimensionality of data malicious code also destroys the system through mathematical operations such as adding, changing, deleting some code, this can be resolved by two approaches, Host based network based, to enhance better performance and to ensure time management, we introduce RBM in DBN. Detection of malicious code can be effectively done by differentiating the code into normal and malicious data. The relevant characteristics and to build a better model, that differentiate malicious data and two criteria which leads to better performance of malicious code. Mainly combines the advantage of Autoencoder and DBN.

According to William S Treaswell [11] he describes the article by that the device contains processor and memory, that has a malicious code software that prevents the malicious code execution by detecting a request for the execution of a file. The file is scanned for a risk before processing a request. The risk is assigned by a score and the execution of file is responsible to the risk score. One of the advantages is each file is scanned to detect the malicious code mainly focus on Detection of malicious software, files typically include a combination of heuristic scanning and signature definition. Related to the above here we make a difference that, Detection of malicious code is based on signature and image.

According to Riaz Ulla Khan [12] He focuses on the dataset contains the collection of files that has two subsets. First, subset containing malicious code and another containing benign code files that identified by antivirus program. All features of files are selected and reduced by feature selection methods. They will be used as training and test sets after determining optimal number of features, the

author finds the benefit that the malicious code files are identified by antivirus program. In the above paper Dataset is created which is a collection of files and code files and the number of steps for detection are high and complicated, based on this we introduce that dataset are mapped to a coloured image, hence reduces complexity and number of steps are less and easy to understand.

According to Michael L santacrose [13] in he came with the novel method of viewing malicious assembly by transforming executable byte code into video rather than image. It uses quantization techniques and establish the best network architecture for classifying the malware which based time distributed convolutional neural network. He mainly focuses on a method for viewing the grey image rather than a video which require source assembly and not entire file. Her we introduce that the images are converted grey scale images to coloured images, and it improves pooling for malware images.

According to EIIMouatez Bilhah Karbab, Mourad debbabi [14] this paper defines popularity of android devices is increasing and becomes the target of malicious app in order to reduce the effect of malicious codes 'Malware' – an automatic android malware detection and family attribution – framework can be established as a precaution of the malicious codes. Malware is capable to automatically extracts and learn being patterns to detect the android malware. It can detect accurately the malware and families which belong them with an FI – score of 96% - 99%. It is used only in android. We introduce the method which can be used in android IOT devices and all the computer appliances.

A Mural Fiskiran [15] this paper introduces malicious codes that can affect the system REM is the method in which used to detect programs flow anomalies with malicious codes. The REM can be performed by verifying the program codes integrity at the hash block level which monitor the behaviour of malwares with performance degradation with REM ranging with 6.4% average increasing the size of LI instructions cache reducing to 5%. He also defines hash block that excludes the single-entry point requirement, and verifies the program code. Related to above said the proposed system includes all program codes as their method may leads to misclassification.

Limin shen [16] he describes that Android is the most targeted system by malwares the two-layer deep learning method comprising the first layer with permissions and component information based static detection method Following the first layer, the second layer takes the input from first layer which cascades CNN and auto encoder for the detection of malware. The feature of two-layer model is that it is capable of verification and classifying the categories and family of the malicious codes the advantage is that he proposed an automatic detection.

Binglong Li [17] mainly focuses on semantic features of malicious codes rather than the detection of codes based on signature and other regular criteria. This can be functioned by a pre-processing method of APK file which generate graphical semantics for graph convolutions network (GCN). GCN model automatically identifies and learns this by extracting the features of malicious code detections. Focus on static features of malicious, that is sensitive application programming face and programming functions. the detection is based on signature. This paper describes that image are used for detecting new malware, and focus on dynamic analysis and static analysis that its feature library is small and not need to update frequently.

III. PROPOSED SYSTEM

This paper presents the improved malicious code variant detection method based on a CNN, which include, the malicious code mapping as the grayscale image; and the CNN design for grayscale image detection. Fig. (2) presents an overview of these two processes. First, the binary files of malicious code are transformed into the grayscale images.

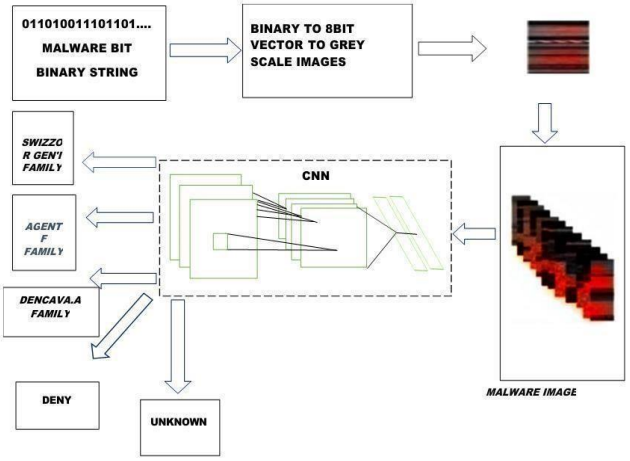

Fig (2): System Architecture

Identification and classification of images are employed through Convolution neural network, which results automatic recognition and classification of malicious software.

A. *Binary to Decimal Conversion*

In general, there are several ways to transform binary code into images. In this paper, we used the visualization of executable malware binary files [18]. A malware binary bit string can be split into a number of substrings that are 8 bits in length. Each of these substrings can be seen as a pixel, because the 8 bits can be interpreted as an unsigned integer in the range 0 to 255. For example, if a bit string is 01100000101011100, the process is 01100000,10101100→01100000,10101100→96,172. An eight-bit binary number B= (b7, b6, b5, b4, b3, b2, b1, b0) can be converted into a decimal number I as follows: I =b0 2^0 + b1 2^1 + b2 ∗ 2^2 + b3 ∗ 2^3 + b4 ∗ 2^4 + b5 ∗ 2^5 + b6 2^6 + b7 2^7(1)

B. *Decimal to Coloured Images*

After the above conversion, bit string has been switched into a 1-D vector of decimal numbers. According to a specified width, this one-dimensional array can be treated as a 2-D matrix of a certain width. Finally, the malicious code matrix interpreted as a grayscale image. For simplicity, the width of the image is fixed, and the height of the image varies depending on the size of the file.

C. *Convolutional Neural Network*

A new CNN was developed to classify malware the structure of the CNN for grayscale image recognition consists of several components, as shown in Fig. (3) First is the input layer, which brings the training images into the neural network. Next are the convolution and sub-sampling layers. The former layer can enhance signal characteristics and reduce noise. The latter can reduce the amount of data processing while keeping the useful information. Then there are several fully connected layers that convert a two-dimensional feature into a one-dimensional feature that conforms to the classifier criteria. Finally, the classifier identifies and sorts the malware images into different families according to their characteristics.

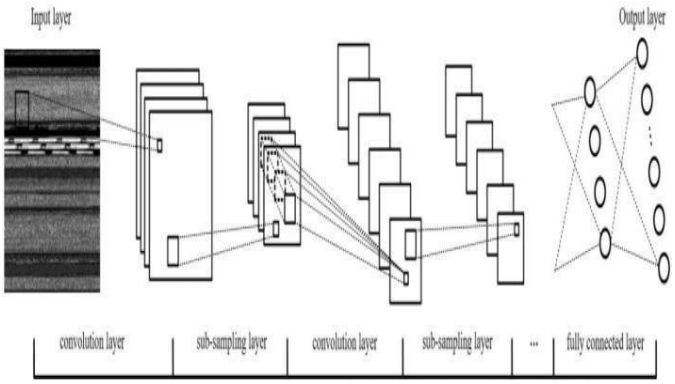

Fig (3): Basic diagram for convolution neural network

Approximately 15 malware families where trained, each family consists of approximately 300 images. Each corresponding malware family are difficult to find, so, an additional two classes where added that is “Deny” and “Unknown”. Deny indicate the images is not a malware. Unknown indicate images is a malware but not included in the discovered families. And the extracted images are automatically mapped to corresponding classes with the help of CNN.

IV.RESULT

This paper mainly used to detect and identify the particular URL code is affected by malware or is it safe from malware and in the case of coloured image i.e., converted grey first and then convert the RGB to coloured does the further process. That is keeps away from malware which is beneficial especially in this pandemic year. Which keeps threat from computer security.

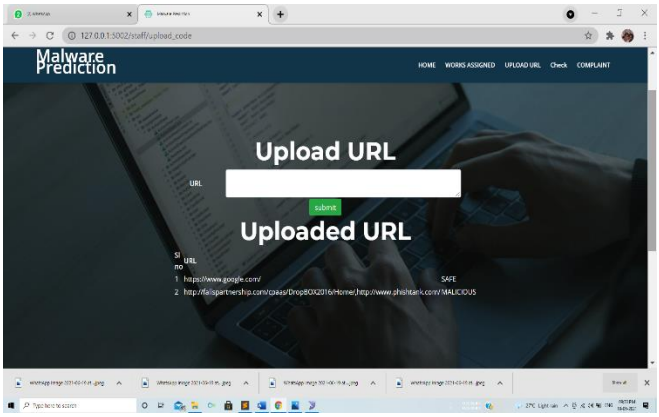

Fig 4

The above fig shows the output for the detection and identification of URL code and specifies that particular code is safe from malware.

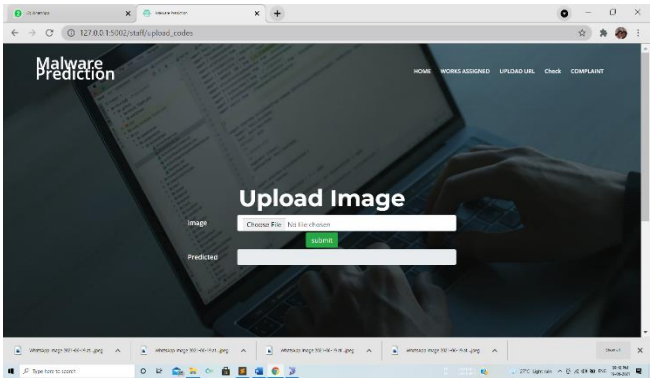

Fig 5

This fig shows detection and classification of an image, whatever the image may be from google or any supported social media particular chosen file is being extracted and converted into step-by-step process as mentioned above using algorithm.

V. CONCLUSION

Due to the pandemic series people were depending on the online platform for each and every moment of the life, purpose for business, transaction and the education system has now been running online. The number of users that use the online sites from each corner of the world has increased, which may affect the computer security, in order to solve this criterion, this paper introduces a method that identify and classifies the malicious code based on deep learning. In this paper, proposed that the images were identified and classifies by CNN and forms into a binary bit and that is converted to coloured image and identifies which type of family that belongs and can be easily identified for what purpose that malware has entered to computer system/software.

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of this project's preliminary report depends largely on the encouragement and guidelines of many others. We would like to take this opportunity to thank everyone who contributed to the successful completion of this project. We would like to show our heartfelt gratitude towards Prof. Dr. ABDUL GAFUR M, Principal, Ilahia College of Engineering and Technology for permitting us to work on this project. Also, we would like to show our greatest gratitude towards our head of the Department of Computer Science & Engineering Dr. Arun E and project guide Mr. Shanavas K A for their valuable advice and guidance. Finally, we express our gratitude and thanks to all our teachers and other faculty members of the Department of Computer Science & Engineering, for their sincere and friendly cooperation in completing this project.

REFERENCES

[1] Saloni Khurana: A Review paper on cyber security
[2] Cyber security report 2020: https://www.ntsc.org
[3] Sam cooks: Malware statistics and facts for 2021: - https://www.comparitech.com/antivirus/malwarestatistics-facts/
[4] https://geekflare.com/website-malware-scanning/
[5] P V Shijo : Integrated static and Dynamic analysis for malware detection
[6] https://www.sciencedirect.com/science/article/pii/S187 7050915002136
[7] Zhihua cui, Fei Xue Et Al :Detection of malicious variants based on deep learning https://www.researchgate.net/publication/3264131 80_Detection_of_Malicious_Code_Variants_Based on_Deep_Learning#:~:text=This%20paper%20 proposed%20a%20novel,malicious%20code%20i nto%20grayscale%20images
[8] Ankur Singh Brist :"Classification and identification of malicious https://citeseerx.ist.psu.edu/viewdoc/download?do I=10.1.1.300.7546&rep=rep1&type=pdf [8] Dongozhi Cao,Yongali yang : An Efficient malicious code detection based on deep learning.
[9] Rajesh Kumar, Zhang Xiasong: Malicious code detection based on image processing. https://www.researchgate.net/publication/3256575 23_Malicious_Code_Detection_based_on_Image_ Processing_Using_Deep_Learning.
[10] Yuanchengi Li, Ruhai : Hybrid malicious code detection based on deep learning. https://www.semanticscholar.org William S Treaswell" Detection and prevention of malicious code execution" https://www.freepatentsonline.com/y2009/0165131
[11] Riaz Ulla Khan"Unknown Malcode detection using classifiers using optimal training sets" https://patents.google.com
[12] Michael L santacrose "Detecting malware code as video with compressed, time distributed neural network". https://www.researchgate.net/publication/3431187
[13] EllMouatez Billah karbab, Mourad debbabi "Android malware detection using deep learning on API method sequences"
[14] https://www.researchgate.net/publication/3220759
[15] A Mural Fiskiran "Runtime execution monitoring to detect and prevent malicious code execution" http://palms.ee.princeton.edu/PALMSopen/fiskira n04runtime.pdf
[16] Limin shen" Two- layer deep learning method for android malware detection using network traffic" https://www.google.com/url?sa=t&source=web&r c t=j&url=http://ieeexplore.ieee.org/document/916685&v
[17] Binglong Li" Malicious code detection based on semantic features" https://www.researchgate.net/publication/3393561
[18] 23_Detecting_Malicious_JavaScript_Code_Based _on_Semantic_Analysis