Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Comparison study of Machine Learning Algorithm and Data Science based Machine Learning Algorithm Malware Detection

Sunita Choudhary[1], Anand Sharma[2]
[1,2]School of Engineering and Technology, Mody University of Science and Technology Lakshmangarh, India
sunitadangi@gmail.com[1], anand_glee@yahoo.co.in[2]

Article Info
Page Number: 01 - 07
Publication Issue:
Vol 71 No. 3s (2022)

Article History
Article Received*: 22 April 2022*
Revised*: 10 May 2022*
Accepted*: 15 June 2022*
Publication*: 19 July 2022*

Abstract
With quick development and advancement of the web, malware is one of major advanced perils these days. Hence, malware discovery is a significant component in the security of PC frameworks. These days, assailants by and large plan polymeric malware, it is typically a kind of malware that ceaselessly changes its unmistakable component to trick recognition strategies that utilizes run of the mill signature-based techniques. For that reason, the requirement for Machine Learning based identification emerges. In this work, we will acquire standard of conduct that might be accomplished through static or dynamic examination, a while later we can apply unique ML strategies to recognize regardless of whether it's malware. Conduct based Detection techniques will be talked about to take advantage from ML calculations in order to approach social-based malware acknowledgment, furthermore, grouping model. In this paper, study related between two major components. First one is machine learning algorithm apply on data set directly. Second is same Machine learning algorithm apply with Data science pre-processing steps.

Keywords: - KNN, Naïve Bayes, SVM, J48, Malware Detection, Classification, Machine learning.

Abbreviation:

KNN: k-nearest neighbors
SVM: Support-vector machines
J48: J standing for Java

Introduction: Malware is any product tenaciously intended to make harm a PC, server, customer or PC organization [1]. With the fast advancement and development of the web, malware has turned out to be one of the major digital dangers in nowadays. In the year 2017, a cybersecurity and anti-virus provider like Kaspersky Labs defined malware as “a kind of PC program planned to taint an authentic client's PC and perpetrate hurt on it in different manners." As the decent variety of malware is expanding nowadays, antivirus tools are not capable of fulfilling the need of protection and results in millions of hosts being hacked. In addition to that, the skills required for malware development is decreasing due to high availability of attacking tools on the Internet [2]. According to data of AV-TEST Institute, they reported over three lac fifty thousand novel malware (malicious projects) and PUA (Potentially Unwanted Applications) consistently. Therefore, to protect

computer system from malware is one significant tasks of cybersecurity for a single user as well as for businesses because even a single attack can result in huge information and financial loss.

A. Types of Malware

Malware can be isolated into a few classes relying upon its purpose. The classes are as per the following [3]:

• Adware: It is the slightest risky and the most rewarding malware, it shows advertisements on PC.
• Spyware: As it implies from the name, the malware that uses for spying. Some run of the mill activities of spyware incorporate following inquiry history to send customized advertisements, following exercises to offer them to the outsiders in this way.
• Virus: This is the most straightforward type of the software. It is basically any bit of programming that is stacked and propelled without client's authorization while replicating itself or contaminating changing other programming. Regularly this is spread by sharing records or programming between PCs.
• Worm: It is a program that imitates itself and obliterates data and records on the PC.
• Trojan: It is a kind of malicious code and computer software to look authorized but can control the system or machine.
• RootKit: An assortment of vindictive programming created to enable access to a framework or on particular area of the system.
• Backdoors: It is a method to convert the bypassing normal authentication and encryption.
• Keyloggers: It is totally depend on Keyboard working style like the action of keyboard typically covertly unaware their actions are being method and monitored.
• Ransomware: This is malware software but extreme use of this software is for Accounts section like access to system until a sum of money is paid [4].
• Browser Hijacker: It is a type of undesirable programming that changes a program's setting without client authorization, to infuse the profitless promoting into program.

B. Malware Discovery Investigation Procedures

All malware discovery procedures can be partitioned into mark based and conduct based strategies. How about we examine a few procedures for the examination.

• Static Method: A static strategy for examination of malware depends on pre- characterized marks. These can be document fingerprints, e.g. file metadata, static strings, MD5 or SHA1 hashes.
• Dynamic Method: A dynamic method of analysis and for resultant of malware relies for the change according to time and classifies the malware-based approach on the hand behalf of the time approach.

Literature:

Uppal et al. (2014) presented a malware identification approach based on features from the API sequences. The method monitors the execution of a binary to keep track of the API calls invoked [5]. Bekerman et al. (2015) presented a system for detecting malware by analyzing network traffic. In their work, they extracted 972 behavioral features from analyzing the network traffic on the Internet, Transport and Application layers [6]. O ’kane et al. (2016) analyzed malicious runtime

traces to determine (1) the optimal set of opcodes necessary to build a robust indicator of maliciousness in software, and to determine (2) the optimal duration of the program's execution to accurately classify benign and malicious software. The proposed approach used a Support Vector Machine on the opcode density histograms extracted during the program's execution to detect malware [7]. Galal et al. (2016) presented an approach to process raw information gathered by API call hooking to produce a set of actions representing the malicious behaviors of malware [8]. Boukhtouta et al. (2016) proposed a malware detection and classification system based on DPI and flow packed headers. Their approach executed malware in a sandbox for 3 min to generate representative malicious traffic [9]. Salehi et al. (2017) proposed a dynamic method to detect malicious activity in Android APKs based on the arguments and return values of API calls. They developed an "in-house" tool consisting of a virtual machine, a hooking tool and a logging system, which was used to analyze the binary files and monitor their behavior. Their approach is based on the hypothesis that API names alone may not represent intent of the operations that the function performs [10]. Yuxin et al., (2019) used a Deep Belief Network (DBN) as an autoencoder to reduce the dimensions of the input feature vectors. As a result, after learning is completed, the last hidden layer of the DBN outputs a new representation or encoding of the N-gram vectors passed as input. By training the DBN with unlabeled data, their classification accuracy outperformed that of the K-Nearest Neighbor, Support Vector Machines and Decision Tree algorithms [11].

Table 01: Related work of Malware Detection

Paper	Feature Type	Used Algorithm/Technique	Feature Selection, Reduction
Uppal et al. (2014) [5]	API Call Traces	NB, RF, DT, SVM	odds ratio
Bekerman et al. (2015) [6]	Network Traffic	Naïve Bayes, J48 DT, RF	Correlation Feature Selection Algorithm
O'kane et al. (2016) [7]	Instruction Traces	SVM	PCA
Galal et al. (2016) [8]	API Call Traces	DT, RF, SVM	Hand-crafted Heuristics
Boukhtouta et al. (2016) [9]	Network Traffic	Boosted J48, J48, NB, Boosted NB, SVM, HMMs	
Salehi et al. (2017) [10]	API Call Traces	RF, J48 DT, Bayesian Logistics Regression, Sequential Minimal Optimization	Fisher Score, SVM based on Recursive Feature Elimination
Yuxin et al. (2019) [11]		DBN, SVM, K-NN, DT	n-grams (opcodes)

Methodology:

The proposed work is presented by the data flow diagram with step-by-step methodology. Firstly, data pre-processing is performed, steps covered split the dataset and pre-processing steps, clean the data set and noise after that null values clearance and then the results are obtained in form of accuracy, comparison with existing results.

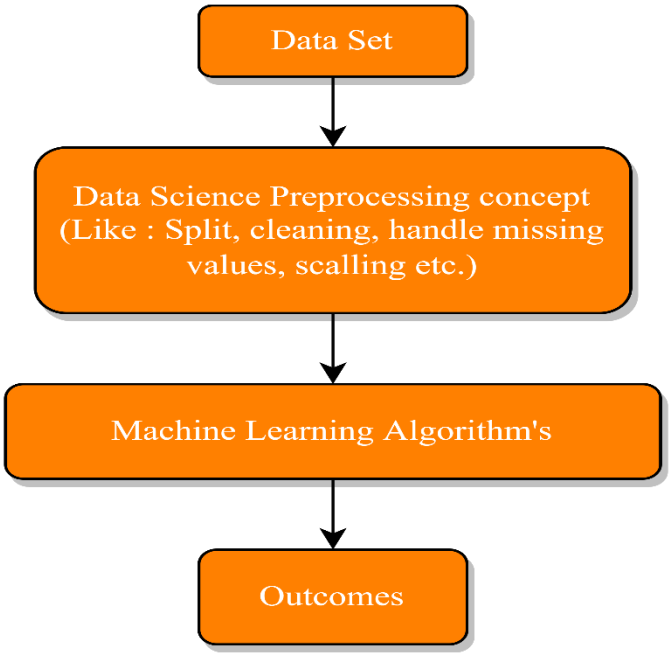

Fig. 01 Data Flow Diagram of Methodology

As per Data flow diagram shown basically 04 steps in the methods:

- **Data set:** A sum of 220 one of kind malware tests are gathered. Additionally gathers clean framework documents from a perfect establishment from framework records of Windows XP Professional. What's more, a report is created by leading conduct observing as for malware documents and clean records [12-13].
- **Concept of Data Science:**
  Spilt the data set into 20% and 80% Test and Training data set randomly. After that remove the noise and clerar the null values in the data set malware columns. then scale the data according to the standard parameter. Now the data proceed to different algorithms one by one for outcome.
- **Machine Learning Algorithm's:**
  SVM : SVM assembles a hyperplane or set of hyperplanes in a high or boundless dimensional space, which can be used for arrangement. A decent partition is practiced by the hyperplane that has the greatest partition to the nearest getting ready data motivation behind any class (called useful edge), since when everything is said in done greater edge, lower speculation mistake of characterizer[14-15].
  • KNN : KNN may be utilized for arrangement and relapse issues. Although in our problem set, it is used to classify malwares in view of those k preparing models or instances, which are in majority with respect to the input i.e. which class it closely associated with. There are only two classes which the input can be associated with one is malware detected or not. KNN basically using for Classification techniques but in Malware Detection many terminally says and according to our study also its evaluate depends on the "ease to interpret output, calculation time and predictive power".

• Naïve Bayes : As we discussed earlier, for malware detection Naive Bayes categorizer may be utilized to characterize or distinguish malware dependent on the conditional probability.
• J48 Decision Tree : Decision tree is a structure that consolidates a root hub, branches and leaf hubs. Each hub connotes a test quality, each branch
implies the consequence of the test, and each leaf hub holds class name. Decision tree doesn't require any area learning yet it utilizes the idea of data entropy, and it is anything but difficult to fathom, and the learning and characterization steps of choice tree are straightforward and quick [16].

- **Outcomes:** Outcome is the form of accuracy based on standard formulas. Machine learning algorithms based data science concept outcomes are KNN achieve 90.2% and 96.3%, Naive Bayes achieve 69.1% and 68.1%, SVM 94.7% and 97.9% , J48 aimed 97.3% and 98.2% respectively.

Result and Discussion:

As per performance of algorithms with pre-processing steps of data science concept achieved accuracy is better than existing accuracy on same algorithms. When comparison shown then the outcome is say Data science concept using algorithm results are better than only algorithm performance outcomes. Comparison shown in table 02.

Table 02: Comparison between ML algorithms and Data science based ML Algorithms

Categorizer	Machine Learning Algorithms		ML Algo's apply with Data Science concept	
	PERFORMANCE METRICS RESULTS			
	BINARY, NO FEATURE SELECTION	TERM FREQUENCY, FEATURE SELECTION	BINARY, NO FEATURE SELECTION	TERM FREQUENCY, FEATURE SELECTION
KNN	87.6%	93.1%	90.2%	96.3%
Naïve Bayes	66.5%	64.9%	69.1%	68.1%
SVM	92.1%	89.8%	94.7%	97.9%
J48	94.7%	93.8%	97.3%	98.2%

Conclusion and future scope:

According to the study and observation we can see the potential of machine learning algorithm over traditional methods that are used by anti- virus tools for malware detection. And we have also discussed about different machine learning algorithms that can be of great help in detecting malware as with the quick development and advancement of web, malware is major threat. Malware are becoming widespread and more complex day by day. In this experiment, the focus lies on analysing and measuring the detection accuracy of the ML algorithm. We were able to train machine-learning algorithms to detect malware and benign files. From this experiment it is clear that by using static analysis based on information and selection of relevant features of the data can also give the best detection accuracy and can accurately represent malware. Furthermore, the

advantages of this method there is no need to execute or run malware and we can understand whether it is malware or not.

References:

[1]. Ye, Y., Chen, L., Wang, D., Li, T., Jiang, Q., Zhao, M., Nov 2008a. Sbmds: an interpretable string based malware detection system using svm ensemble with bagging. J. Comput. Virol. 5 (4), 283.

[2]. Ye, Y., Wang, D., Li, T., Ye, D., Jiang, Q., Nov 2008b. An intelligent pe-malware detection system based on association mining. J. Comput. Virol. 4 (4), 323–334.

[3]. Rieck, K., Trinius, P., Willems, C., Holz, T., Dec. 2011. Automatic analysis of malware behavior using machine learning. J. Comput. Secur. 19 (4), 639–668.

[4]. Sami, A., Yadegari, B., Rahimi, H., Peiravian, N., Hashemi, S., Hamze, A., 2010. Malware detection based on mining api calls. In: Proceedings of the 2010 ACM Symposium on Applied Computing. SAC 10. ACM, New York, NY, USA, pp. 1020–1025

[5]. Uppal, D., Sinha, R., Mehra, V., Jain, V., Sep. 2014. Malware detection and classification based on extraction of api sequences. In: 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2337–2342

[6]. D. Bekerman, B. Shapira, L. Rokach and A. Bar, "Unknown malware detection using network traffic classification," *2015 IEEE Conference on Communications and Network Security (CNS)*, 2015, pp. 134-142,

[7]. Okane, P., Sezer, S., McLaughlin, K., May, 2016. Detecting obfuscated malware using reduced opcode set and optimised runtime trace. Security Informatics 5 (1), 2.

[8]. Galal, H.S., Mahdy, Y.B., Atiea, M.A., May 2016. Behavior-based features model for malware detection. Journal of Computer Virology and Hacking Techniques 12 (2).

[9]. Boukhtouta, A., Mokhov, S.A., Lakhdari, N.-E., Debbabi, M., Paquet, J., May 2016. Network malware classification comparison using dpi and flow packet headers. Journal of Computer Virology and Hacking Techniques 12 (2), 69–100.

[10]. Salehi, Z., Sami, A., Ghiasi, M., 2017. Maar: robust features to detect malicious activity based on api calls, their arguments and return values. Eng. Appl. Artif. Intell. 59, 93–102.

[11]. Yuxin, D., Siyi, Z., Feb 2019. Malware detection based on deep learning algorithm. Neural Comput. Appl. 31 (2), 461–472.

[12]. Choudhary, S. and Sharma, A., 2021, February. Data Science Approach for Malware Detection. In Journal of Physics: Conference Series (Vol. 1804, No. 1, p. 012196). IOP Publishing.

[13]. Hafidi, S., F. Amounas, L. E. . Bermi, and M. . Hajar. “An Innovative Approach for Enhancing Cloud Data Security Using Attribute Based Encryption and ECC”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 9, no. 5, May 2021, pp. 01-06, doi:10.17762/ijritcc.v9i5.5471.

[14]. Choudhary, S. and Sharma, A. (2020). Malware detection & classification using machine learning. In 2020 International Conference on Emerging Trends in Communication, Control and Computing, ICONC3, pages 1–4.
(Accessed on Jan. 23, 2022: https://ieeexplore.ieee.org/abstract/document/9117547)

[15]. Deep, V., and H. Sharma. “SVM Classifier on K-Means Clustering Algorithm With Normalization in Data Mining for Prediction”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 7, no. 6, June 2019, pp. 29-34, doi:10.17762/ijritcc.v7i6.5318.

[16]. Sunita Choudhary, Anand Sharma “Malware detection in IoT using Machine Learning enabled Data Science Approach” CEUR Workshop Proceedings Vol.2823 (Accesses on Oct 16, 2021: http://ceur-ws.org/Vol-2823/Paper15.pdf)

[17]. Aditya Atreya, Khushbu Garg. (2021). Numerical Simulation and Design of Copy Move Image Forgery Detection Using ORB and K Means Algorithm. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 7(12), 30–41. Retrieved from http://www.ijfrcsce.org/index.php/ijfrcsce/article/view/2060

[18]. Puneet Sharma, Pawan Kumar Tiwari. (2022). Numerical Simulation of Optimized Placement of Distibuted Generators in Standard Radial Distribution System Using Improved Computations. International Journal on Recent Technologies in Mechanical and Electrical Engineering, 9(5), 10–17. Retrieved from http://www.ijrmee.org/index.php/ijrmee/article/view/369

[19]. Ahmed Cherif Megri, Sameer Hamoush, Ismail Zayd Megri, Yao Yu. (2021). Advanced Manufacturing Online STEM Education Pipeline for Early-College and High School Students. Journal of Online Engineering Education, 12(2), 01–06. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/47

[20]. N. A. Libre. (2021). A Discussion Platform for Enhancing Students Interaction in the Online Education. Journal of Online Engineering Education, 12(2), 07–12. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/49