International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 01 | Jan 2024 www.irjet.net p-ISSN: 2395-0072

INTELLIGENT MALWARE DETECTION USING EXTREME LEARNING MACHINE

G MAHESH CHALLARI [1], P SWAPNA [2], T SOUMYA [3]

[1, 2 & 3] *Assistant Professor in Department of cse at Sree Dattha Institute of Engineering & Science*

Abstract - *Security breaches due to attacks by vicious software (malware) continue to escalate posing a major security concern in this digital age. With multitudinous computer stoners, pots, and governments affected due to an exponential growth in malware attacks, malware discovery continues to be a hot disquisition content. Current malware discovery results that adopt the static and dynamic analysis of malware signatures and gets patterns are time consuming and have proven to be ineffective in relating unknown malwares in real- time. Recent malwares use polymorphic, metamorphic, and other fugitive ways to change the malware conduct snappily and to induce a large number of new malwares. Analogous new malwares are generally variants of being malwares, and machine knowledge algorithms (MKAs) are being employed recently to conduct an effective malware analysis. Therefore, this work proposes the combined visualization and deep knowledge architectures for static, dynamic, and image processing predicated crossbred approach applied in a big data terrain, which is the first of its kind toward achieving robust intelligent zero- day malware discovery. Overall, this work paves way for an effective visual discovery of malware using a scalable and cold- thoroughbred extreme knowledge machine model named as ELM Net for real- time deployments*

Key Words: Naive Bayes, DNN, Deep Learning, FNN, Protocol.

1. INTRODUCTION

In this digital world of Assiduity4.0, the rapid-fire advancement of technologies has affected the diurnal conditioning in businesses as well as in particular lives. Internet of effects (IoE) and operations have led to the development of the ultramodern conception of the information society. Still, security enterprises pose a major challenge in realizing the benefits of this artificial revolution as cyber miscreant's attack individual PC's and networks for stealing nonpublic data for fiscal earnings and causing denial of service to systems. Similar bushwhackers make use of vicious software or malware to beget serious pitfalls and vulnerability of systems (1). The major challenge in similar classical approaches is that new variants of malware use antivirus elusion ways similar as law obfuscation and hence similar hand- grounded approaches are unfit to descry zero- day malwares(2). Hand- grounded malware discovery system requires expansive sphere position knowledge to reverse mastermind the malware using Static and dynamic analysis and to assign a hand for that. Also, hand- grounded system requires larger time to reverse mastermind the malware and during that time a bushwhacker would worm into the system. In addition, hand- grounded system fails to descry new types of malware.

2. LITERATURE REVIEW

Machine Literacy Algorithms calculate on the point engineering, point selection and point representation styles. The set of features with a corresponding class is used to train a model in order to produce a separating aero plane between the benign and malwares. This separating aero plane helps to descry a malware and classify it into its corresponding malware family. Both point engineering and point selection styles bear sphere position knowledge. The colorful features can be attained through stationary and dynamic analysis. Stationary analysis is a system that captures the information from the double program without executing. Dynamic analysis is the process of covering malware gets at run time in an isolated terrain. The complications and colorful issues of Dynamic analysis are bandied in detail by (10). Dynamic analysis can be an effective long- term result for malware discovery system. The Dynamic analysis cannot be stationed in end- point real time malware discovery due to the reason that it takes important time to dissect its gets , during which vicious cargo can get delivered. Malware discovery styles grounded on Dynamic analysis are more robust to obfuscation styles when compared to statically collected data. Utmost generally, the market able anti-malware results use a mongrel of Static and Dynamic analysis approaches. The major issue with the classical machine literacy grounded malware discovery system is that they calculate on the point engineering, point literacy and point representation ways that bear an expansive sphere position knowledge(11),(12),(13).

Also, once a bushwhacker comes to know the features, the malware sensor can be finessed fluently (14). To be successful, MLAs bear data with a variety of patterns of malware. The intimately available standard data for malware analysis exploration is veritably less due to the security and sequestration enterprises. Though many datasets live, each of them has their own harsh examines as utmost of them are outdated. Numerous of the published results of machine literacy grounded malware analysis have used their own datasets. Indeed however intimately available sources live to

crawl the malware datasets, preparing a proper dataset for exploration is a daunting task. These issues are the main downsides behind developing general machine literacy grounded malware analysis system that can be stationed in real time. More importantly, the compelling issues in applying data wisdom ways were bandied in detail by (15).

3. SYSTEM ANALYSIS

3.1.1 Naive Bayes

Naive Bayes algorithm is a probabilistic literacy system that's substantially used in Natural Language Processing (NLP). The algorithm is grounded on the Bayes theorem and predicts the label of a textbook similar as a piece of dispatch or review composition. It calculates the probability of each label for a given sample and also gives the label with the loftiest probability as affair. Naive Bayes classifier is a collection of numerous algorithms where all the algorithms partake one common principle, and that's each point being classified isn't related to any other point. The presence or absence of a point doesn't affect the presence or absence of the other point. Naive Bayes is an important algorithm that's used for textbook data analysis and with problems with multiple classes. To understand Naive Bayes theorem's working, it's important to understand the Bayes theorem conception first as it's grounded on the ultimate. Bayes theorem, formulated by Thomas Bayes Eq (1), calculates the probability of an event being grounded on the previous knowledge of conditions related to an event. It's grounded on the following formula

$$(A|B) = (A) * P(B|A)\, P(B) \text{---------- Eq(01)}$$

Where we're calculating the probability of class A when predictor B is formerly handed.

P (B) = previous probability of B

P (A) = previous probability of class A

P (B| A) = circumstance of predictor B given class A probability

3.1.2 Deep Learning

Deep Literacy or deep neural networks (DNNs) takes alleviation from how the brain works and forms a sub module of artificial intelligence. The main strength of deep literacy infrastructures is the capability to understand the meaning of data when it's in large quantities and to automatically tune the deduced meaning with new data without the need for a sphere expert knowledge. Convolutional neural networks (CNNs) and intermittent neural networks (RNNs) are two types of deep literacy infrastructures generally applied in real- life scripts. Generally, CNN infrastructures are used for spatial data and RNN infrastructures are used for temporal data. The combination of CNN and LSTM is used for spatial and temporal data analysis.

3.1.2.1 DNN

A feed forward neural network (FFN) creates a directed graph in which a graph is composed of bumps and edges. FFN passes information along edges from one knot to another without conformation of a cycle. Multi-layer perceptron (MLP) is a type of FFN that contains 3 or further layers, specifically one input subcase, one or further retired subcase and an affair subcase in which each subcase has numerous neurons, called as units in fine memorandum. The number of retired layers is named by following a hyperactive parameter tuning approach. The information is converted from one subcase to another subcase in forward direction without considering the history values. Also, neurons in each subcase are completely connected.

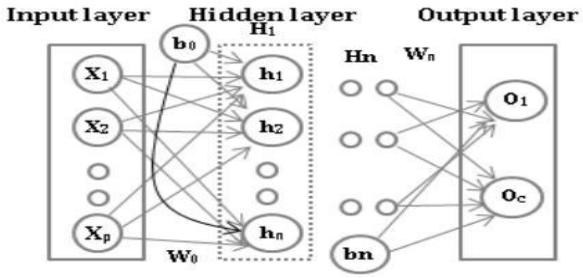


Fig. 1: DNN architecture

3.1.2.2 CNN

Convolutional network or convolutional neural network or CNN is supplement to the classical feed forward network (FFN), primarily used in the field of image processing. It's shown in Figure 2, where all connections and retired layers and its units aren't shown. Then, m denotes number of pollutants, ln denotes number of input features and p denotes reduced point dimension, it depends on pooling length. In this work, CNN network composed of complication 1D subcase, pooling 1D subcase and completely connected subcase. A CNN network can have further than one complication 1D subcase, pooling 1D subcase and completely connected subcase. In convolutional 1D subcase, the pollutants slide over the 1D sequence data and excerpts optimal features. The features that are uprooted from each sludge are grouped into a new point set called as point chart. The number of pollutants and the length are chosen by following a hyperactive parameter tuning system. This in turn uses on-linear activation function, ReLU on each element. The confines of the optimal features are reduced using pooling 1D subcase using either maximum pooling, min pooling or average pooling.

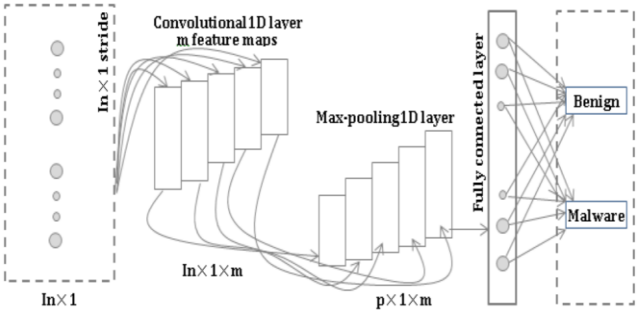


Fig. 2: Architecture of CNN for malware detection.

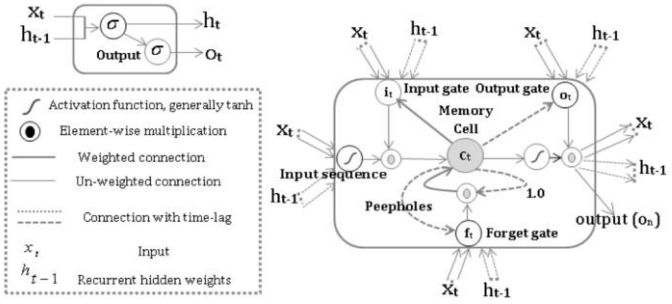


Fig. 3: Architecture of RNN unit (left). LSTM memory block (right).

4. PROPOSED METHODOLOGIES

Deep Literacy or deep neural networks (DNNs) takes alleviation from how the brain works and forms a sub module of artificial intelligence. The main strength of deep literacy infrastructures is the capability to understand the meaning of data when it's in large quantities and to automatically tune the deduced meaning with new data without the need for a sphere expert knowledge. Convolutional neural networks (CNNs) and intermittent neural networks (RNNs) are two types of deep literacy infrastructures generally applied in real- life scripts. Generally, CNN infrastructures are used for spatial data and RNN infrastructures are used for temporal data. The combination of CNN and LSTM is used for spatial and temporal data analysis.

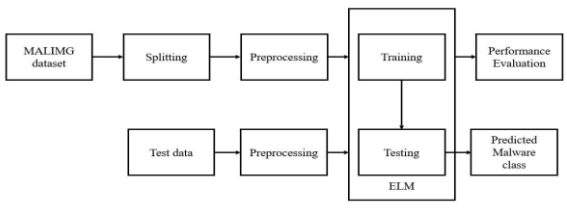


Fig. 4: Block Diagram of Proposed System

4.1.1 MALIMG dataset

CICDDoS2019 contains benign and the most over- to- date common DDoS attacks, which resembles the true real- world data. It also includes the results of the network business analysis using CICFlowMeter- V3 with labeled overflows grounded on the time stamp, source, and destination IPs, source and destination anchorages, protocols and attack (CSV lines). Generating realistic background business was our top precedence in erecting this dataset. We've used our proposed B- Profile system to outline the abstract gets of mortal relations and generates natural benign background business in the proposed testbed. For this dataset, we erected the abstract gets of 25 druggies grounded on the HTT, HTTPS, FTP, SSH and dispatch protocols.

Need of Data Preprocessing A real- world data generally contains noises, missing values, and maybe in an useless format which can't be directly used for machine knowledge models. Data preprocessing is demanded tasks for drawing the data and making it suitable for a machine knowledge model which also increases the delicacy and effectiveness of a machine knowledge model. Getting the dataset.

1. Handling Missing data

2. Garbling Categorical Data.

3. Point scaling

Encoding Categorical data: Categorical data is data which has some orders similar as, in our dataset; there are two categorical variables, Country, and Bought. Since machine literacy model fully works on mathematics and figures, but if our dataset would have a categorical variable, also it may produce trouble while erecting the model. So, it's necessary to render these categorical variables into figures.

Feature Scaling: Point scaling is the final step of data preprocessing in machine literacy. It's a fashion to regularize the independent variables of the dataset in a specific range. In point scaling, we put our variables in the same range and in the same scale so that no variable dominates the other variable. A machine literacy model is grounded on Euclidean distance, and if we don't gauge the variable, also it'll beget some issue in our machine learning model. Euclidean distance is given

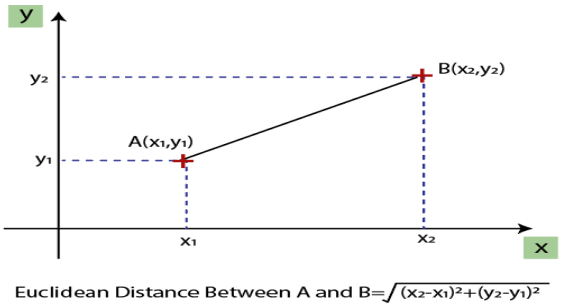
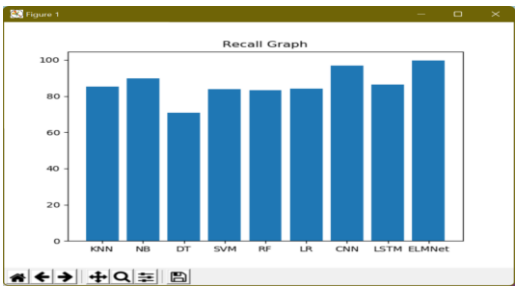
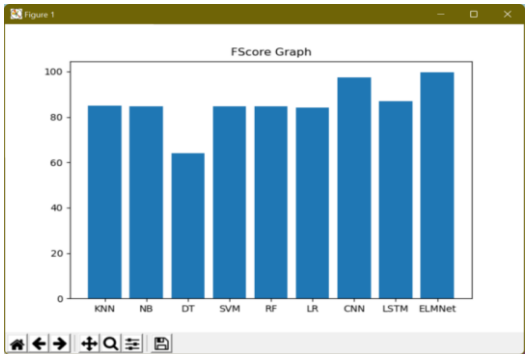


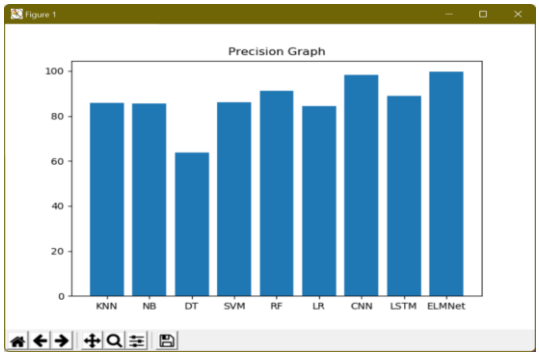
Fig. 5: Feature scaling



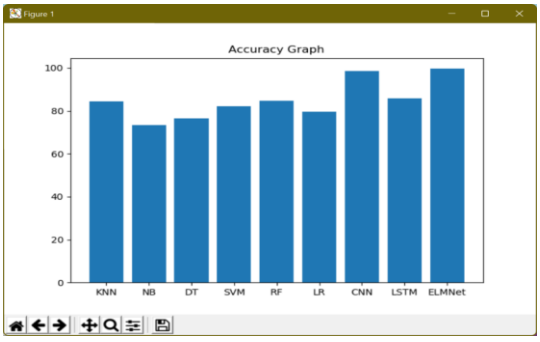
(a) Recall Graph



(b) F-Score Graph



Precision Graph



Screenshot 6.5-Acuracy Graph

Graph/ Algorithm	Recall Graph	Fscore Graph	Precision Graph	Accuracy Graph
KNN Algorithm	85.1	84.8	85.7	84.3
Naïve Bayes	89.8	84.5	85.3	73.2
CNN Algorithm	96.9	97.5	98.2	98.4
ELMNET	99.6	99.6	99.6	99.6

Extreme Learning Machines (ELMs) are a type of machine learning algorithm that originated in the field of neural networks. ELMs are known for their fast learning speed and computational efficiency compared to traditional gradient-based learning algorithms. They achieve this by randomly initializing the parameters of the hidden layer neurons and solving the output weights analytically.

Compared to traditional grade- grounded algorithms since they don't bear an iterative optimization process.

1. This can be profitable when dealing with large datasets.

2. Computational effectiveness ELMs have shown good computational effectiveness, making them suitable for real-time or high- speed malware discovery operations.

3. Conception Capability ELMs have been observed to have good conception capabilities, allowing them to perform well on unseen malware samples during the testing phase.

4. Resistance to Overfitting ELMs have been reported to parade better resistance to overfitting, which occurs when a model becomes too technical to the training data and performs inadequately on new data.

5. Scalability ELMs can handle large- scale datasets effectively due to their effective literacy process, making them potentially suitable for assaying expansive malware collections.

However, it's important to note that the effectiveness of any malware detection method, including ELMs, depends on various factors such as the quality and diversity of the training dataset, the choice of features, and the ability to adapt to emerging malware techniques. It's possible that there have been further advancements and refinements in the field of intelligent malware detection since my knowledge cutoff, so it's worth exploring the latest research papers and publications to gather up-to-date information on the performance of ELMs for this specific application.

5. CONCLUSION

To apply this design and to estimate machine literacy algorithms performance author is using double malware dataset called 'MALIMG '. This dataset contains 25 families of malware and operation will convert this double dataset into argentine images to induce train and test models for

machine literacy algorithms. These algorithms converting double data to images and also generating model, so they're called as Malcom CNN and Malcom LSTM and other algorithm refers as EMBER. Operation convert dataset into double images and also used 80 dataset for training model and 20 dataset for testing. Whenever we upload new test malware double data also operation will apply new test data on train model to prognosticate malware class

REFERENCES

1] R. Anderson et al., ''Measuring the cost of cybercrime,'' in The Economics of Information Security and Privacy. Berlin, Germany: Springer, 2013, pp. 265–300.

[2] B. Li, K. Roundy, C. Gates, and Y. Vorobeychik, ''Large-scale identification of malicious singleton files,'' in Proc. 7th ACM Conf. Data Appl. Secur. Privacy. New York, NY, USA: ACM, Mar. 2017, pp. 227–238.

[3] M. Alazab, S. Venkataraman, and P. Watters, ''Towards understanding malware behaviour by the extraction of API calls,'' in Proc. 2nd Cybercrime Trustworthy Comput. Workshop, Jul. 2010, pp. 52–59.

[4] M. Tang, M. Alazab, and Y. Luo, ''Big data for cybersecurity: Vulnerability disclosure trends and dependencies,'' IEEE Trans. Big Data, to be published.

[5] M. Alazab, S. Venkatraman, P. Watters, and M. Alazab, ''Zero-day malware detection based on supervised learning algorithms of API call signatures,'' in Proc. 9th Australas. Data Mining Conf., vol. 121. Ballarat, Australia: Australian Computer Society, Dec. 2011, pp. 171–182.

[6] M. Alazab, S. Venkatraman, P. Watters, M. Alazab, and A. Alazab, ''Cybercrime: The case of obfuscated malware,'' in Global Security, Safety and Sustainability & e-Democracy (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), vol. 99, C. K. Georgiadis, H. Jahankhani, E. Pimenidis, R. Bashroush, and A. Al-Nemrat, Eds. Berlin, Germany: Springer, 2012.

[7] M. Alazab, ''Profiling and classifying the behavior of malicious codes,'' J. Syst. Softw., vol. 100, pp. 91–102, Feb. 2015.

[8] S. Huda, J. Abawajy, M. Alazab, M. Abdollalihian, R. Islam, and J. Yearwood, ''Hybrids of support vector machine wrapper and filter based framework for malware detection,'' Future Gener. Comput. Syst., vol. 55, pp. 376–390, Feb. 2016.

[9] E. Raff, J. Sylvester, and C. Nicholas, ''Learning the PE header, malware detection with minimal domain knowledge,'' in Proc. 10th ACM Workshop Artif. Intell. Secur. New York, NY, USA: ACM, Nov. 2017, pp. 121–132.

[10] C. Rossow, et al., ''Prudent practices for designing malware experiments: Status quo and outlook,'' in Proc. IEEE Symp. Secur. Privacy (SP), Mar. 2012, pp. 65–79.