IJIRMET ISSN (Online): 2456-0448
International Journal of Innovative Research in Management, Engineering, and Technology
Vol. 7, Issue 3, March 2022

A Novel Method For Strapping Malware Classification On Real Time Environments Using Deep Learning Techniques

[1] B.Varsha,[2]T.Hemalatha

[1 Student, [2]Assistant Professor, Department of CSE

[1][2]KrishnasamyCollegeofEngineeringandTechnology,Cuddalore,

***Abstract** In the recent years IOT becomes an emerging technology it provides high to accessibility wireless network and at the same time it produces privacy issues. The devices using IoT are needs to be connected to the internet constantly. Due to this, it becomes vulnerable to the malware which makes easier to the hackers to access the devices. Malware or malicious software is designed to cause the device or network. It introduce evasive techniques like polymorphic or metamorphic malware which can change its behavior frequently and it is difficult to predict . Malware classification can be carried out to understand about the malware by categorizing the datasets. The malware classification acts like a preventive technique to the malware behavior and it's types can be analyzed. To deal with security concerns deep learning techniques are launched. Deep learning works with neural networks which imitates how human train and gain knowledge. The datasets are trained and tested. The RNN- LSTM classifier is used for classification of malware because of its good prediction and the evaluation metrics are increased efficiently. The evaluation metrics such as accuracy, recall, precision, F1Score, and confusion matrix are evaluated with better performance.*

***Index Terms**—Accessibility, Malware, Metamorphic, Polymorphic, RNN-LSTM.*

1. INTRODUCTION

Recent technological developments in computers and internet provides convenience to the malware attacks. In the emerging IOT technologies allows the high accessibility to the user and at the same time malicious software invasion. The fast development of malware may leads to create threats.The security of the system is attacked and hacked by the malware programs. Malwares can steal confidential data in order to produce danger to the computer systems.

Malware can change its behaviours frequently in order to make the detectors fool and hack the data if incase the malware can be identified it can be grouped to its family which is known as malware classification.

In order to reduce the crucial works of human and intelligent systems are developed to understand the way the human behaves to the situation is called AI.

Machine Learning is a part of Artificial Intelligence in this machines are trying the based on the given dataset data set is large with the high number of input and output and since difficulty in solving it is a part of machine learning it works efficiently with minimum guidance and solved diamond analytic reduction problems and also it mimic the human brain. It provides many hidden layers in the hidden layer interconnect which each other with predicts based on the height probability they got output from the input.

The deep in the deep learning denotes that it contains many layers in the network to get high level features from raw data.In this paper, malware can be classified with combination of RNN-LSTM. The RNN is a generalization of feed forward neural network that has an internal memory. In the recurrent neural network every data is predicted based on its past values. To save the past data for long time LSTM is used to predict the large datasets. It allows to predict both past and future values.

In malware classification this hybrid approach provides easiest way to predict the datasets by training and testing the datasets. The dataset is converted into arrays by min-max scaler and sequential long information is passed to the RNN_LSTM network and this paper provides the effective evaluation metrics to measure the performance by using recall, precision, f-score and confusion matrix.

This paper focusses on increasing the accuracy with effective hybrid approach of RNN-LSTM algorithm as a classifier used to remember huge amount of data and for effective prediction.

The contributions of our project is,

(1) The datasets are collected from the CTC website containing malware and benign data samples.

(2) The pre processing of data is captured by using min-max scaler.

(3) The features are extracted and classifier makes a classification of dataset of malware and benign samples.

The objective of the project is to classify the malware and benign samples in order to analyze their dynamic changing behavior and to prevent the attacks by the malware in future. The RNN-LSTM classifier is better in predicting past and future values so that by analyzing the past behaviors of the malware and predict the system or internet.

This application is mainly used for classifying the malwares that mainly causes the system or internet by its attacks. Some of the existing techniques detecting that the malware is present or not but this would help to predict and act as a prevention techniques for the malware attacks. With the help of this we can able to protect resources from security concerns.

2.PROPOSED SYSTEM

Malware attacks can be significantly increased by using some evasive techniques such as changing its behavior rapidly to fool the system and acts as benign software. It becomes a crucial step if it is not properly identified as malware then it attacks the entire system or network. In the existing they used convolutional neural network as a classifier to classify the datasets and group them according to their belonging family but it shows some issues in classifying large amount of sequential data. The dataset may contain huge volume of malware and benign sample and classification plays a vital role to classify it. The data samples may be a files,email attachments or images. Here, we uses a files with both malware and benign sample data. The datasets are collected from IoT23. In order to train our neural network need a preprocessing step to convert it into a array and feature extraction used to remove the outliers and then introducing a RNN-LSTM classifier to classify the malware samples. The benefit of this classifier that can predict the past and future action by using gates. The forget gate plays a main role among the gates, if that particular data is needed it will move to the next step or else it will move out of it.

The evaluation metrics provides high efficient results in accuracy, precision, recall and confusion matrix using this classification approach.

2.1.SYSTEMARCHITECTURE

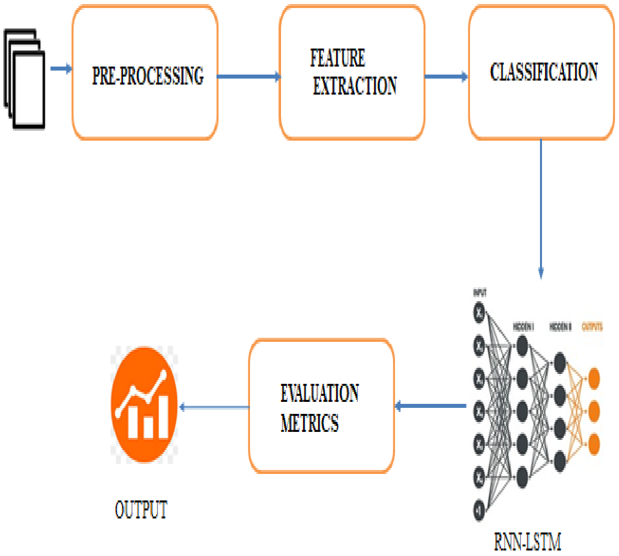


FIGURE 1 PROPOSED SYSTEM ARCHITECTURE

FIGURE 1: First collect the malware and benign data samples from the website named CTU. In that website we are collecting the datasets and it moved to the preprocessing stage in which the loaded data can uses EDA process and remove the unwanted or repeated data samples in the dataset.

The classifier RNN-LSTM classifies the seven types of malware to its family and denotes the appearance in it. The evaluation metrics is a final state to analyze the performance.

3.DATA REPRESENTATION AND LEARNING

3.1 DATASET

The datasets are collected from the website IOT 23 and are loaded on to the drive.The datasets are represented in CSV format. Large number of data used in the datasets in a tabular form this CSV format helps to read data efficiently.

3.2 PREPROCESSING

After gathering the data need to look after for pre processing in this we can split the data set into training testing and validating. The data may contain missing faulty values. To avoid this processing is done to reduce the noise it can be done by using feature extraction in order to predict the future values splitting of the data set into training and tested in training.

In training data set we can learn the machine by the data set done. By the train data set we can check whether the model is valid or not them all were can be arranged based on its behaviour to the family. The family consist of different models with the parameters research for the model it will provide accurate results after finishing training testing can be taken place in which the new object or applied the model makes prediction.

The preprocessing technique use min-max scaler method. The min-max scaler is used to preserve the shape of the dataset. The shape of the dataset is evaluated it consists of (34288,21). The scaler range from [0,1].

3.3 FEATURE EXTRACTION

In this process it can transform the raw data into the array format. The datasets may contain irrelevant data or repetition of data. Feature extraction is used to reduce the computing and processing of large number of data using dimensionality reduction by the name itself denote that the required features can be extracted and also it avoids repeated data.

The input is taken from the preprocessing and output will act us and input to the classification stage they helps to increase the speed of the mission and reduces taken by the machine. This appears at the beginning stage which makes use of a large dataset.

3.4 CLASSIFICATION

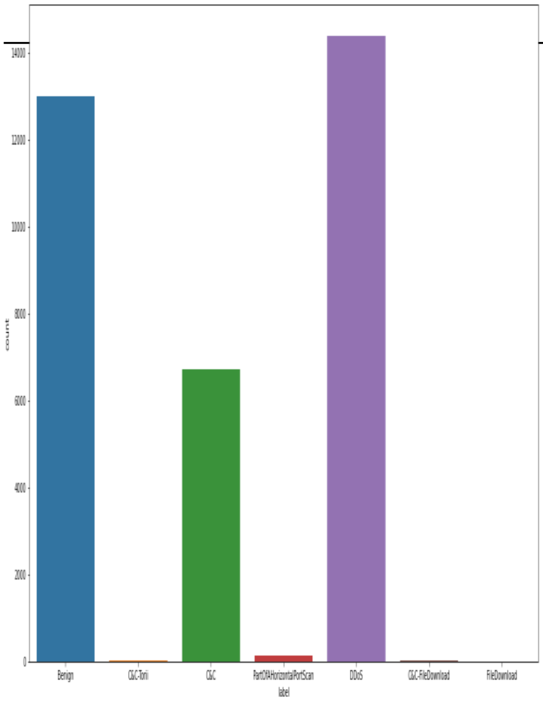
Classification is used to define the malware family by categorising based on its behaviour. Malware can be classified by using RNN-LSTM algorithm. RNN-LSTM algorithm makes benefits over the deep learning. RNN which is defined as Recurrent Neural Network which can take input values and remembers it then predict the next values which is suitable in time series. LSTM refers to long short term memory which usually enlarges the memory and it allows to kept for a long time. It allows to read, write, or delete based on the usage of gates. There can be three gates, Input gate, Output gate, Forget gate. The gate allows to decide whether it can moves to the next state. Forget gate acts as a terminator which removes the unwanted data from the memory. The output gate produces the output based on its previous action. The input gate takes the data and proceeds in the form of using sigmoid function as activation function. That activation function lets to manage whether the data can be passed through it or not by using 1's or 0's.

3.5 EVALUATION METRICS

Evaluation metrics is the method to evaluate the datasets about the presence of malware by using precision, recall, confusion matrix and F1 score. In precision, divides the sum of true values and false values by total number of samples present in the dataset.

In recall, which provides the way to high produce in true values in the dataset. If recall is increases with increase in the true value. In confusion matrix, it evaluates N*N matrix to measure targeted values based on the actual values. In F1 score, which is more useful than accuracy because it can manage both precision and recall in order to provide the most exact results in uncommon situations also.

4.RESULTS



The FIGURE 2 represents the seven families as label and based on the count value graph drawn.

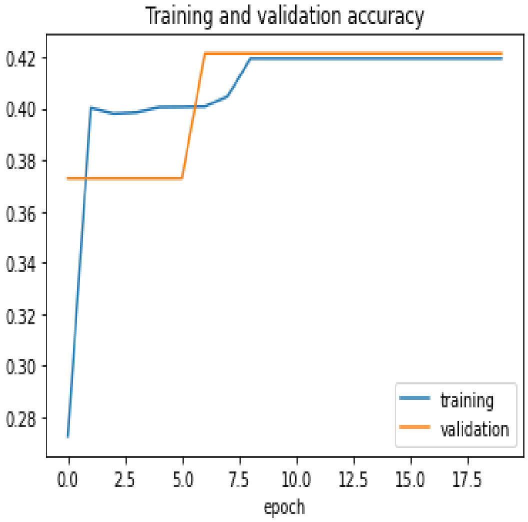


FIGURE 3 TRAINING AND VALIDATION

The FIGURE 3 represents the training and validation accuracy based on the epochs.

5.CONCLUSION AND FUTURE WORKS

The purpose of the work is to classify the malware based on hybrid deep learning algorithms. In this phase, identifying the unique values based on the gathered data are represented via graph. EDA process allows to visualize the information of the data from the dataset. The graph represents the information about 7 types of malware with an accuracy of 95%.

The future exploration includesclassifying the malware samples based on the RNN-LSTM classifier and the evaluation metrics are calculated. To increase the efficiency this classifier predict both past and future actions and enhance the performance.

6.REFERENCES

[1] H. Alasmary, A. Abusnaina, R. Jang, M. Abuhamad, A. Anwar, D. Nyang, and D. Mohaisen, “Soteria: Detecting adversarial examples in control flow graph-based malware classifier,” in 40th IEEE International Conference on Distributed Computing Systems, ICDCS, 2020.

[2] D. Park and B. Yener, “A survey on practical adversarial examples for malware classifiers,” 2020.

[3] A. Abusnaina, A. Khormali, H. Alasmary, J. Park, A. Anwar, and A. Mohaisen, “Adversarial learning attacks on graph-based iot malware detection systems,” in 39th IEEE International Conference on Distributed Computing Systems, ICDCS 2019, Dallas, TX, USA, July 7-10, 2019.

[4] A.Azmoodeh, A. Dehghantanha, and K.-K. R. Choo, “Robust malware detection for Internet Of (Battlefield) Things devices using deep eigenspace learning,” IEEE Transactions on Sustainable Computing,vol. 4, no. 1, pp. 88–95, 2019.

[5] S. Siby, R. R. Maiti, and N. O. Tippenhauer, “IoTScanner: Detecting privacy threats in IoT neighborhoods,” in Proceedings of the 3rd ACM International Workshop on IoT Privacy, Trust, and Security, 2017, pp.23–30.

[6] A New Malware Classification Framework Based on Deep Learning Algorithms Ömer Aslan 1 And Abdullah Asim Yilmaz 2019.

[7] Hybrid Malware Classification Method Using Segmentation-Based Fractal Texture Analysis and Deep Convolution Neural Network Features Maryam Nisa , Jamal Hussain Shah , ShansaKanwal , Mudassar Raza ,Muhammad Attique Khan , RobertasDamaševiˇcius and Tomas Blažauskas,2020.

[8] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, “On detecting adversarial perturbations,” in 5th International Conference on Learning Representations, ICLR, 2017.

[9] Y. M. P. Pa, S. Suzuki, K. Yoshioka, T. Matsumoto, T. Kasama, and C. Rossow, “IoTPOT: A novel honeypot for revealing current IoT threats,” Journal of Information Processing JIP, vol. 24, no. 3, pp. 522–533, 2016.

[10] X. Chen, C. Li, D. Wang, S. Wen, J. Zhang, S. Nepal, Y. Xiang, and K. Ren, “Android hiv: A study of repackaging malware for evading machine-learning detection,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 987–1001, 2019.