ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(03), March 2023, pp. 28-36

Article Title: Image-Based Malware Classification Using Deep Learning Techniques

Image-Based Malware Classification Using Deep Learning Techniques

R.R. Ramya[1], J. Banumathi[2*]

[1] Research Scholar, Department of Information Technology, University College of Engineering, Nagercoil, Tamilnadu, remmi.ky@gmail.com.

[2] Assistant professor, Research Scholar, Department of Information Technology, University College of Engineering, Nagercoil, Tamilnadu, jbanu85@yahoo.co.in.

ABSTRACT

The number of malware extending along through World Wide Web and endangering computer system as well as other internet connections has drastically risen over the past few years. To deadline, multiple methods and approaches for detecting and minimizing these malware entities have been suggested. Nevertheless, as innovative and integrated malware strategies arise, a large amount of malware is formed that can circumvent a variety of advanced malware tracking systems. As an outcome, in an effort to safeguard information from harmful activity, this article suggests an image-based malware classification utilizing Deep Learning (DL) techniques. To facilitate the process, the input dataset is pre-processed using the Keras pre-processing method. Data preparation and data processing are two aspects of data pre-processing. The preprocessed data is fed into the data visualization block after pre-processing. These visualizations aid in determining how data is handled as well as possible irregularities in the data. Finally, for improved accuracy, the convolution neural network (CNN) DL classification method is used. The obtained outcomes are validated using python platform. The proposed CNN attained high accuracy in contrasted with other approaches.

Keywords: Malware, Keras pre-processing, CNN, DL

1 Introduction

Malware is malicious software that is designed to affect information systems. It is employed to harm, invade, or steal information to a certain digital products which might be extremely delicate to access, or to lead to damage or unforeseen consequences in a system. This is essential to establish defense systems in opposition to malicious code attacks, which can interrupt entrepreneurs by infiltrating their own cloud applications, information, and implementations without obtaining user permission and authorization [1-3]. It is a significant difficulty to properly identify such modern obscured malicious files that are being quietly developed in big scale, considering the enormous range of open-source software platform and the convenience with which software can be deployed through the Web [4, 5]. Thus, malware monitoring systems are developing in the sector of cyber security to tackle this malware conflict, and this work makes a modest advancement in this research path.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(03), March 2023, pp. 28-36

Article Title: Image-Based Malware Classification Using Deep Learning Techniques

The Internet of Things (IoT) is susceptible to web-based threats and viruses that can cause unauthorized access, manipulation of data, and other significant damage to societies and humans. Among most frequently utilized business systems for the identification and reduction of malware have been anti-virus software applications for many years [6, 7]. Such systems have historically relied on malware detection, which needed human involvement to keep and modify the database schemes when computer virus is detected and addressed. Furthermore, machine learning methods have been invented that can address the shortcomings of signature-based systems [8].

The application of both dynamic and static evaluation for detecting attacks is considered to be among the most often used methods. Whereas dynamic analysis involves monitoring the activity of malware while it is running, static analysis gathers data from malicious without ever operating them [9, 10]. Within big scheme of things, static and dynamic estimation are both considered to be more dependable and effective, but they both had significant drawbacks.

These self-learning systems have the ability of using data mining and DL techniques to acquire complicated malware sequences and differentiate among both benign and malicious binaries [11]. However, various self-learning methods display varying degrees of unpredictability in their capacity to identify emerging malware as well as variations of existing malware. Researchers have looked into ML approached for hybrid malware estimation. The above algorithm have shown promise in identifying obscured malware, and their own consequences were examined, inspiring additional study.

By experimenting with various classifiers, such as Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Hidden Markov Models (HMM) and Nave Bayes (NB), a number of static malware investigative techniques have separated their work [12, 15]. In general, dynamic research methods analyses different malware execution traces to determine its behavior patterns. But nevertheless, the significant number of these methods depend on extensive feature engineering to perform the use of shows the importance expertise to generate the features. It is an issue because the rate at which new malware is developed is extremely high, and this technique is unable to maintain.

Therefore, in this paper suggests an image-based malware classification is implemented to ensure the data against harmful behavior. For simplify the procedure, the input dataset is pre-processed using the Keras pre-processing method. With help of visualizations, the structure of the data along abnormalities of the system is determined. Furthermore, for better efficacy, the CNN based DL classification approach is applied. CNN provide better results, in terms of accuracy.

2 Proposed System Description

Figure 1 represents the schematic diagram of suggested image based malicious classification system.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(03), March 2023, pp. 28-36

Article Title: Image-Based Malware Classification Using Deep Learning Techniques

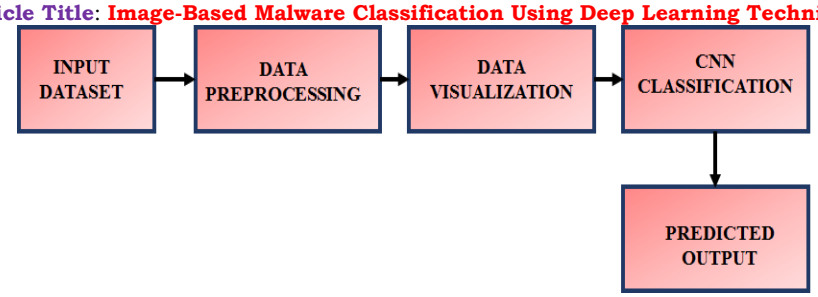


Figure 1: *Block diagram of proposed system*

Here, the input dataset is prepared using the Keras pre-processing technique. Data preprocessing is utilized to transform the information into an image which is more usable and approachable. Following pre-processing, the dataset is transferred to the data visualization module. These visualizations help to analyze the data's organization and the presence or absence of errors. The proposed CNN classification method is used with the dataset. After that, CNN's classification approach is implemented for higher precision. At the conclusion, a predicted output is obtained**.**

2.1 Input Dataset

Use different data sets to replace malware with an image, as the Malimg sample includes malware images. A hexadecimal version of the digital content of each file is accessible in the source data for each file. The goal is to convert those files into PNG images so that the CNN can use them as input**.**

2.2 Data Preprocessing

The kind of processing done on original data to get it prepared for another data processing operation is described as data preprocessing, which is a part of data preparation. It has historically been a significant first stage in the data mining process. Since accurate data is certainly more significant than excellent models, and because the accuracy of the information is of the highest concern, it's necessary to perform data preparation. As a result, organizations and individuals spend a lot of time cleansing and getting the data ready for modelling. The information that is accessible in the real world has several reliability issues, is noisy, unreliable, and incomplete. It might not have precise, pertinent attributes, and inaccurate information. Preprocessing is crucial to enhancing the quality of the data. By removing any errors or anomalies, standardizing the data so that it can be compared, and enhancing the outcome's quality, preprocessing assists in ensuring that the data is reliable.

2.3 Keras Preprocessing

The data preprocessing and data augmentation module of the Keras deep learning package is called Keras Preprocessing. It offers tools for interacting with picture, text, and sequence data. Preprocessing with Keras enables you to convert raw data stored on disc into a dataset object that can be used to train a model. Here's a brief illustration: If you wish to train a classifier that

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(03), March 2023, pp. 28-36

associates an image with its category and you have 10 folders, each of which contains 10,000 images.

2.4 Image data generator

The input of the initial data is obtained using the Keras Image Data Generator, which then transforms the input data randomly and outputs a result that solely contains the newly altered data. The data are not added. The original data inputs are used to feed the Keras Image Data Generator, which transforms the data randomly and outputs a result that only contains the newly altered data. The data are not included.

2.5 Data Visualization

To make the analysis easier for the human brain to comprehend and make conclusions from, data visualization is the practice of putting information into a visual context, like a map or graph. Data visualization's major objective is to make it simpler to spot patterns, trends, and outliers in big data sets. The work is visually represent data is known as data visualization. It effectively communicates findings from data by graphically plotting the data.

Also it is obtain a visual summary of our data via data visualization. The human mind processes and comprehends any given data more easily when it is presented with images, maps, and graphs. Both small and large data sets can benefit from data visualization, but enormous data sets are especially beneficial because it is impossible to see all of the data, let alone manually interpret and understand it.

2.6 Types of Data Visualizations

- **Tables:** It have columns and rows that are used to compare different variables. Tables can display a lot of data in a structured manner, but they can also be overwhelming for consumers who are only searching for broad trends.
- **Pie charts and stacked bar charts:** Parts of a whole are represented by the portions that make up these graphs. They offer an easy method for compiling data and evaluating the relative sizes of the various components.
- **Line charts and area charts:** These graphics, which are commonly used in predictive analytics, depict a succession of data points over time to demonstrate change in one or more quantities. Whereas area charts connect data points with line segments, stack elements on top of each other, and use colour to differentiate between variables, line graphs use lines to show these changes.
- **Histograms:** This graph shows the amount of data that falls inside a specific range by plotting a distribution of numbers using a bar chart (without spaces between the bars). With the help of this visual, a user can quickly find outliers in a given dataset..
- **Scatter plots:** These representations are frequently used in regression data analysis because they are useful in revealing the relationship between two variables. However, these can occasionally be mistaken for bubble charts, which use the x, y, and bubble size to represent three different variables.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(03), March 2023, pp. 28-36

- **Heat maps:** These visualisations are useful for displaying behavioural data according to location. It could be a place on a map or even a website.
- **Tree maps:** in which hierarchical data is displayed as a collection of nested forms, often rectangles. With their large surface areas, tree maps are excellent for analysing the ratios across groups.

2.7 CNN Classification

The idea of classification, which functions as a display method for elements, classes, and interfaces. A classifier defines a collection of cases that have similar structural and behavioural characteristics. In this work, CNN classifications are employed.

2.7.1 CONVOLUTIONAL NEURAL NETWORKS

Similar towards how people recognize images with help CNN function. They are based on investigations carried out to get better mammals experience the visual world, as detailed in. The research found that visual stimuli like edges and forms cause brain cells to activate. CNNs are well-known DL models that are essentially a feed-forward neural network and are frequently employed for image categorization issues. CNNs were implemented using Keras for this study. The following three types of layers are typically found in CNNs:

- Convolutional layer
- Pooling layer
- Fully-connected layer

Convolutional Layer

Convolutional layers are a crucial component of CNNs. Two functions are combined mathematically to form a convolution. The input and a filter are combined in this layer. It is possible to employ several filters of various diameters. The input is passed through each filter until it has been completely covered. Stride, which determines how far the filter moves in each step from left to right and top to bottom, can be used to govern the direction that the filter advances in. The feature maps created using all filters are the output of the convolutional layer.

Pooling Layer

A pooling layer appears after a convolutional layer. The result of the convolutional layer is down sampled in this layer. This drop directly corresponds to a decrease in calculations, and therefore, a decrease in training time. There are various methods for putting pooling into practice. For this research, the max pooling is employed

Fully Connected Layers

Among the most fundamental varieties of layers in a convolutional neural network is the fully-connected layer (CNN). Each neuron in a layer that is fully linked is, as the name implies, completely connected to every other neuron in the layer below. The final Fully Connected Layer offers a classified image based on the training data set. When the objective is to use the features learnt by the preceding layers to produce predictions, fully linked layers are often used towards the end of a CNN.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(03), March 2023, pp. 28-36

For image identification and classification applications, CNNs are frequently utilised. CNNs can also be utilised for more difficult tasks, like creating descriptions for images or locating the focal points in images. Furthermore, time-series data, such audio or text data, can be employed with CNNs. CNNs are an effective deep learning technology that have been utilised to produce cutting-edge outcomes in a variety of applications.

3 Results and Discussion

The obtained results of suggested system is implemented in python platform to verify its performance. There are 9339 malware images in the Malimg Dataset, which are divided into 25 families/classes. The input dataset is shown in Figure 2.

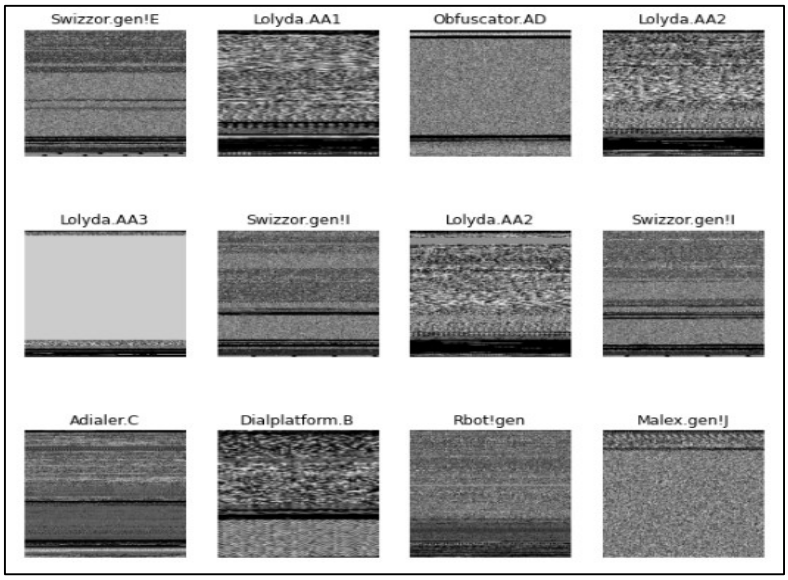


Figure 2: *Input Dataset*

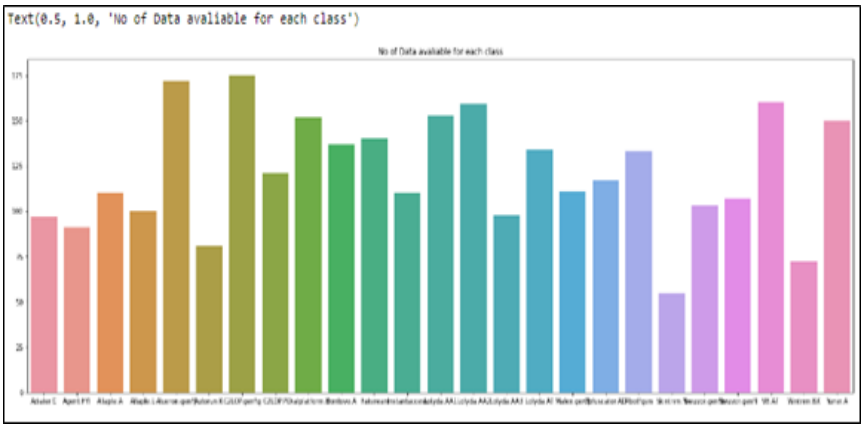


Figure 3: *Data Visualization*

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(03), March 2023, pp. 28-36

Figure 3 represents the data visualization of dataset for malware classification. Similarly, Figure 4 & 5 indicates the CNN classification with accuracy. From the Figure 5, is it clear that the proposed CNN classifier attained high accuracy in contrast to other techniques.

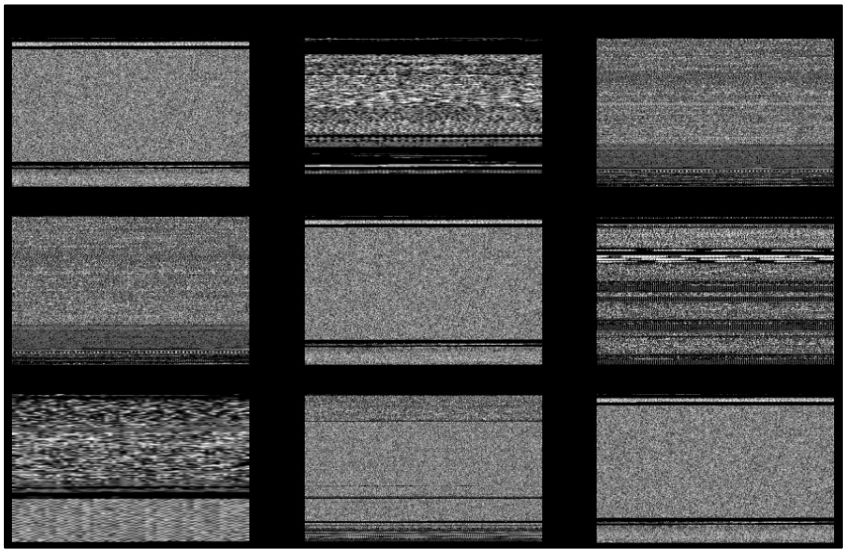
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 254, 254, 30)      840

 max_pooling2d (MaxPooling2D  (None, 127, 127, 30)     0
 )

 conv2d_1 (Conv2D)           (None, 125, 125, 15)      4065

 max_pooling2d_1 (MaxPooling  (None, 62, 62, 15)       0
 2D)

 dropout (Dropout)           (None, 62, 62, 15)        0

 flatten (Flatten)           (None, 57660)             0

 dense (Dense)               (None, 128)               7380608

 dropout_1 (Dropout)         (None, 128)               0

 dense_1 (Dense)             (None, 50)                6450

 dense_2 (Dense)             (None, 25)                1275

=================================================================
Total params: 7,393,238
Trainable params: 7,393,238
Non-trainable params: 0
_________________________________________________________________
```

Figure 4: *CNN Classification*

```
Epoch 6/10
85/85 [==============================] - 220s 3s/step - loss: 0.7297 - accuracy: 0.8010 - val_loss: 0.5533 - val_accuracy: 0.87
92
Epoch 7/10
85/85 [==============================] - 202s 2s/step - loss: 0.6805 - accuracy: 0.8223 - val_loss: 0.4935 - val_accuracy: 0.88
89
Epoch 8/10
85/85 [==============================] - 205s 2s/step - loss: 0.5934 - accuracy: 0.8385 - val_loss: 0.4973 - val_accuracy: 0.88
75
Epoch 9/10
85/85 [==============================] - 193s 2s/step - loss: 0.5467 - accuracy: 0.8488 - val_loss: 0.4633 - val_accuracy: 0.88
61
Epoch 10/10
85/85 [==============================] - 202s 2s/step - loss: 0.4871 - accuracy: 0.8624 - val_loss: 0.5576 - val_accuracy: 0.87
08
```

```
9/9 [==============================] - 33s 517ms/step - loss: 0.2609 - accuracy: 0.9444
[0.2609328627586365, 0.9444444179534912]
```

Figure 5: CNN Classification Accuracy



ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(03), March 2023, pp. 28-36

Article Title: Image-Based Malware Classification Using Deep Learning Techniques

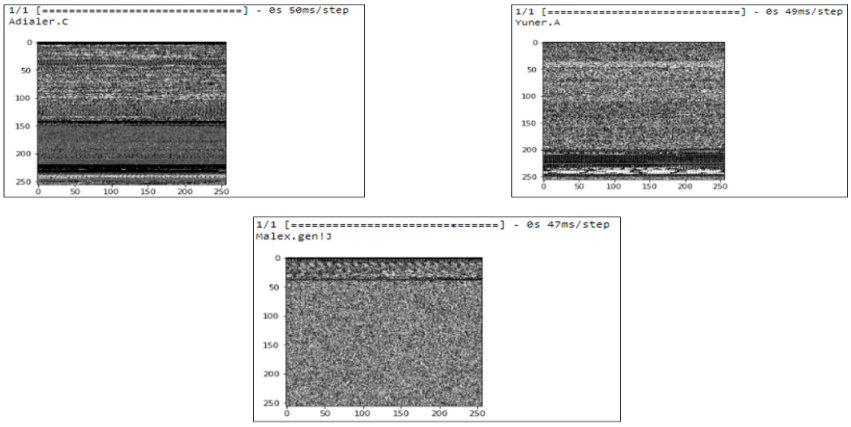


Figure 6: *Predicted Output*

The suggested method provides a clear classification of the many types of malware. Figure 6 displays the malware classification's result images individually.

4 Conclusion

The classification of malicious activity is a significant and difficult issue in information security. Targeted malware operations are typically made by a small number of criminal organizations. To effectively develop the variants for sophistication and prevent malware issues, the effective approach is needed. Hence, in this paper propose an efficient image-based malware classification be utilized to protect data from malicious activity. Also this pattern highlights the significance of accomplishing malware family classification in order to deploy effective prevention techniques for malware functions. The input dataset is pre-processed using the Keras pre-processing approach to make the process simpler. Visualizations are used to analyses the structure of the data and identify systemic errors. Moreover, the CNN-based DL classification approach is used for greater efficacy. For more accurate findings, CNN is excellent compared to other classifier methods.

References

1. Vasan Danish; Mamoun Alazab; Sobia Wassan; Babak Safaei; Qin Zheng, Year: 2020, "Image-Based malware classification using ensemble of CNN, “Image-Based malware classification using ensemble of CNN architectures (IMCEC),” Computers & Security, Vol: 92, pp. 101748.

2. Bhodia Niket; Pratikkumar Prajapat, Fabio Di Troia; Mark Stamp, Year: 2019, “Transfer learning for image-based malware classification,” arXiv preprint arXiv: 1903.11551.

3. Vasan Danish; Mamoun Alazab; Sobia Wassan; Hamad Naeem; Babak Safaei; Qin Zheng, Year: 2020, “IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture,” Computer Networks, Vol: 17, pp. 107138.

4. Awan Mazhar Javed; Osama Ahmed Masood; Mazin Abed Mohammed; Awais Yasin; Azlan Mohd Zain; Robertas Damaševičius; Karrar Hameed Abdulkareem, Year: 2021, "Image-based malware classification using VGG19 network and spatial convolutional attention," Electronics, Vol: 10, no: 19, pp. 2444.

5. Xiao Mao; Chun Guo; Guowei Shen; Yunhe Cui; Chaohui Jiang, Year: 2021, "Image-based malware classification using section distribution information," Computers & Security, Vol: 110, pp. 102420.

6. Son Tran The; Chando Lee; Hoa Le-Minh; Nauman Aslam; Moshin Raza; Nguyen Quoc Long, Year: 2020, "An evaluation of image-based malware classification using machine learning," Advances in Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30–December 3, 2020, Proceedings 12, Springer International Publishing.

7. O'Shaughnessy Stephen; Stephen Sheridan, Year: 2022, "Image-based malware classification hybrid framework based on space-filling curves," Computers & Security, Vol: 116, pp. 102660.

8. Rahali Abir; Arash Habibi Lashkari; Gurdip Kaur; Laya Taheri; Francois Gagnon; Frédéric Massicotte, Year: 2020, "DIDroid: Android malware classification and characterization using deep image learning," 2020 The 10th international conference on communication and network security.

9. Demirezen Mustafa Umut, Year: 2021, "Image Based Malware Classification with Multimodal Deep Learning," International Journal of Information Security Science, Vol: 10, no: 2, pp. 42-59.

10. Venkatraman Sitalakshmi; Mamoun Alazab; R. Vinayakumar, Year: 2019, "A hybrid deep learning image-based analysis for effective malware detection," Journal of Information Security and Applications, Vol: 47, pp. 377-389.

11. Demirezen Mustafa Umut, Year: 2021, "Image Based Malware Classification with Multimodal Deep Learning," International Journal of Information Security Science, Vol: 10, no: 2, pp. 42-59.

12. Singh Ajay; Anand Handa; Nitesh Kumar; Sandeep Kumar Shukla, Year: 2019, "Malware classification using image representation," Cyber Security Cryptography and Machine Learning: Third International Symposium, CSCML 2019, Beer-Sheva, Israel, June 27–28, 2019, Proceedings 3. Springer International Publishing.

13. Tang Mingdong; Quan Qian, Year: 2019, "Dynamic API call sequence visualisation for malware classification," IET Information Security, Vol: 13, no: 4, pp. 367-377.

14. Dang Dennis; Fabio Di Troia; Mark Stamp, Year: 2021, "Malware classification using long short-term memory models," arXiv preprint arXiv: 2103, pp. 02746.

15. Roseline S. Abijah; S. Geetha; Seifedine Kadry; Yunyoung Nam, Year: 2020, "Intelligent vision-based malware detection and classification using deep random forest paradigm," IEEE Access, Vol: 8, pp. 206303-206324.