iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** III **Month of publication:** March 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

A Survey on Malware Classification using Supervised Learning Methods

Afrah Mehar[1], Ajini A[2]
[1]*Mtech IT student, Dept.of IT, GEC Idukki,*
[2]*Assistant Professor, Dept. of IT, GEC Idukki*

Abstract: ***Machine learning techniques contribute so much to these dynamic detection of malware. Machine learning helps in dynamic analysis of malware signatures. Different studies and papers are proposed for this classification of malware using Supervised Learning methods. Some of these methods are high performance in detecting malware. Using Covolutional Neural Network(CNN) this detection process made more accurate. These CNN has a detailed feature extraction capability which made it more accurate. Some new architectures are proposed by authors which consist a step wise feature extraction and classification of malware. This paper discuss about some previous methods that used Supervised Learning techniques to detect malware.***
Keywords: CNN, Supervised Learning, Malware

I. INTRODUCTION

Different malwares like Internet worms, computer viruses and Trojan horses cause a major threat to the security of networked systems. The rapid rise of the Internet and the resulting increase in malware meant that manually generated detection rules were no longer feasible and modern, sophisticated security technologies were needed. But the Classical signature based detection methods are not applicable in today's scenario. The anti-virus industry has implemented traditional methods, like hash-based, signature-based, and heuristic-based detection techniques to detect malware. But each of these has its own collection of disadvantages that restrict its ability to detect malware efficiently. Due to the dynamic nature of this malware there is essential for a better detection strategy.

Machine learning techniques contribute so much to these dynamic detection of malware. Machine learning helps in dynamic analysis of malware signatures. Current malware detection solutions adopt Static and Dynamic analysis of malware signatures and behavior patterns that are time consuming and ineffective in identifying unknown malware. Recent malwares use different methods to change the malware behaviors quickly and to generate large number of malwares. Since new malware are generated from existing malware and they are variants of existing ones so machine learning algorithms are being employed recently to conduct an effective malware analysis.

II. SUPERVISED LEARNING METHODS IN MALWARE DETECTION

Machine learning algorithm explores the concepts underlying the data it sees and formalizes them. Using this knowledge the algorithm can think the properties of samples that were previously unseen. A sample previously encountered in malware detection may be a new file. Its secret property can be either malicious or benign.The model is considered a mathematically formalized set of principles that underlie the data properties. Machine learning encompasses a wide range of solutions to a solution rather than a single process. Such approaches have different capacities and tasks which best suits them.Malware detectors based on signatures will work well on previously identified malware, which some vendors of antivirus have already discovered. Nevertheless, it cannot detect polymorphic malware that has the potential to alter its signatures, as well as new malware that has not yet been developed for signatures.

From the point of view of machine learning, malware detection can be seen as a classification or clustering problem: unknown forms of malware should be clustered into multiple clusters, based on some properties that the algorithm is detecting. At the other hand, we can reduce this problem to classification, having trained a model on the large dataset of malicious and benign files.

III.LITERATURE REVIEW

Konrad Rieck et al. [1] proposed Learning and Classification of Malware,Proposed a method that proceeds in three stages first one is behavior of collected malware is monitored in a sandbox environment second is based on a corpus of malware labeled by an AV scanner a malware classifier is trained using learning techniques and finally differentiated features of the behavior models are ranked for explanation of classification decisions. This learning approach of the proposed method consist of different steps.

1) *Data Acquisition:* Different malware binaries currently spreading in the wild is collected using a variety of techniques, such as honeypots and spam-traps. An antivirus engine is used to identify known malware instances and to enable learning and classification of family-specific behaviour.
2) *Behavior Monitoring.* Malware binaries are monitored in a sandbox environment. Based on state changes, a behavior based analysis report is generated.
3) *Feature Extraction.* Features reflecting behavioral pat- terns are extracted from the analysis reports and embed the malware behavior into a high-dimensional vector space.
4) *Learning and Classification.* Machine learning techniques are enforced for identifying the shared behavior of each malware family. Finally, a mixed classifier for all families are constructed and applied to different testing data.
5) *Explanation.* The segregated model for each malware family is analyzed using the weight vector expressing the donation of behavioral patterns. The most prominent patterns bring in insights into the classification model and expose relations between malware families.

This proposed method can give over 3,000 previously un- detected malware binaries, our system correctly predicted almost 70% of labels assigned by an anti-virus scanner four weeks later. This method also detects anonymous behavior, so that malware families not present in the learning corpus are correctly identified as unknown.

Stefano Schiavoni et al. [2] proposed work has focused on the analysis of DNS traffic to identify botnets based on their DGAs. Here proposed method is Phoenix. This mechanism can differentiate DGA generated and non-DGA domains using a mix of string and IP-based features, characterizes the DGAs behind them, and also finds groups of DGA-generated domains that are representative of the respective botnets. It contains discovery module, detection module and Intelligence and Insights module. The Detection module will get one or more domain names with the corresponding DNS traffic, and uses the models developed by the Discovery module to tell whether such domain names appear to be automatically developed. The Detection module evaluate a stream of DNS traffic and notice the (previously unknown) domains that resemble a known DGA. The Intelligence and Insights module contribute the analyst with information useful, for instance, to track a botnet. The system contain steps as Filtering, Clustering, Fingerprinting and Detection Module. filtering step works like,DGA-generated domains show a certain degree of linguistic randomness, as numerous samples of the same randomized algorithm exist. The main aim of clustering step is to cluster the domains according to their similarity. This proposed method has an advantage as successfully used in real-world settings to find malicious domains. But have a limitation like results in a less-responsive detection of previously unseen DGAs.

Zhixing Xu et al. [3] introduce a new framework for hardware-assisted malware detection based on tracking and classifying memory access patterns using machine learning. Here for the system call epoch for kernels, do feature election using F-score, a commonly used measure of the discriminative power of features.

TABLE I
Comparison Of **Cnn** And **Lstm**

Method	TPR	FPR
CNN	72.89%	0.31%
LSTM	74.05%	0.54%

Here selected the top 10% F-score features for the training phase. Then proposed a classifier architecture that consider two design points for the classifiers. First one is Direct Classification of Memory Access Histograms, this performs binary classification directly on the histograms computed in each epoch. This was effective in detecting kernel- level malware. The second one is Weighted Classification of Memory Access Histograms, for user-level malware the epoch boundaries correspond to function calls, which unlike system calls are not limited in number and are different across programs. Therefore it's very difficult to identify malware by analyzing the histograms. The proposed solution is to classify the different functions in a program separately and consider a weighted sum of classification results. The framework has a detection rate of 99% but here consider histogram bin size, it needs to be chosen carefully.

Bashari et al. [4] proposed method that concentrate on the investigation and implementation of a neural network binary malware classifier. This classifier can classify an unseen file as malicious or benign. The aim of the method has been limited down to classify Windows Portable Executable (PE) files based on their imported library function calls. proposed solution collects different malicious samples. The obtained samples are initially placed into their respective folders, for instance, make a folder called "malicious samples" and the malicious samples will place into it. all the benign samples will be placed into a folder called "benign

samples". In feature extraction the imported Windows library function calls for each sample need to be excerpted. These can be represented within a binary feature vector matrix. This model give very promising results, as they signify the model's ability to generalize against an independent set.But this paper focused on samples that were not packed with known packers, it affect the quality of the model due to the usage of static features.

B. Yu et al. [5] Proposed and explored a way of applying supervised learning to DGA detection in real time. Here used simple filtering steps to obtain sufficiently pure DGA/non- DGA samples from real DNS traffic .This paper also leveraged deep learning for the benefits of automated extraction of features and the ability for online learning to keep up with changes in the trends of DGA domain. This proposed methods models for a target false positive rate as low as 0.01%. It obtained a better performance and it collected large volume of data from real traffic as DGA/non-DGA. But it used only fixed width padded input. This paper compares the CNN and LSTM method for DGA domain classification, from that it found LSTM has more TPR(True Positive Ratio) than CNN.Table show the different TPR and FPR values of CNN and LSTM.

TABLE II
Detection Rates Of Different Methods

Method	AUC
Recurrent SVM	0.9969
Bidirection LSTM	0.9964
CNN+LSTM	0.9959
LSTM	0.9955

Hieu et al. [6] investigated various supervised learning methods, including Hidden Markov Model, C4.5 decision tree, Support Vector Machines, Extreme Learning Machine, Long Short-Term Memory network, Recurrent SVM, CNN+LSTM and Bidirectional LSTM for detection of DGA. Table II show the detection rate achieved by these methods. The highest de- tection rate is achieved by the Recurrent SVM is AUC=0.9969. Followed by the Bidirectional LSTM achieved AUC=0.9964,CNN+LSTM achieved AUC=0.9959. But using these super- vised learning methods eight DGA malwares not able to detect.

Alazab et al. [7] proposed an architecture Scale MalNet, it's a malware analysis system that can collect data internally from various data resources and uses self-learning techniques such as classical machine learning, deep learning and image processing techniques to detect, classify and categorize malware to their corresponding malware family accurately. The framework is highly scalable which facilitates collection of malware samples from different sources and applies preprocessing in a distributed way. The framework incorporates self-learning techniques to malware analysis such as malware detection, classification and categorization. The performance of deep learning architectures are evaluated over classical machine learning algorithms (MLAs) and an improvement in performance is observed consistently. But the proposed work not discussed the robustness of deep learning architectures.

Y. Li et al. [8] proposed the machine learning framework with the development of a deep learning model to handle DGA threats. The proposed machine learning framework consists of a dynamic blacklist, a feature extractor, a two-level machine learning model for classification and clustering, and a prediction model, this model have Detection speed is high and provide effective solution for the data imbalance problem among different malware family. But this proposed method have a drawback, as size of data increase it effect performance.

Cui et al. [9] Proposed an advanced method for detection of malware. Here a Deep Learning method is used which has a better performance. It translates the malicious codes into gray scale image and then given this image to a deep learning model. This model can automatically extract the features of image. Based on these a classification performed. This method uses image as input to the CNN has achieved 94.5% of accuracy. And precision of 94.6% with good detection speed.

TABLE III
Malware Classification Results Using Different Classifiers

Classification algorithms	Accuracy
K-Nearest Neighbor	*96.46%*
Decision Tree	*99.11%*
Support Vector Machine	*86.72%*
Random Forest	*88.23%*

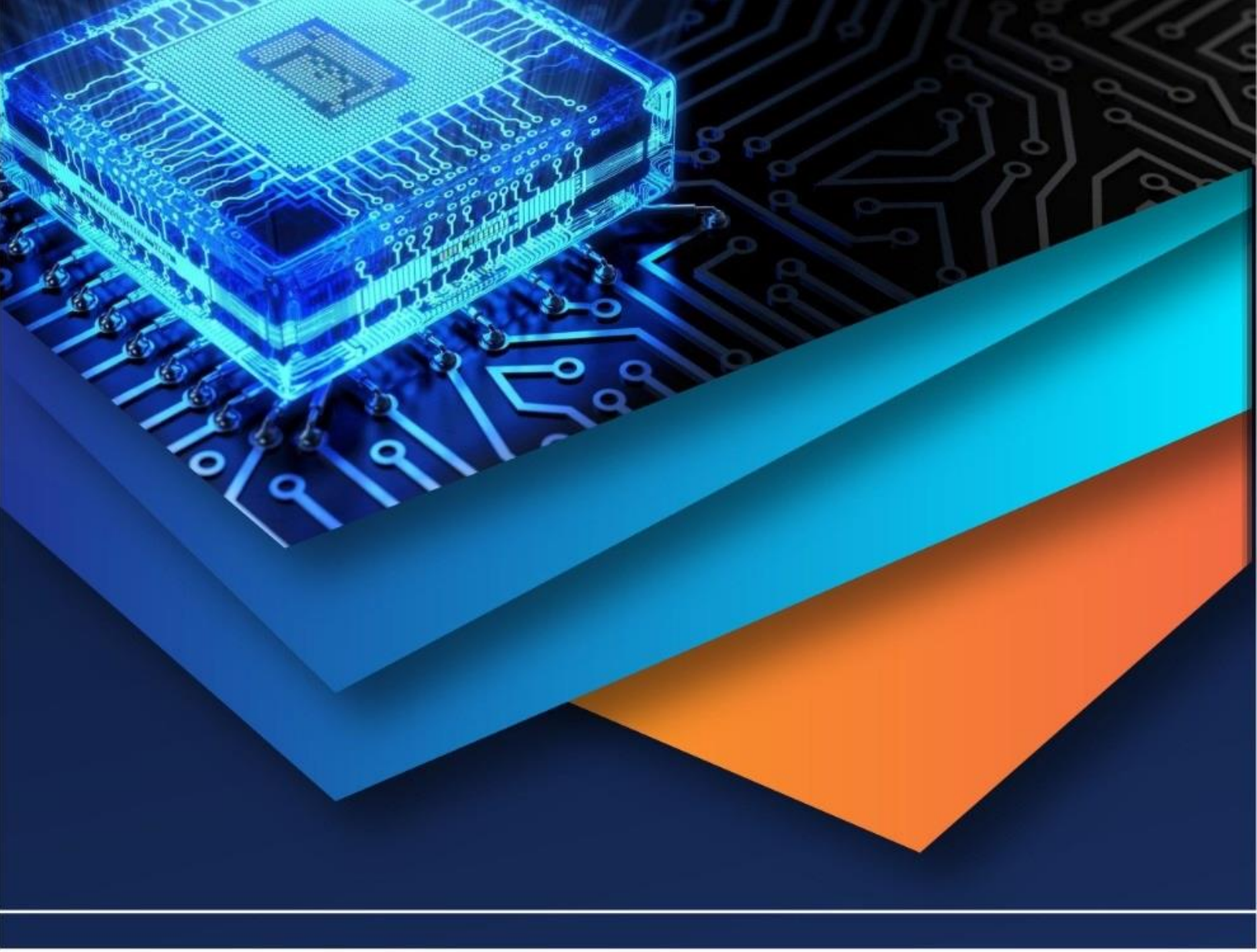
Sethi et al. [10] proposed a novel system for malware detection which can effectively identify and recognize malware. The proposed solution uses two separate feature selection algorithms to extract most important features that minimize training time and increase the precision of the detection and classification. The proposed method shows that the Decision tree has a high detection and classification accuracy compared with that of other classifiers. It also classifies malware based on their families, and has tested the accuracy of each family of malware. This paper also compare different classifier algorithms using proposed method. Table III shows the accuracy given by different classifier algorithms. It shows Decision Tree algorithm have highest accuracy of 99.11%.
When analyze the techniques used in these proposed papers it shows that some algorithms or architectures are efficient for malware classification.

IV. CONCLUSIONS

There are different detection methods are existing for malware detection. In this survey, different supervised learning methods which are used for malware detection are compared. But classical signature-based detection methods are not appli- cable in today's scenario.So some supervised learning methods are proposed by different authors for improved detection of malware. These supervised learning methods gives a dynamic analysis of malware and some of them allow real time analysis of malware. Different supervised learning algorithms are used by these methods and some of them are very effective in detection of malware. Some recent proposed methods are used CNN(Convolutional Neural Network) which is more accurate.

REFERENCES

[1] Rieck, Konrad Holz, Thorsten Willems, Carsten Düssel, Patrick Laskov, Pavel. " Learning and Classification of Malware Behavior". springer. 2008.
[2] Schiavoni, Stefano Maggi, Federico Cavallaro, Lorenzo Zanero, Ste- fano. (2014). "Phoenix: DGA-Based Botnet Tracking and Intelligence". 8550. 192-211.
[3] Z. Xu, S. Ray, P. Subramanyan and S. Malik, "Malware detection using machine learning based analysis of virtual memory access patterns," . Design, Automation Test in Europe Conference Exhibition (DATE), 2017, Lausanne, 2017, pp. 169-174. doi: 10.23919/DATE.2017.7926977
[4] Bashari Rad, Babak Shahpasand, Maryam Nejad, Mohammad. (2018)." Malware classification and detection using artificial neural network." Journal of Engineering Science and Technology. 13. 14-23.
[5] B. Yu, D. L. Gray, J. Pan, M. D. Cock and A. C. A. Nascimento, "Inline DGA Detection with Deep Networks,"2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, 2017, pp. 683-692. doi: 10.1109/ICDMW.2017.96
[6] Mac, Hieu Tran, Duc Tong, Van Nguyen, Giang Tran, Hai- Anh. (2017). "DGA Botnet Detection Using Supervised Learning Methods."SoICT 2017: Proceedings of the Eighth International Sym- posium on Information and Communication Technology. 211-218. 10.1145/3155133.3155166.
[7] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran and S. Venkatraman, "Robust Intelligent Malware Detection Using Deep Learning," in IEEE Access, vol. 7, pp. 46717-46738, 2019.
[8] Y. Li, K. Xiong, T. Chin and C. Hu, "A Machine Learning Frame- work for Domain Generation Algorithm-Based Malware Detection," in IEEE Access, vol. 7, pp. 32765-32782, 2019.
[9] Z. Cui, F. Xue, X. Cai, Y. Cao, G. Wang and J. Chen, "Detection of Ma- licious Code Variants Based on Deep Learning," in IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3187-3196, July 2018.
[10] K. Sethi, R. Kumar, L. Sethi, P. Bera and P. K. Patra, "A Novel Machine Learning Based Malware Detection and Classification Framework "2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY