International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217518

Evaluating Malware Detection System using Machine Learning Algorithms

S. Bhaskara Naik[1], B. Mahesh[2]

[1] Lecturer, S.V.B.Government Degree College, Koilakuntla, Kurnool(Dist), Andhra Pradesh, India

[2]Associate Professor, Department of CSE, Dr. K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India

ABSTRACT

Article Info

Volume 7, Issue 5

Page Number: 43-48

Publication Issue :

September-October-2021

Article History

Accepted : 15 Sep 2021

Published : 30 Sep 2021

Malware, is any program or document that is unsafe to a PC client. Kinds of malware can incorporate PC infections, worms, Trojan ponies and spyware. These noxious projects can play out an assortment of capacities like taking, scrambling or erasing touchy information, adjusting or commandeering center processing capacities and observing clients' PC action. Malware identification is the way toward checking the PC and documents to distinguish malware. It is viable at distinguishing malware on the grounds that it includes numerous instruments and approaches. It's anything but a single direction measure, it's very intricate. The beneficial thing is malware identification and evacuation take under 50 seconds as it were. The outstanding development of malware is representing an extraordinary risk to the security of classified data. The issue with a significant number of the current order calculations is their small presentation in term of their capacity to identify and forestall malware from tainting the PC framework. There is a critical need to assess the exhibition of the current Machine Learning characterization calculations utilized for malware identification. This will help in making more hearty and productive calculations that have the ability to conquer the shortcomings of the current calculations. As of late, AI methods have been the main focus of the security specialists to distinguish malware and foresee their families powerfully. Yet, to the best of our information, there exists no complete work that looks at and assesses a sufficient number of machine learning strategies for characterizing malware and favorable examples. In this work, we led a set of examinations to assess AI strategies for distinguishing malware and their classification into respective families powerfully. This investigation did the presentation assessment of some characterization calculations like J45, LMT, Naive Bayes, Random Forest, MLP Classifier, Random Tree, AdaBoost, KStar. The presentation of the calculations was assessed as far as Accuracy, Precision, Recall, Kappa Statistics, F-Measure, Matthew Correlation Coefficient, Receiver Operator Characteristics Area and Root Mean Squared Error utilizing WEKA AI and information mining recreation device. Our test results showed that Random Forest calculation delivered the best exactness of 99.2%. This decidedly shows that the Random Forest calculation accomplishes great precision rates in identifying malware.

Keywords : Malware, Machine Learning, Deep Learning, classification algorithms, Random Forest

I. INTRODUCTION

The huge development of administrations and possessions has expanded the quantity of Internet users during an assortment of gadgets going from systems to implanted frameworks. This Internet connectivity has offered numerous types of assistance to the end users, like simple and speedy correspondence. These days, end users can appreciate online administrations anyplace any time through an Internet associated gadget like mobiles, tabs, and so on. This expanding number of Internet clients additionally enacted the pernicious programmers to foster vindictive applications or programs generally called malware. In the new years, a enormous amount of malware has been seen as portrayed in Fig 1.

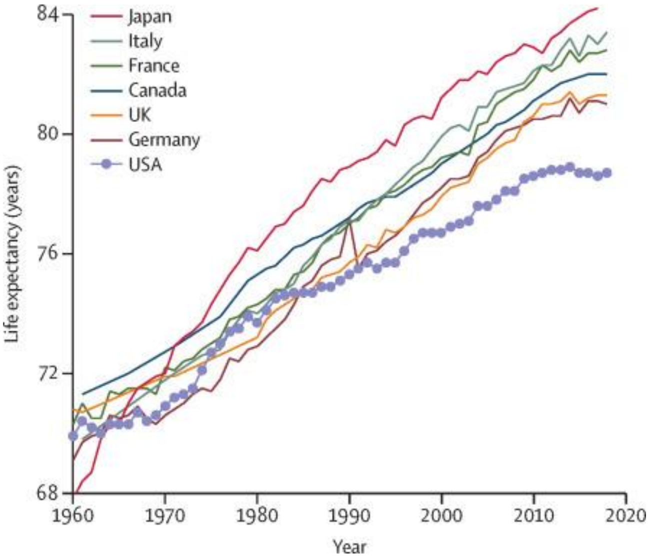

Figure 1. Trends in malware growth

Numerous antivirus, interruption location frameworks and additional malware identification frameworks have been produced for the avoidance of harm brought about by these pernicious projects. All things considered, there exist a few issues that need quick consideration. Due to varying nature of malware and flaw in existing programming. Different procedures from various controls have been projected for compelling malware discovery. The strategies can be sorted extensively into two classifications, to be specific the inert, signature based methods and active, conduct based procedures. Static procedures examine malware dependent on its design, control flow, and so on without executing it. These methods include foundation of a mark data set. The significant constraint is that these methods neglect to distinguish a novel malware until its mark is refreshed. Though, dynamic procedures examine the malware tests through its execution. These strategies dissect conduct of malware tests from their execution reports. As of late, malevolent software engineers are growing more intricate and progressed malware utilizing muddling and encryption strategies. Static procedures neglect to identify malware precisely. Though, dynamic procedures have benefit over static methods, since it is more hard to cover the conduct of malware during its execution. Mulling over the benefits of dynamic methods, the focal point of ebb and flow explore has moved to dynamic and robotized procedures for malware detection.

A malware can basically be characterized as a noxious program which the client accidentally introduce on their appliance and later on these projects can start to upset the appropriate activity of the machine or may proceed undetected and do pernicious activities without been taken note. At the point when the assailant deals with the machine, he would then be able to approach any data put away on the machine. A portion of the misleading methodologies used to introduce malware on the PC framework during the web incorporate repackaging the product, update assault or longing for download. The assailant utilizes any of the strategies referenced before to make malignant programming by embeddings a particular kind of malware into it prior to transferring it to the web. Malware can be portrayed as different sorts of programming which have the ability to unleash ruin on a PC framework or unlawfully utilize this data without the assent of the clients. Malware can be ordered in different types, for example, Botnet, Backdoor, Ransomware, Rootkits, Virus, Worms, and Trojan Horse, Spyware, Adware, Scareware and Trapdoor. They are utilized to assault PC frameworks and for performing crimes like trick, phishing, administration abuse and root access.

Here, we assess execution of delegate AI methods from various classes like choice tree based, likelihood based utilizing a genuine malware dataset as far as an assortment of execution measurements. Assessment of ML procedures on various measurements is significant, on the grounds that distinctive ML methods have been intended to advance an alternate arrangement of models. Along these lines, they act contrastingly in a comparative climate.

In this research, we recognize the best performance techniques for dynamic malware identification based upon a capable arrangement of highlights separated from implementation reports of malware and generous examples. most important commitments of this work are:

- mining of dynamic conduct of genuine malware sample from Virus Total by executing them in a virtually controlled climate of Cuckoo Sandbox.
- Collection of highlights addressing dynamic behavior of malware to create a genuine malware dataset.
- Assessment of ML procedures class shrewd, for example, choice tree based, likelihood based techniques, and so forth to distinguish the promising techniques using a genuine malware dataset.
- Experiential near investigation of the ML methods to recognize the best performing technique for successful malware identification, so as to use it as an applicant strategy for mounting dynamic malware location frameworks.

II. METHODOLOGY

This segment presents the general system continued in this research as portrayed in Figure 2. It incorporates dynamic malware discovery utilizing ML strategy comprising of information age stage, information extraction stage, classification stage, and execution metric calculation stage. The information executes the generous and malware PE in a controlled climate of Cuckoo sandbox and produces its execution report as a Javascript Object Notation(JSON) file. The information extraction stage extricates highlights from JSON files that addresses the unique conduct of tests and marks each example as considerate or malware. It produces a genuine malware dataset that is additionally utilized as preparing and test dataset by the classification stage. The exhibition metric calculation stage registers malware location brings about terms of assortment of measurements.

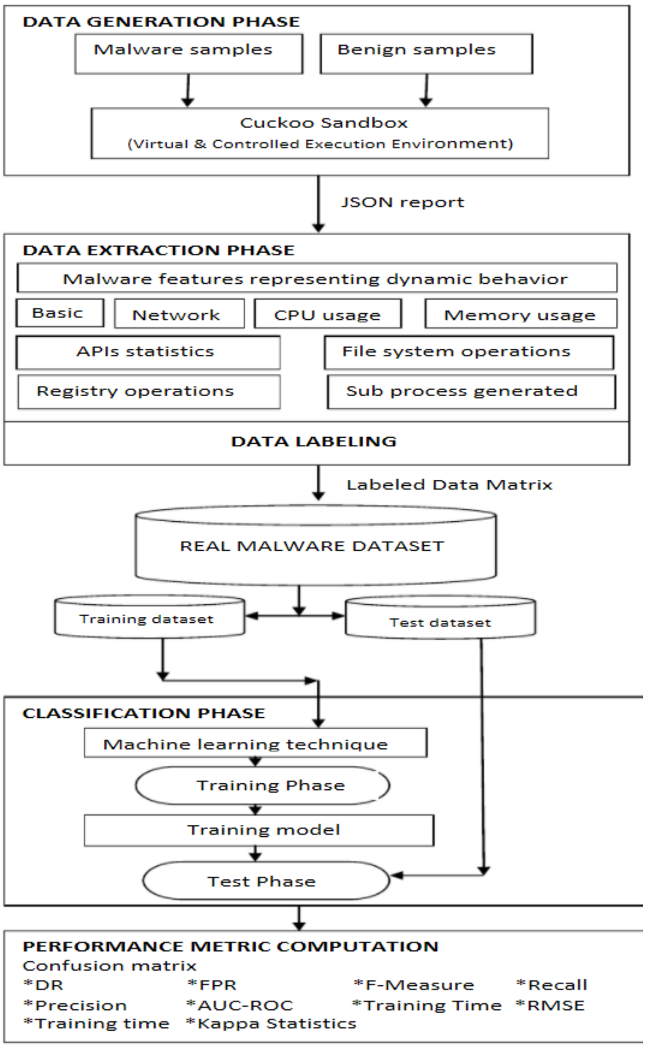

Figure 2: Proposed System to Evaluate Malware Detection System

- *Data Generation Phase:* In this stage information is created through malware tests, Cuckoo Sandbox, Anubi's. These environment permit the execution of malware and benign binaries inside a segregated climate, dissect and record their conduct.
- *Data Extraction Phase:* The information gathered through Cuckoo Sandbox, Anubi's will be available in the form of JSON objects. The primary strides for information extraction phase are as below:

1. Peruse areas of JSON file
2. Concentrate highlights
3. Naming of malware tests

- *Classification Phase:* An enormous numeral of regulated ML methods have been intended for characterizing dataset into a bunch of malware classes. For example, Artificial Neural Networks are intended to impersonate the human mind. They have the ability to gain proficiency with any non-straight connection among input and wanted yield even within the sight of loud preparing information. Intrigued per users may investigate audit of ML methods referenced in the examinations. The ML methods from various classes carried out in ML device WEKA are utilized to create prepared models for dataset having divergence of 70% as preparing and 30% test dataset. In the current work, we have been utilizing default boundaries of various ML techniques executed in WEKA. Notwithstanding, fine tuning of the boundaries may lead in additional improvement of classification aftereffects of ML procedures. The prepared model of ML procedure is additionally used to foresee malware group of obscure examples. The yield of this stage is a report comprising of disarray grid and different subtleties. The created report is utilized by the security specialists for additional register other execution measurements and infer strategy decisions.
- *Performance Computation segment:* The performance metric calculation stage calculates the identified execution measurements from the confusion grid in the wake of testing stage. The confusion network gives the upsides of FP, TN, FN,and TP. It determines the weighted average of identified execution measurements like TPR(also known as Recall), FPR, RMSE, Detection Accuracy, Precision, F-measure, AUC-ROC from the values of TN, FN, TP, and TN for relative evaluation of the ML strategies. We utilized a biased average of various measurements for determining the metrics like AUC-ROC by following one versus rest approach.

III. EVALUATION MALWARE DATASET

Here, we utilized the malware and kind samples from VirusTotal. VirusTotal is a site that offers the investigation of dubious documents and URLs to identify kinds of malware together with viruses, worms, and trojans. VirusTotal totals numerous antivirus items and online sweep motors to check for infections that the client's own antivirus may have missed, or to confirm against any bogus positives. Documents up to 512 Mega Bytes can be transferred or send through email to the site. Antivirus programming sellers can get duplicates of files that were flagged by other scans, but passed by their own motor, to improve their software and, likewise, VirusTotal's own capacity. Clients can likewise examine suspect URLs and search through the VirusTotal dataset. VirusTotal uses Cuckoo sandbox for dynamic investigation of malware. In coming about dataset, an enormous number of malware families were found. For assessment motivation behind ML procedures, we ordered malware samples into various families, according to their fundamental functions. For uniform and far reaching analysis of the proposed work, malware dataset is divided arbitrarily into preparing and test dataset. The preparation dataset contains 70 percentage of tests and test dataset contains 30 percentage examples. The classification wise number of tests in the preparation informational collection and test information set are as portrayed in Table 1.

Sr No	Class name	Number of samples
Training dataset		
1	Benign	3433
2	Trojan	4447
3	Virus	265
4	Worm	326
5	Packed	430
6	Backdoor	233
7	Hoax	41
8	DangerousObject	28
9	Adware	62
	Total	9265
Test dataset		
1	Benign	1451
2	Trojan	1889
3	Virus	115
4	Worm	140
5	Packed	206
6	Backdoor	109
7	Hoax	20
8	DangerousObject	13
9	Adware	28
	Total	3971

Table 1: Categories and Number of Samples in dataset

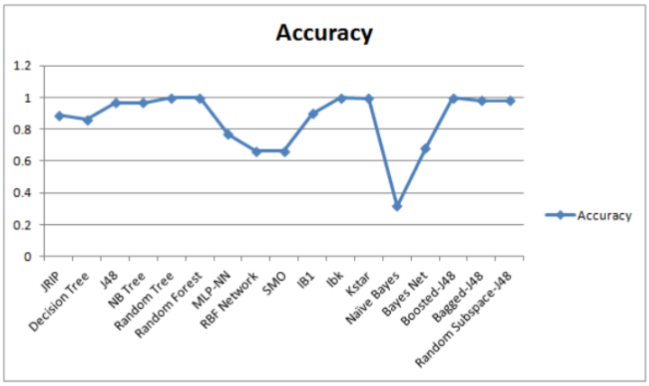

Figure 3: Performance comparison in terms of accuracy

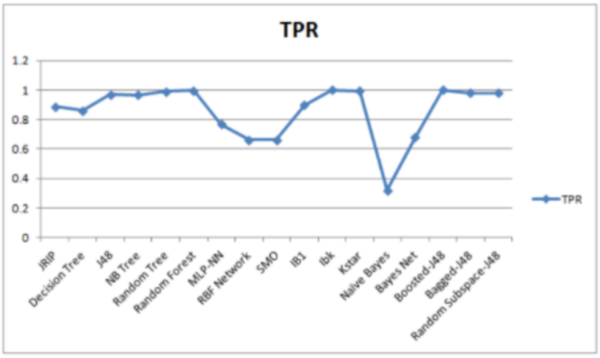

Figure 4: Performance comparison in terms of TPR

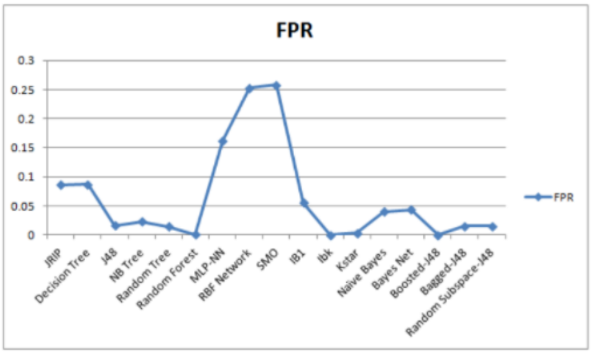

Figure 5: Performance comparison in terms of FPR

IV. CONCLUSION

In this work, an assessment of regulated ML techniques is done experimentally for recognizing malware using a genuine malware dataset as far as an assortment of evaluation measurements. The significant inspiration driving using an assortment of assessment measurements is that dissimilar techniques are created to advance distinctive set of models. To assess ML methods widely, a promising arrangement of highlights has been extract from malware and kindhearted executable examples using a Cuckoo Sandbox and a Python based automated system to shape a genuine malware dataset. In this work, we distinguish the best strategies for effective malware location dependent on a genuine mal-product dataset as far as identified execution metrics.

V. REFERENCES

[1]. SF Ahmad, SZ Ahmad, SR Xu, and B Li. Next gen-eration malware analysis techniques and tools. InElectronics, Information Technology and Intellectu-alization: Proceedings of the International Confer-ence EITI 2014, Shenzhen, China, 16-17 August 2014,page 17. CRC Press, 2015.

[2]. Ulrich Bayer, Andreas Moser, Christopher Kruegel,and Engin Kirda. Dynamic analysis of malicious code.Journal in Computer Virology, 2(1):67–77, 2006.

[3]. R. Bellman. Adaptive control processes: a guidedtour princeton university press. Princeton, New Jer-sey, USA, 1961.

[4]. Silvio Cesare and Yang Xiang. Software similarityand classification. Springer Science & Business Me-dia, 2012.

[5]. Gianluca Dini, Fabio Martinelli, Andrea Saracino, andDaniele Sgandurra. Madam: a multi-level anomalydetector for android malware. In International Con-ference on Mathematical Methods, Models, and Ar-chitectures for Computer Network Security, pages240–253. Springer, 2012.

[6]. Manuel Egele, Theodoor Scholte, Engin Kirda, andChristopher Kruegel. A survey on automated dynamicmalware-analysis techniques and tools. ACM Comput-ing Surveys (CSUR), 44(2):6, 2012.

[7]. Christian Gorecki, Felix C Freiling, Marc K ¨uhrer, andThorsten Holz. Trumanbox: Improving dynamic mal-ware analysis by emulating the internet. In Stabi-lization, Safety, and Security of Distributed Systems,pages 208–222. Springer, 2011.

[8]. Kent Griffin, Scott Schneider, Xin Hu, and Tzi-CkerChiueh. Automatic generation of string signatures formalware detection. In Recent advances in intrusiondetection, pages 101–120. Springer, 2009.

[9]. Chun-Ying Huang, Yi-Ting Tsai, and Chung-HanHsu. Performance evaluation on permission-based de-tection for android malware. In Advances in Intelli-gent Systems and Applications-Volume 2, pages 111–120. Springer, 2013.

[10]. Youngjoon Ki, Eunjin Kim, and Huy Kang Kim. Anovel approach to detect malware based on api call se-quence analysis. International Journal of DistributedSensor Networks, 2015:4, 2015.

[11]. Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayi-otis E Pintelas. Machine learning: a

review of classi-fication and combining techniques. Artificial Intelli-gence Review, 26(3):159–190, 2006.

[12]. G. Kumar and K. Kumar. Ai based supervised clas-sifiers: an analysis for intrusion detection. In Proc.of International Conference on Advances in Comput-ing and Artificial Intelligence, pages 170–174. ACM,2011.

[13]. G. Kumar and K. Kumar. An information theoreticapproach for feature selection. Security and Commu-nication Networks, 5(2):178–185, 2012.

[14]. G. Kumar, K. Kumar, and M. Sachdeva. The use ofartificial intelligence based techniques for intrusiondetection: a review. Artificial Intelligence Review,34(4):369–387, 2010.

[15]. Andreas Moser, Christopher Kruegel, and EnginKirda. Limits of static analysis for malware detec-tion. In Computer security applications conference,2007. ACSAC 2007. Twenty-third annual, pages 421–430. IEEE, 2007.

[16]. S. Mukkamala and A.H. Sung. A comparative studyof techniques for intrusion detection. In Proc. of 15thIEEE International Conference on Tools with Artifi-cial Intelligence, 2003, pages 570–577. IEEE, 2003.

[17]. Fairuz Amalina Narudin, Ali Feizollah, Nor BadrulAnuar, and Abdullah Gani. Evaluation of machinelearning classifiers for mobile malware detection. SoftComputing, 20(1):343–357, 2016.

[18]. Philip O'Kane, Sakir Sezer, and Keiran McLaughlin.Obfuscation: The hidden malware. Security & Pri-vacy, IEEE, 9(5):41–47, 2011.

[19]. Konrad Rieck, Thorsten Holz, Carsten Willems,Patrick D¨ussel, and Pavel Laskov. Learning and clas-sification of malware behavior. In Detection of In-trusions and Malware, and Vulnerability Assessment,pages 108–125. Springer, 2008.

[20]. Cuckoo Sandbox. Automated malware analysis, 2013.

[21]. Bhaskar Pratim Sarma, Ninghui Li, Chris Gates,Rahul Potharaju, Cristina Nita-Rotaru, and Ian Mol-loy. Android permissions: a perspective combiningrisks and benefits. In Proceedings of the 17th ACMsymposium on Access Control Models and Technolo-gies, pages 13–22. ACM, 2012

Cite this article as :

S. Bhaskara Naik, B. Mahesh, "Evaluating Malware Detection System using Machine Learning Algorithms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 5, pp. 43-48, September-October 2021. Available at doi : https://doi.org/10.32628/CSEIT217518
Journal URL : https://ijsrcseit.com/CSEIT217518