A Data-Driven and Probabilistic Methodology for Malicious URL Prediction

[1] Surya V, [2] Ashitosh Arun Gupta, [3] Bhagya Shree, [4] Sree Sita Kanaka Naga Sai Sree Kandalam
[1]*Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, TN, India*
[2,3,4] *Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, TN, India*

Abstract **- This paper presents a novel implementation of applying the latest machine learning algorithms to detect a cyber intrusion or cyber-attack in web applications. Cyber Security is an ever-growing field, which with the emerging era of data science, has increased its manifold tenfold. Thus are the implications of data security and applications to detect the attacks on cyber medium. Whomsoever it may be, whether companies, businesses, or authorities from the government, it is essential to formulate an application to detect and prevent user data and information from cyber-attacks. This paper implements the latest growing pace of machine learning to implement the same.**

Index Terms **- Cyber Attack, Machine Learning, STM Algorithm, Random Forest, Cybersecurity.**

INTRODUCTION

With the ever-increasing use of the internet, user information security is a significant concern. Web applications may steal the user data such as mobile numbers, emails, bank details, transactional passwords, and much more are on high roads. With more than 2.9 M apps in the google play store, so is the threat and necessity to detect the malicious or faulty applications in the play store to protect the authentic data of the users pose under threat. Apart from apps, web applications' cyber-attacks on user data or information are a growing concern for companies. The following
Figure 1 gives the details of the ever-increasing attacks in the cyber net. [1] [2]

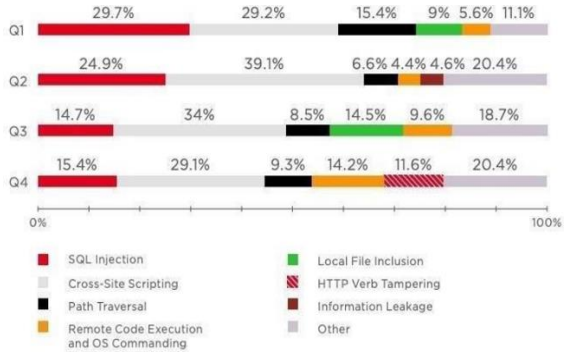
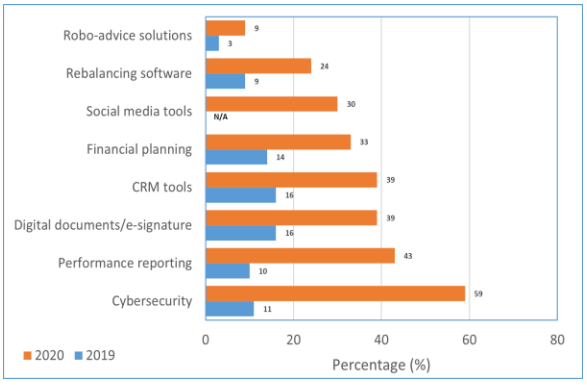


Figure 1: Web attack statistics in the year 2020 [3]

Information is the essential vita for any business or opportunity. A threat to the information or an attack on the user information or data is, on the other hand, an attack on the company itself. Hence, it is essential to formulate an implementation to detect and prevent cyber-attacks in the existing web applications. Companies and business units have started investing in cybersecurity to prevent data theft, cyber-attack using web applications on user data or client information etc. The following represents the major investments planned by the organization during the year 2019-2020. This will be able to provide a better understanding of the area of cybersecurity.

Figure 2: Investments in 2019-2020 [3]



STATE-OF-THE-ART

The existing works in detecting cyber-attacks on websites are merely based on programming techniques. Most of them employed open-loop technologies, i.e., feedback recognition systems implementation was not deployed. Systems used randomly employed specific datasets on anomaly applications to be categorized and fed into the systems for comparison with the existing application. The comparative results were used to detect whether the application is malignant or benign. These methods, though they seemed to be effective, had their advantages and disadvantages. The main disadvantage is that; it will not add new malignant applications to the existing database. Else the system will check for cyber-attack only with the existing old datasets. Moreover, the accuracy level in this type of cyber-attack detection in web applications is around 50-60 %. This level of accuracy detection shall not be applicable for companies and organizations dealing with credit cards, money transfers, bank, military operations etc., wherein data security is the most important, and it is not affordable to lose 40-50 %, if data is sent to cyber-attacks. [3] [4]

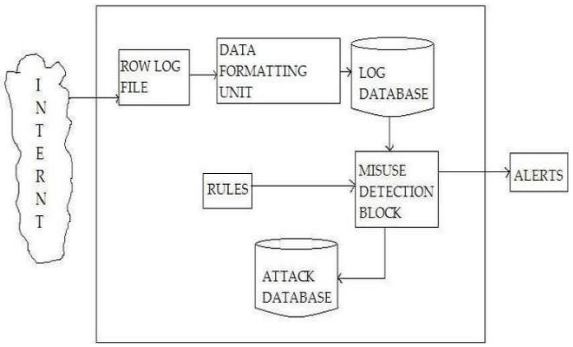


Figure 3: Block diagram of existing systems [7]

The block diagram (Figure 3) of existing cyber-attack detection systems can be described as follows: [4] [5]

- Cyber-attacks originated from the internet, creates a log file, which is then processed using the standard data processing systems and is stored in the database. It is then compared to the malignant database that has been stored in the server by simple comparative algorithms and a standard set of rules. The cyber-attack is being detected, and the user or the company is alerted for the cyber-attack on the same, so that the companies can take preventive measures.
- The vulnerability of attacks on websites is increasing. Given the above vulnerability ratio for the last decade, it has become a daunting task for developers to detect and classify web applications without specialized machine learning techniques. The existing systems propose checking for web apps from the list of applications in the cloud and then filtering the benign apps from the list. This process is time-consuming, and the results are not reliable. Moreover, no confusion matrices are applied to cross-check the correctness of the products.

CURRENT RESEARCH

In this paper, we have come up with the idea of using machine learning approaches for detecting cyber-attacks on websites. First, we have to gather a dataset of past malicious websites as a training set. With the help of the Support Vector Machine (SVM) algorithm and Decision Tree (DT) algorithm makeup comparison with the training dataset and trained dataset, we can detect the cyber-attack on web applications with close around 90 % result.

A novel implementation scheme to collect the app dataset, process them using the DT and SVM algorithm, prune the available data and classify them into malware and benign apps. With the help of machine learning algorithms such as SVM and DT, we intend to compare the training dataset and the trained dataset. SVM algorithms act as a classifier issued to classify the malicious application and benign app.

Improvised percentage of detection of the defective/faulty websites, implementation of ML algorithms such as SVM and Decision Tree, less time consumption and graphical results are the significant advantages of the proposed system. The gathered results are then added to the malignant and the benign datasets via a closed-loop system, making the system with the updated comparable dataset (see Figure *4*). [6]

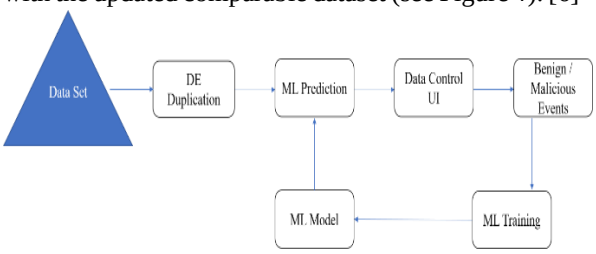


Figure 4: Block diagram of proposed website attack detection system

1.1 Data-set collection

The datasets for web applications are received and collected by multiple methods, ML approaches, existing system datasets, data from companies with firewalls, protected data classified as benign or malignant, and all the data sorted and stored the dataset to be compared. This is the primary step towards the detection of cyber-attacks in web applications.

1.2 File-level deduplication

Document-level information deduplication analyzes a record to be upheld up or chronicled with duplicates that are now put away. This is finished by checking its credits against a file. If the document is remarkable, it is put away, and the file is refreshed. Just a pointer to the current record is put away. The outcome is that just one example of the document is saved, and ensuring duplicates are supplanted with a reference that focuses on the first record (see Figure *5*). [7]

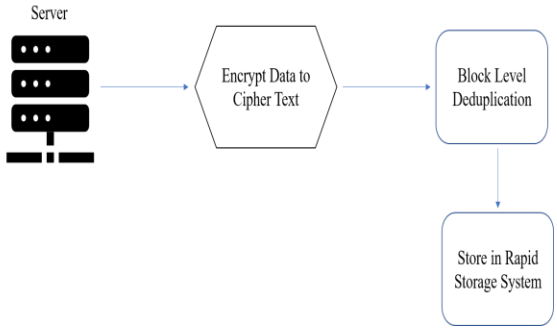


Figure 5: Detection deduplication application

Another mark match checking searches inside a record and saves exceptional cycles of each square. Every one of the squares is broken into lumps with a similar fixed length. The primary idea behind the data deduplication of the existing comparable datasets is to reduce the load on the server or application. This deduplication also reduces the memory consumed by the application, making the cyber-attack detection process less time-consuming. [8]

1.3 ML prediction

Machine Learning (ML) Prediction is the prime component of the detection system. This will compare the existing dataset with the new set using the Random Forest (RF) algorithm and SVM algorithm. Both the algorithms are deployed in the same prediction component, and the data is then curated for storage in the classification component (see Figure *6*). [7]

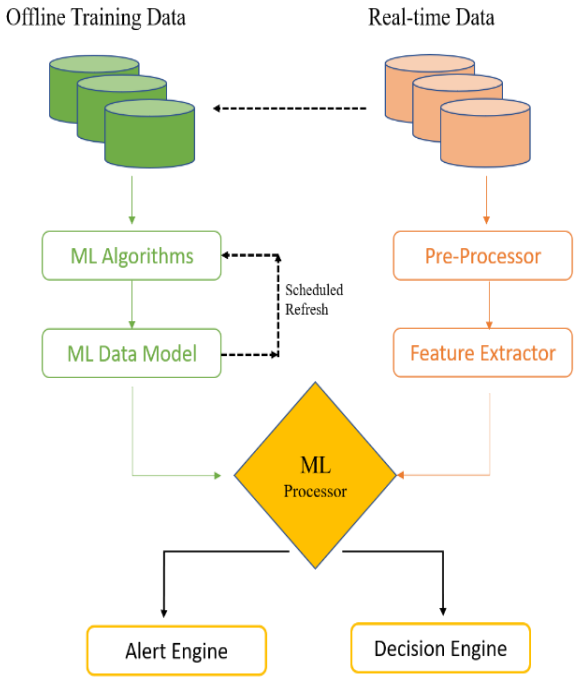


Figure 6: Machine learning implementation of URL phishing detection

The SVM and RF algorithm applications do the same job; it is necessarily important to compare the accuracy levels of both and save the information for future reference in the benign/malignant application dataset. The dataset is then used to compare the real-time dataset to conclude the attack. The results of these can be imparted to the alert or the decision engine. [10]

1.4 Random forest classification

An arbitrary backwoods classifier calculation is an indicator comprising of an assortment of randomized base relapse trees. These random trees are joined to frame the totaled relapse gauge. Due to multi-tree assessment, arbitrary woods grouping is moderately more reasonable to diminish the inclination or overfitting without affecting the trademark fluctuation. [10]

1.5 SVM classification

In the SVM classification, we plot every information thing as a point in n-dimensional space (where n is the number of highlights you have), with the estimation of each component being the estimation of a specific organization. At that point, we perform grouping by tracking down the hyper-plane that separates the two classes quite well (see Figure *7*). [11] [13]

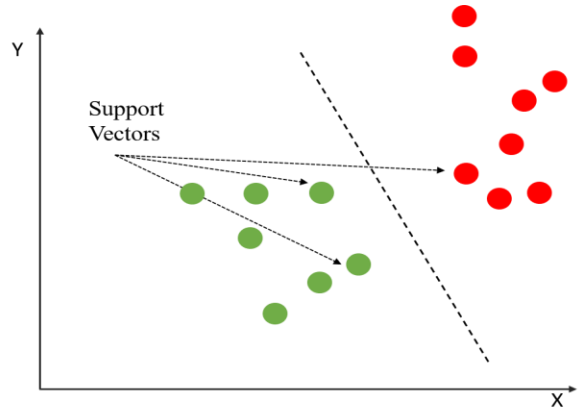


Figure 7: SVM classification of cyber-attack data [18]

1.6 Accuracy comparison

The obtained results are better when compared, and accuracy checked. An insight into the cyber-attack data classification detected 90 % accuracy for RF classification and 96 % accuracy for SVM classification. Figure 8 is an illustration of the same. Whatever be the accuracy detections, the data that are classified as malignant and benign are sorted out and stored in a separate dataset that can be used for future predictions. This will improve the efficiency of the system and the classification predictions. [14] [15]

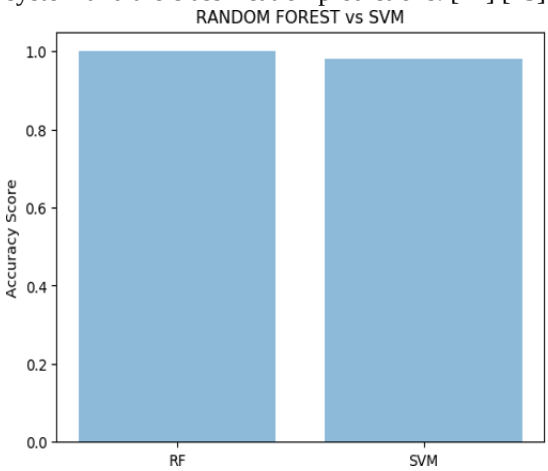


Figure 8: Result comparison for random forest and SVM

1.7 Data curation

As the data processing and the classification of the cyber-attack have been identified and processed, it is essential to store the processed or analyzed data in the form of log files. The data can be accessed for reuse from time to time. This is done in the data curation component. The processed data, classified data, benign and malignant data are stored in the server or disk with the proper annotation to be reused for better results. This can be done using the normal storage and indexing mechanisms but ensured that it is done.

1.8 Software system (Python 2.5 or 3.5)

Python is a significant level, deciphered, intelligent and object-arranged scripting language. Python is intended to be profoundly discernible. Python is an object-oriented programming language. Python upholds an object-oriented style or method of programming that exemplifies code inside objects. This open-source language can be used efficiently for various machine language algorithms, implementation, and effective classification. RF algorithm, SVM algorithm and data curation implementation can be effectively done using open-source language idle.

RESULTS

Various results can be categorized from the total list of websites that have been tested. The trained dataset of applications is checked using a graphical illustration and as a confusion matrix as listed below (see Figure *9*).

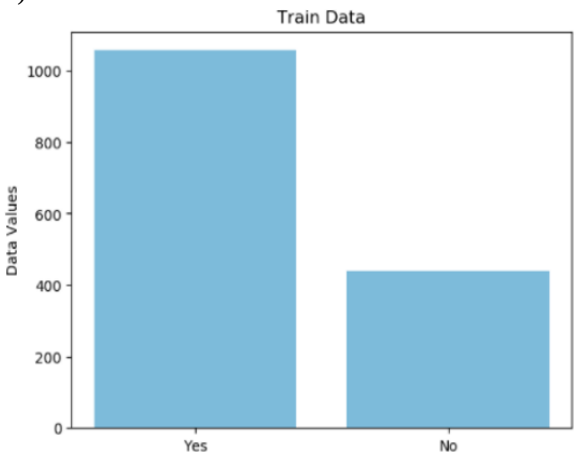


Figure 9: Screenshot of trained dataset graphics

The count and classification of malignant and benign web applications as per the raw data taken and tested (see Figure *10*).

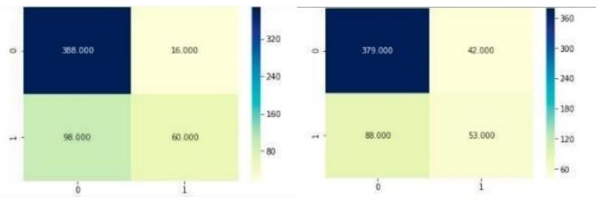


Figure 10: Graphical representation of cyber-attack detection

CONCLUSION

We got from learning security occasions typical conduct baselines and prompting the insignificant number of bogus positive security cautions. Moreover, ML investigations are most appropriate to dissect a tremendous volume of safety occasions and feed deviations from typical baselines into proactive danger chasing measures as pointers or leads of possible malevolent movement - besides, directed (double or multi). We have introduced various models for utilizing machine learning examination to upgrade digital protection checking alongside investigation on ideal calculations for normal digital dangers cases. An ML-based examination is a superb device to give setting class grouping calculations. ML has restricted task to carry out later on due to operational expense of getting ready great and terrible preparing information to prepare calculations models. Semi-directed (one-class grouping) calculations like One-Class SVM (OC-SVM) are moderately simpler to prepare, more financially savvy and more qualified to empower SOC.

REFERENCE

[1] J. J and C. M, "Cyberattacks—The Instability of Security and Control Knowledge," ISACA, vol. 5, pp. 1-5, 2016.

[2] C. R. Hollingsworth, "Auditing for FISMA and HIPAA: Lessons Learned Performing an In-house Cybersecurity Audit," ISACA, vol. 5, 2016.

[3] Positive Technologies, "Web application attack statistics: 2017 in review," 2018. [Online]. Available: https://www.ptsecurity.com/ww-en/analytics/web-application-attack-statistics-2017/?hcb=1. [Accessed 14 May 2021].

[4] TD Ameritrade Institutional, "RIA Sentiment Survey," 2020. [Online]. Available: https://s1.q4cdn.com/959385532/files/doc_downloads/research/2019/2019-RIA-Sentiment-Survey.pdf.

[5] X. Li, J. Wang and X. Zhang, "Botnet detection technology based on DNS," MDPI - Future Internet, vol. 9, no. 55, 2017.

[6] X. Yan, Y. Xu, B. Cui, S. Zhang, T. Guo and C. Li, "Learning URL Embedding for Malicious Website Detection," IEEE Transactions on Industrial Informatics, vol. 16, no. 10, pp. 6673-6681, 2020.

[7] P. B. Ambhore and B. R. Meshram, "Intrusion Detection System for Intranet Security," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 7, pp. 626-631, 2014.

[8] Y. Hu and Z. Ling, "DBN-based Spectral Feature Representation for Statistical Parametric Speech Synthesis," IEEE Signal Processing Letters, vol. 23, no. 3, pp. 321-325, 2016.

[9] D. MonDivakaran, K. W. Fok, I. Nevat and V. L.L.Thing, "Evidence gathering for network security and forensics," Science Direct - Digital Investigation, pp. S56-S65, 2017.

[10] M. Khan, "Managing Data Protection and Cybersecurity—Audit's Role," ISACA, vol. 1, 2016.

[11] S. Bromander, A. Josang and M. Eian, "Semantic Cyberthreat Modelling," STIDS, 2016.

[12] D. Marchette, "A Statistical Method for Profiling Network Traffic," in USENIX, USA, 1999.

[13] S. Fong, R. Wong and A. V. Vasilakos, "Accelerated PSO Swarm search feature selection for data stream mining big data," IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 33-45, 2016.

[14] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K_means Clustering," International Journal of Advance Research in Computer Science and Management Studies, vol. 1, no. 6, pp. 90-95, 2013.

[15] V. Mavroeidis and S. Bromander, "Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence," in European Intelligence and Security Informatics Conference: IEEE Society, Nowray, 2017.

[16] G. Gu, R. Perdisci, J. Zhang and W. Lee, "BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection," USENIX, 2008.

[17] J. Paparrizos and L. Gravano, "k-Shape: Efficient and Accurate Clustering of Time Series," SIGMOD- ACM org., pp. 1855-1870, 2015.

[18] Blackbelt digital, "Choosing the Best Classification Model for Machine Learning," [Online]. Available: https://www.blackbelt.

digital/choosing-the-best-classification-model-for-machine-learning/?hcb=1. [Accessed 14 May 2021].

[19] M. E. Celebi, H. A. Kingravi and P. A. Vela, "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm," Expert Systems with Applications - Elsevier, vol. 40, no. 1, pp. 200-210, 2013.

[20] C. Yang, F. Wang and B. Huang, "Internet Traffic Classification Using DBSCAN," IEEE - WASE International Conference on Information Engineering, 2009.