*International Journal of Scientific Research and Engineering Development-– Volume 2 Issue 6, Nov-Dec 2019*

*Available at www.ijsred.com*

RESEARCH ARTICLE OPEN ACCESS

# Survey on Robust Intelligent Malware Detection Using Deep Learning

Bhagyashree Bhoyar*, Preeti Shinde**, Nikita Suryawanshi***, Mansi Sawaj****, Nikita Shinde*****

**(Department of Computer Engineering, SPPU/Dr. D. Y. Patil Institute of Technology & Engineering, Pimpri Email:bhagyashree05bhoyar@gmail.com)*

*** (Department of Computer Engineering, SPPU/Dr. D. Y. Patil Institute of Technology & Engineering, Pimpri Email: shindepreeti9526@gmail.com)*

**** (Department of Computer Engineering, SPPU/Dr. D. Y. Patil Institute Of Technology & Engineering, Pimpri Email: nikitasuryawanshi087@gmail.com)*

***** (Department of Computer Engineering, SPPU/Dr. D. Y. Patil Institute Of Technology & Engineering, Pimpri Email: mansisawaj1990@gmail.com)*

****** (Department of Computer Engineering, SPPU/Dr. D. Y. Patil Institute Of Technology & Engineering, Pimpri Email: nikitashinde2307@gmail.com)*

----------**********----------

## *Abstract:*

Security breaches due to attacks by malicious software (malware) continue to expand posing a major security concern in this digital age. Current malware detection solutions that support the static and dynamic analysis of malware signatures and behavior patterns are time consuming and have proven to be ineffective in identifying unknown malwares in real-time. Therefore there is a need to implement a zero-day detection and attack blockage system so that the user transactions are harm-free/safe. Different datasets are used for detection, classification and categorization of malwares. Dataset bias is removed in the experimental analysis by splitting the public and private datasets to train and test the model in a disjoint way using different timescales. Hybrid architecture is designed by implementing static, dynamic and image processing to detect malwares. A comparative study shows that deep learning architectures outperform classical machine learning architectures. The combination of visualization and deep learning architectures for static, dynamic, and image processing-based hybrid approach applied in a big data environment is the first of its kind toward achieving robust intelligent zero-day malware detection. Overall, this method paves way for an effective visual detection of malware using a scalable and hybrid deep learning framework for real-time deployments.

*Keywords* **—Malware Detection, Static & Dynamic Analysis, Image Processing, Zero-Day Detection, Hybrid Architecture, Deep Learning.**

----------**********----------

  

## I. INTRODUCTION

IOT & Applications have led to the development of modern concept of the information society. However, security concerns pose a major challenge in realising the benefits of industrial revolution as cyber-criminals attack individual PC's and networks for stealing confidential data for financial gains and causing DOS attacks to systems. Such attackers make use of malicious software or malware to cause serious threats and vulnerabilities of system. A malware is computer program with the purpose of causing harm to the OS.

Deep Learning is an artificial intelligence (sometimes known as machine intelligence) function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Deep Learning is a subgroup of Machine Learning in Artificial Intelligence (AI) that has networks adequatefor unsupervised learning from data that is unlabelled. Also known as Deep Neural Network.

Malware analysis involves two key techniques: static and dynamic analysis. Static analysis examines malware without actually running it. Dynamic analysis executes malware in a controlled and observed environment to observe its behaviour. Each technique includes elements in addition categorized as basic or advanced. Although there are benefits for conducting static and dynamic analysis as separate tasks, an analyst can realise the value provided by conducting both techniques when reverse engineering complex malware.

Several studies conducted have taken advantage of the fact that most malware variants are similar in structure, with digital signal and image processing techniques used for malware categorization. They have transformed the malware binaries into gray scale images and report that malware from the same malware family seem to be quite similar in layout and texture.The visualization techniques use imageprocessing approaches to classify themalware. The malware executable binary files are transformed into image and these images areused to detect the family of malware. In image processing the strand is divided in three phases that are: Malware Image Generation, Feature extraction and classification.

### *PROJECT AIM*

Design a system that performs zero-day malware detection with the help of static, dynamic and image processing and effective visual detection using scalable and hybrid deep learning framework for real-time deployment.

## II. LITERATURE SURVEY

***Ebenuwa, Solomon H., et al. "Variance ranking attributes selection techniques for binary classification problem in imbalance data." IEEE Access 7 (2019): 24649-24666. [1]***

A new attribute selection technique called variance ranking for handling imbalance class problems in a dataset was proposed. The results achieved were correlated to two well-known attribute selection techniques: the Pearson correlation and information gain technique.A novel similarity measurement technique ranked order similarity-ROS to evaluate the variance ranking attribute selection correlated to the Pearson correlations and information gain. Farther validation was carried out using three binary classifications: logistic regression, support vector machine, and decision tree. The ROS technique provided an admirable means of grading and measuring the similarities.

***Anderson, Hyrum S., and Phil Roth"Ember: an open dataset for training static PE malware machine learning models." arXiv preprint arXiv: 1804.04637 (2018). [2]***

EMBER: a labelled benchmark dataset for training machine learning models to statically detectmalicious Windows portable executable files was planned. This dataset fills a gap in the information security machine learning association: a malicious dataset that is large, open and general enough to cover several interesting use cases. The comparisonof a baseline gradient boosted decision tree model, trained using LightGBM with default settings to MalConv, an end-to-end (featureless) deep learning model for malware detection is demonstrated in one use case. Results showed that even without hyper-parameter optimization, the baseline EMBER model outperforms MalConv.The dataset, code and baseline model provided by EMBER helped invigorate machine learning research for malware detection, in much the same way that benchmark datasets have progressive computer vision research.

***Agarap, Abien Fred, and Francis John Hill Pepito"Towards building an intelligent anti-malware system: a deep learning approach using support vector machine (SVM) for malware classification." arXiv preprint arXiv: 1801.00318 (2017). [3]***

This is an anti-malware system which detects newly released malware by using mathematical generalization .That is,it finds the relationship between given malware and its corrosponding family.This system has used maling dataset which consisted of malware images that were proccessed from malware binaries and some DL models were trained to classify each malware family.DL models are:CNN-SVM,GRU-SVM and MLP-SVM.Empirical evidence has shown that GRU-SVM model has higher peredictive accuracyof~~84.92%[3]. Hence,mentioned model had sophisticated architecture design among others.

***Damodaran, Anusha, et al. "A comparison of static, dynamic, and hybrid analysis for malware detection." Journal of Computer Virology and Hacking Techniques 13.1 (2017): 1-12. [4]***

Comparison of malware detection techniques based on static, dynamic and hybrid analysis was done. Hidden Markov Models (HMMs) were trained on both static and dynamic feature sets and the results were compared based on detection rates over a substantial number of malware families. In hybrid cases, dynamic analysis was used in the training phase, with static techniques used in detection phase, and vice versa. It was concluded that a fully dynamic approach generally yields the best detection rates.

***Saxe, Joshua, and Konstantin Berlin"Deep neural network based malware detection using two dimensional binary program features." 2015 10th International Conference on Malicious and Unwanted Software (MALWARE) IEEE, 2015[5]***

  

Machine learning provides the work automation required for detection of newly discovered malware families. It could potentially learn rationalizations about malware and benignware that supports the detection of unknown and newly released malware family.Very few ML based methods for malware detection had achieved low false positive rates and high scalability that is required to deliver deployable detectors.The author introduced an approach that addressed these issues. The system has achieved a usable detection rate at an extremely low false positive rate and scales to real world training example volumes on commodity hardware. Results are achieved by directly learning on all binaries, without any filtering, unpacking, or manually separating binary files into categories. Further, the confirmation of false positive rates was done directly on a live stream of files coming in from Invincea's deployed endpoint solution. It provided an estimate of how many new binary files are expected to see a day on an enterprise network, and describe how that relates to the false positive rate and translates into an intuitive threat score.

## III. ADVANTAGES

- High level of accuracy by taking asset of many classifiers.
- Low level of false positives in static analysis.
- Good in detecting unknown malwares with the help of dynamic analysis.
- Easy to determine and classify malwares into malware family.
- Provides high security as attacks can be blocked early using zero-day detection.

## IV. DISADVANTAGES

- Hybrid approach can be time consuming because it has many layers to make the final result.
- Analysing multipath malwares can be difficult.

## V. CONCLUSION

The system evaluates classical machine learning algorithms (MLA's) and deep learning architectures based on static analysis, dynamic analysis and image processing for malware detection and designs a highly scalable framework to detect, classify and categorize zero-day malwares. This framework applies deep learning on collected malwares from end user hosts and follows a two-stage process for malware analysis. In the first stage, a hybrid static & dynamic analysis is applied for malware classification. In second stage, image processing will be done to group the malwares according to their category. The framework is capable of analysing large number of malwares in real-time, and scaled-out environment to analyse even larger number of malwares by stacking a few more layers to the existing architectures.
Future research entails exploration of these variations with new features that could be added to the existing data.

## VI. REFERENCES

[1] Ebenuwa, Solomon H., et al. "Variance ranking attributes selection techniques for binary classification problem in imbalance data." IEEE Access 7 (2019): 24649-24666.

[2] Anderson, Hyrum S., and Phil Roth. "Ember: an open dataset for training static PE malware machine learning models." arXiv preprint arXiv: 1804.04637 (2018).

[3]Agarap, Abien Fred, and Francis John Hill Pepito. "Towards building an intelligent anti-malware system: a deep learning approach using support vector machine (SVM) for malware classification." arXiv preprint arXiv:1801.00318 (2017).

[4] Damodaran, Anusha, et al. "A comparison of static, dynamic, and hybrid analysis for malware detection." *Journal of Computer Virology and Hacking Techniques* 13.1 (2017): 1-12.

[5] Saxe, Joshua, and Konstantin Berlin. "Deep neural network based malware detection using two dimensional binary program features." *2015 10th International Conference on Malicious and Unwanted Software (MALWARE)* IEEE, 2015

[6] Anderson, Ross, et al. "Measuring the cost of cybercrime." *The economics of information security and privacy* Springer, Berlin, Heidelberg, 2013 265-300

[7] Rossow, Christian, et al. "Prudent practices for designing malware experiments: Status quo and outlook." *2012 IEEE Symposium on Security and Privacy*, IEEE, 2012

[8] Alazab, Mamoun, et al. "Zero-day malware detection based on supervised learning algorithms of API call signatures." *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*,Australian Computer Society, Inc., 2011.

  