Malware Detection and Classification

Armaan Sait
Kumaraguru College of Technology
Coimbatore, India
armaan.17cs@kct.ac.in

Kiruthika J
Kumaraguru College of Technology
Coimbatore, India
kiruthika.17cs@kct.ac.in

Nivetha R
Kumaraguru College of Technology
Coimbatore, India
nivetha1.17cs@kct.ac.in

Chandrakala D
Kumaraguru College of Technology
Coimbatore, India
chandrakala.d.cse@kct.ac.in

***Abstract*—The Growth of Technology used on the Internet, Computers, Smartphones, and Tablets have been favourable to the emergence and spread of cyber threats, resulting in cyber attacks. The number of attacks has grown exponentially and has resulted in discovering various malware detection approaches. Multiple machine learning and big data technologies have been used for the detection of malware. Current malware detection solutions that adopt traditionalMachine Learning techniques take time but have been shown to be successful at detecting unknown malware in real time. The feature engineering process can be absolutely eliminated by employing advanced Machine Learning Algorithms such as deep learning.Various malware classification and identification methods are discussed in this paper. To identify the sample as benign or malware, machine learning and deep learning-based solutions have been addressed.**

***Keywords*—Cybersecurity, Malware detection, Machine learning, Ransomware**

I. INTRODUCTION

Cyberattack is the most common problem in current technology. It usually involves hacking into a vulnerable device and stealing, changing, or destroying a specific target.. They have become increasingly sophisticated and dangerous.

Malware is one such cyber attack.It refers to any programme that is created with the intent of causing harm to a device, server, client, or network.

Viruses, Trojan horses, ransomware, spyware, adware, rogue applications, wiper, scareware, and other forms of malware exist.. Programs are called malware if they perform actions against the authorization of the computer user.

Malware impacts on the digital systems also include:

- Disrupts operations.
- Steals sensitive information.
- Hardware Failure.
- Allowing unauthorized access to the system and all its resources.
- decreases computer or web browser speeds.
- Creates issues and problems connecting to networks.
- Results in crashing or frequent freezing of the system.

A Malware detection module determines if a program or connection is a threat, based on the collected and trained data. Preferably a machine-learning algorithm that can discover and formalize the principles that underlie the data it sees. A machine-learning algorithm is a program with a specific way to adjust its own parameters based upon the feedback on its previous performance in making predictions about a test dataset.

II. LITERATURE REVIEW

Daniel Gibert et al in [1] represented as images elucidates the conversion of PE Files to Images. then these are converted into grayscale images of (256 x 256). The maximum number of bits in the given image for a PE file is 256x256x8. Portable executable files that are malignant include four major types and benign are given as input. 8 bits are read at a time and converted to integer following which we get 65 KB if it's less than that then it's padded with zeros. Thus after this, a model was trained on these images which gets an accuracy of 66%. E files to Image.

Large-scale malware classification **[2]** using random projections and neural networks was published in 2013. Logistic regression and Neural Networks are primarily used. The input vector's dimensionality is reduced from 179 thousand to 4000 features using random projections. It is used to train 2.6 million labelled malware using neural networks, resulting in a 0.49 percent error rate for a single neural network and 0.42 percent for an ensemble of neural networks. Adding more hidden layers did not result in a major improvement in accuracy. As compared to the one-layer neural network, the two and three hidden layer models perform marginally worse. It's believed that the explanation for the poor results is that there aren't enough errors to properly learn additional layers.

A study in M. Ijaz, et al (2019) **[3]** explains the traditional machine learning approaches peculiarly in malware detection. The process is generally categorized into two groups depending on the type of analysis as static and dynamic analysis. Static analysis involves examining the malware sample without running it. On the other hand,

Dynamic analysis is carried out systematically in a controlled environment where the malware is executed in a controlled system.

Malware detection based on the deep learning algorithm was published in" The Natural Computing Applications Forum 2017" in July 2017. In this paper, they [4] portrayed malware as opcode sequences, used a deep belief network (DBN) to detect malware, and compared the performance of DBNs to three baseline malware detection models that used vector machines, decision trees, and the KNN algorithm. This system has a 96 percent accuracy rating.. They have proposed that using unlabeled data can improve the accuracy of malware detection models.

Robust Intelligent Malware Detection Using Deep Learning was published by IEEE Volume 7 of 2019. This paper evaluates various Machine learning Algorithms and Deep Learning algorithms for the categorization of multiple private and public datasets. **[5]** Experimental analysis is done to remove dataset bias using time scales. Finally, they propose an Image processing technique to arrive at an efficient zero-day model. By adopting CNN 2 layer + LSTM, they have reached an accuracy of 98.8%. In the work which was proposed, the robustness of the deep learning architectures was not being discussed.

Machine Learning Methods for Malware Detection was published by Kaspersky. It is a two-stage design **[6]** that reduces the number of false positives. In the first stage, unbiased regions are detected and second stage classifiers are trained only on a single bucket. Hence these regional classifiers used to detect the malware are quite efficient.

Bazrafshan, Z et al in **[7]** explains the three main strategies used in the detection of malware files: Signature based, Behavioral based, and Heuristic based detection. Here Pattern matching method and signature based detection techniques are widely used for malware detection. Data Collector- Interpreter- Matcher. The major benefits of Behaviour based malware detection techniques is the ability to detect unknown and polymorphic malware variants. Heuristic malware detection methods use various kinds of machine learning approaches to render the pattern of an executable file.

Malware and Detection Techniques: A Survey was published on 12 December 2013. In this paper **[8]**, they have compared the benefits and limitations of malware detection techniques. Various issues are discussed. **Signature-based** approaches can't identify unknown malware variants, and extracting specific signatures takes a lot of manpower, time, and resources. It was impossible to identify mutated codes**. Behaviour-based** approaches have a high scanning period and a non-availability of promising False Positive Ratio (FPR). Handling a large number of genera is one of the heuristic features constraints.

Malware Classification with Deep Convolutional Neural Networks was published in 2018(IEEE). This paper targets **[9]** convolution neural networks used to classify malware samples. These binary samples are first converted to Grayscale images through visualization. They have developed a CNN model and trained the same. This method has achieved 98.52% and 99.97% on Microsoft datasets that are available for public use.

Gupta, P et al **[10]** elucidated the hybrid approach to malware detection, which incorporates static and dynamic analysis aspects. Experts believe that AI-powered anti-malware software would aid in the detection of new malware attacks and the improvement of scanning engines. Neural networks have recently made a name for themselves in learning features from raw inputs in a variety of fields.

III. MALWARE DATA ANALYSIS

A. *Data Description*

The data was taken for entire analytics obtained from the Canadian Institute for Cybersecurity. Dataset possesses pcap files with log details of applications under each malware category. These log features can be extracted and thus trained for various models. The samples come from 42 unique malware families. It contains over 10,854 samples (4,354 belong to malware class while 6,500 belong to benign class) from several sources. Malware types include Ransomware, Adware, Scareware, and SMS malware.

B. Exploratory Data Analysis

This entire project was carried out using Python. Importing all necessary libraries for model deployment includes Pandas, sci-kit learn, seaborn, Matplotlib, NumPy, TensorFlow. The dataset contains 88 attributes based on the structure and network configuration of malware. The correlation matrix for all the features has been drawn and based on that, certain columns are dropped. The entire dataset is truncated to 77 attributes. A Label encoder was used to mutate the Non-numerical data into a machine-readable format for easy interpretation by the system. Normalization was carried out on Flow ID, Source IP, Destination IP, Label, and source values.

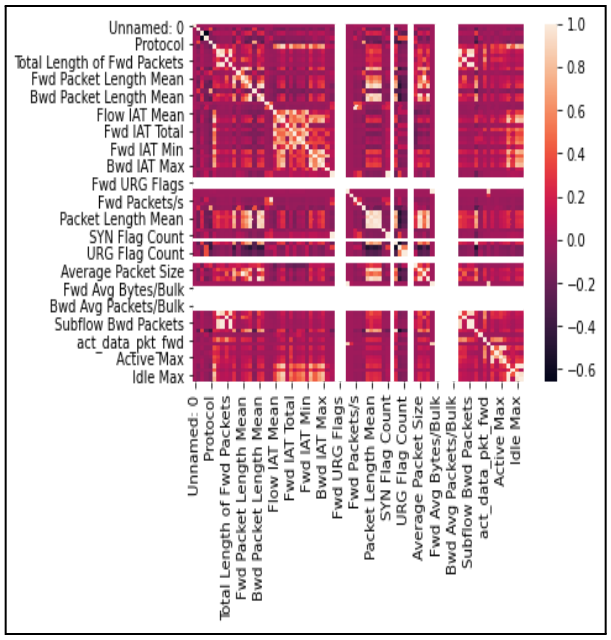

Fig.1 Correlation Matrix

C. *Predictive Data Analysis*

The robust classification of the malware type is imperative to backtrack the cybersecurity threats. The comparative analysis of all algorithms and fine-tuning methods are given. Since the given dataset is abstract and highly skewed, Multi-Layer Perceptron is deployed with 10 layers and an input of 72 attributes. The Activation function used here is Sigma and optimizer as adam and the loss function is Mean Absolute Error (MAE). The accuracy of the MLP obtained was 29.6 percent.

Then, Machine Learning techniques like random forest classifier and KNN (K Nearest Neighbour) are implemented, accuracy about 50 percent and 39.7 percent are attained. Train and test split ratio are given as 70:30, The number of attributes passed is the same as deployed in the MLP model.

Fine-tuning the Random Forest algorithm with K-fold, grid search results in increased accuracy of about 62.2 percent. Further, the algorithm is tuned by adjusting the optimization parameters like max depth with an increased number of layers. Finally, 95.6 percent accuracy has been achieved through the Random Forest classification method.

D. Results and Discussion

TABLE I. PREDICTIVE RESULTS

Result Overview	**Model**	**Accuracy score (Percentage)**
	Multi-Layer Perceptron	29.6
	K Nearest Neighbour	37.9
	Random Forest classifier (Before tuning)	62.2
	Random forest classifier (after tuning)	95.6

Data is the primary source for all applications functioning in the digital world. The measures to be taken to protect the data from many cybersecurity threats are crucial. Indeed, Machine Learning techniques are an inventive approach that paves the way for accurate prediction, although feature selection is a challenging task to build any kind of model. The algorithm which can also process irregular data has to be circumvented.

Future attacks can be controlled and stopped only by analyzing the malware samples and its behaviour that exists now. This is done by cybersecurity experts by making use of some professional malware analysis tools.

Cybersecurity teams can use a malware detection tool to identify and analyse malware samples and see whether they are malicious or not, and if they are, they can be removed from the system and prevented from spreading further. These tools can be used to monitor security alerts and avoid malware attacks in the future. Organizations are adopting new security measures as malware attack vectors become more sophisticated..

REFERENCES

[1] Daniel Gibert, Carles Mateu, Jordi Planes & Ramon Vicens, “Using convolutional neural networks for classification of malware represented as images” -Journal of Computer Virology and Hacking Techniques volume.

[2] George E. Dahl, Jack W. Stokes, Li Deng, Dong Yu, Microsoft Research,” Large-scale Malware Classification Using Random Projections And Neural Networks”-IEEE.

[3] M. Ijaz, M. H. Durad and M. Ismail, "Static and Dynamic Malware Analysis Using Machine Learning," 2019 16th International Bhurban Conference on Applied Sciences and Technology

(IBCAST), Islamabad, Pakistan, 2019, pp. 687-691, doi: 10.1109/IBCAST.2019.8667136.
[4] Ding Yuxin, Zhu Siyi, “Malware detection based on the deep learning algorithm”, Springer.
[5] R. Vinaya Kumar, Mamoun Alazab, K.P.Soman, Prabaharan Poornachandran, And Sitalakshmi Venkatraman, “Robust Intelligent Malware Detection Using Deep Learning”, IEEE.
[6] “Machine Learning Methods for Malware Detection”, Kaspersky, https://media.kaspersky.com/en/enterprise-security/Kaspersky-Lab-Whitepaper-Machine-Learning.pdf.
[7] Bazrafshan, Z., Hashemi, H., Fard, S. M. H., & Hamzeh, A. (2013),” A survey on heuristic malware detection techniques.”

[8] “Malware and Malware Detection Techniques: A Survey”, Jyoti Landage, Prof. M. P. Wankhade Professor, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 12, December - 2013
[9] Mahmoud Kalash, Mrigank Rochan, Noman Mohammed, Neil D. B. Bruce, Yang Wang, Farkhund Iqbal, “Malware Classification with Deep Convolutional Neural Networks”, -IEEE.
[10] S. Gupta, P. Bansal, and S. Kumar, "ULBP-RF: A Hybrid Approach for Malware Image Classification," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India, 2018, pp. 115-119, doi: 10.1109/PDGC.2018.8745989.
[11] Qublai K. Ali Mirza, Fatima Hussain, Irfan Awan, Muhammad Younas, Salah Sharieh, "Taxonomy-Based Intelligent Malware Detection Framework", Global Communications Conference (GLOBECOM) 2019 IEEE, pp. 1-6, 2019.
[12]Ahmad Azab, Mahmoud Khasawneh, "MSIC: Malware Spectrogram Image Classification", Access IEEE, vol. 8, pp. 102007-102021, 2020.