International Journal of Gender, Science and Technology
ISSN: 2040-0748
UGC Care Group I Journal.
Vol-13 Issue-02 July 2024
https://ijgst.com.2024.v13.i2.pp1067-1074

A new ordering system with AI classifiers to analyze and identify malicious Websites

Dr. PEDAPUDI SAMBASIVA RAO[1] B SUNIL KUMAR[2] CH.RAMA[3] M VENKATA RAMANA[4]
[1]ASSOC.PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
[2] ASST. PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
[3]ASST. PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
[4]ASST.PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
[1,2,3,4] SRI MITTAPALLI COLLEGE OF ENGINEERING

Abstract:

Nowadays, going online is an essential part of our daily routine. Therefore, in a bid to attract customers, various program merchants compete to build up new features and improved functionality, which in turn provide a source of attacks for the gatecrasher and put the sites at risk. However, current methods fall short in protecting users, who want a quick and accurate model with the capability to identify benign or malicious websites. This research paper details our plans for a new ordering system that will use artificial intelligence classifiers like random forest and support vector machine to analyze and identify malicious websites. Innocent Using Bayes, computed scores, and a few unique URLs (Uniform Resource Locators) based on hidden features, the classifiers can predict malicious website URLs. According to the first results, as compared to other AI classifiers, the arbitrary woods classifier achieves a higher level of accuracy (95 percent). Keyphrases: malicious website, artificial intelligence, detection, address, malicious websites.

Introduction

Clients are able to access an increasing number of services—e-learning, online banking, online commerce, long-distance interpersonal communication, online shopping, bill payment, and so on—through programs [4] or web applications as the web continues to rapidly evolve. There is a risk of losing personal and sensitive data since the apps are considered to have unique, advanced features and functions [3]. Because unsuspecting users aren't aware of the unique malware, malicious websites can easily capture them with just a single click. Once in, the intruders can exploit the page's vulnerabilities and inject the payloads to gain remote access to the victim's website. Consequently, in a dynamic online environment, the precise identifier of web pages is crucial. To address the issues, boycotting services were included into the programs; nonetheless, there are certain drawbacks, such as incorrect listing, [3]. Here we take a look at a self-learning approach to page order based on a small set of skills. We divide the site into two categories, benign and malicious, using four AI classifiers.

Related Work

According to the research specialists, there are three different ways to identify malicious pages: boycotting, static analysis, and unique analysis. We have covered a number of different approaches in rapid succession, and each one has a purpose. Using controlled AI techniques, Tao et al. [1] presented a new framework for automatically classifying the website as cruel or kind. It was determined that the

sites were malicious or not based on highlights. We culled helpful websites from a database. To address the problem of malicious website ordering based on highlights, Aldwairi et al. [3] introduced an additional lightweight self-learning method. An organised framework was developed that used the Genetic Algorithm (GA) to train classifiers capable of detecting malicious web sites. The Alexa database was considered for the friendly site and the Phis Tank database for the malicious site. An accuracy of 87% was found to be typical for the method.

A "Versatile SVM(aSVM) AI procedure" is used by Hwang et al. [5]. Due to its adaptability, the aSVM is prepared to handle fresh training data. One goal of aSVM is to make it less likely that newly created web pages would be misclassified.

Using the AI calculations K-NN and SVM, Yue et al. [6] suggested a method for organising harmful sites using 30 highlights. Compared to SVM, the output from K-NN was better. In order to determine the harmful online pages and specific forms of hazard, two classification methods were used.

Separately, Yoo et al. [4] suggested two methods for finding malicious websites: abuse detection and inconsistency discovery. Still, the false positive rate was somewhat high at 30.5%, even if the recognition rate was rather high at 98.9%. They used the WEKA instrument with the RafaBot dataset to guide their tests.

In order to identify the malicious website, Krishnaveni et al. [7] developed a collaborative tool called SpiderNet. The gadget was implemented in MatLab. Using three sets of inputs—specific normal highlights, diversion highlights, JavaScript highlights, and XSS attack highlights—the device ran two AI classifiers, mult I-SVM and ELM. ELM achieved a greater accuracy of 96.62% than multi-SVM, which came in at 93.22%.

Programed boycott generator (AutoBLG), developed by Sun et al. [8], can probe both previously unknown and malicious URLs. Using methods for URL Expansion, URL Filtration, and URL Verification, AutoBLG completed the task. A hybrid approach to the problem of harmful page identification was suggested by Wang et al. [10]. To train the classifier to detect potentially malicious websites, static analysis first isolated their static highlights. In the program motor, dynamic analysis runs the pages and deconstructs their dynamic behaviour. For the first, more assets had to be registered, and for the second, fake negatives were generated needlessly.
The WebMon page locator device, developed by Kim et al. [11], outperformed the standard tools by a margin of 7.6 t imes. An ML- and YARA-based commonsense model was suggested for malicious website page location. Additionally, a call tree computation was used to create a malicious divert particle tree with the capability to uncover the evil path.
The authors of this study "fostered a novel methodology of setting delicate and watchword thickness based for characterising the site pages with the assistance of regulated AI calculations (SVM, greatest entropy, ELM)." Despite the fact that many approaches to identifying potentially dangerous areas have been suggested. The main problem with these approaches is that they used tens of thousands of tests to get their findings, but they didn't address the challenges of collecting a diverse set of tests or how to identify harmful URL redirection, which is always changing.

https://ijgst.com.2024.v13.i2.pp1067-1074

To indentify the pages that are malevolent, three various methods i.e boycotting, static analysis , what's more, unique analysis are recommended by research experts. Each approach has some goal to fulfill and we have talked about a portion of these methods consecutively. Tao et al. [1] introduced a novel structure for distinguishing the page as vindictive or kind naturally utilizing managed AI approaches. The pages were recognized as malevolent or not founded on highlights. Benevolent web pages were gathered from dataset Aldwairi et al. [3] presented another lightweight self - learning way to deal with ordering the malevolent web page dependent on the highlights ordered structure was created which utilized the Genetic Algorithm(GA) to prepare classifiers that can identify the vindictive website pages. Dataset Alexa for Amiable site and Phis Tank for vindictive web locales were thought of. The normal system accuracy was discovered at 87%.

Hwang et al. [5] utilizes "Versatile SVM(aSVM) AI procedure" The aSVM can ready to manage new preparing information because of its versatile capacity. The target of aSVM is to lessen the likelihood of misclassification of new website pages.

Yue et al. [6] proposed a technique for arranging noxious pages utilizing 30 highlights with the assistance of AI calculation K-NN and SVM. The consequence of K-NN was superior to SVM. Two grouping models were utilized for identifying the pernicious website pages and explicit danger types.

Yoo et al. [4] proposed two kinds of discovery techniques: abuse identification and inconsistency discovery for distinguishing known and obscure malignant web pages individually. However the recognition rate was relative ly high up to 98.9% it's the bogus positive rate was high which is 30.5%. They have directed their test in WEKA instrument with dataset RafaBot.

Krishnaveni et al. [7] fostered a collaboration instrument, SpiderNet which had the option to recognize the malevolent site page. The device was carried out in MatLab. Two AI classifiers, mult I-SVM, and ELM were carried out in the device by taking three include sets to be specific normal highlights, divert highlights, JavaScript highlights , and XSS assault highlights showing higher precision in ELM(96.62%) than multi-SVM(93.22%).

Sun et al. [8] construct a system, programmed boycott generator(AutoBLG) which can have the option to investigate new and already obscure and malevolent URL. AutoBLG played out the undertaking by URL Expansion, URL Filtration, and URL Verification techniques. Wang et al. [10] proposed a cross breed way to deal with identifying pernicious site pages. Static analysis separated the static highlights of site pages and prepared the classifier to anticipate whether the page is malignant or not. Dynamic analysis executes the pages in the program motor and breaks down the dynamic conduct of the page. The first created unnecessary bogus negative and the

second one required additional registering assets.

Kim et al. [11] fostered a page locator WebMon apparatus which was 7.6 t imes quicker than the conventional instruments. For malevolent website page location, a commonsense model was proposed which comprises of WebKit-2, ML, and YARA based system. Too, a call tree calculation was introduced to make a pernicious divert particle tree that had the option to discover the malevolent way.

Altay et al. [12] "fostered a novel methodology of setting delicate and watchword thickness based for characterizing the site pages with the assistance of regulated AI calculations (SVM, greatest entropy, ELM)." Despite the fact that there are a few methodologies have been proposed for recognizing dabgerous locales. The principle burden of these methodologies is: to accomplish their results they utilized tens and hundreds and thousands of tests, , utilized no approach to recognize noxious URL redirection which is continually evolving, confronting the troubles in gather ing a wide range of tests.

System Analysis

Existing System

Internet banking, online commerce, informal communication, online shopping, online bill payment, e-learning, and countless more services are becoming available to customers as the web continues to rapidly evolve. These services are accessed through various web applications or programs [4]. Due to the applications' advanced features and functions, there is a risk of losing personal and sensitive data. Due to the fact that unsuspecting customers are unaware of the various forms of malware, they are easily caught by the gatecrasher with just a single click on malicious websites. The intruders then exploit the vulnerabilities on the victim's page by injecting payloads, gaining remote access to their website. Lost control over potential dangers. Expensive part since it is a reactive technique

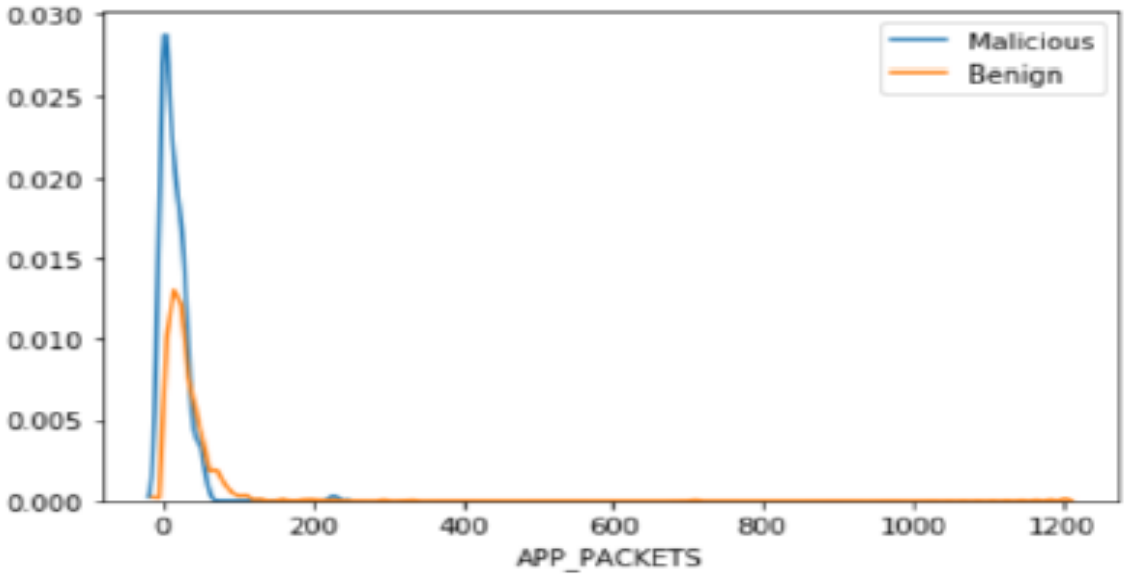
Proposed System

Using artificial intelligence classifiers like random forest and support vector machine, we want to develop a new ordering method to inspect and detect harmful sites. With the use of novel Bayes, selective relapse, and a few unusual URLs (Uniform Resource Locators), classifiers can now anticipate malicious websites based on the eliminated highlights. Identify and manage potential hazards at work. Ensure that your reps are attentive - and use it as a tool for preparation as well. In light of respectable safety procedures and reasonable requirements, establish risk the board standards. Minimise incidents in the workplace. To save money, avoid being reactive and instead take the initiative.

Methodology:

Dataset

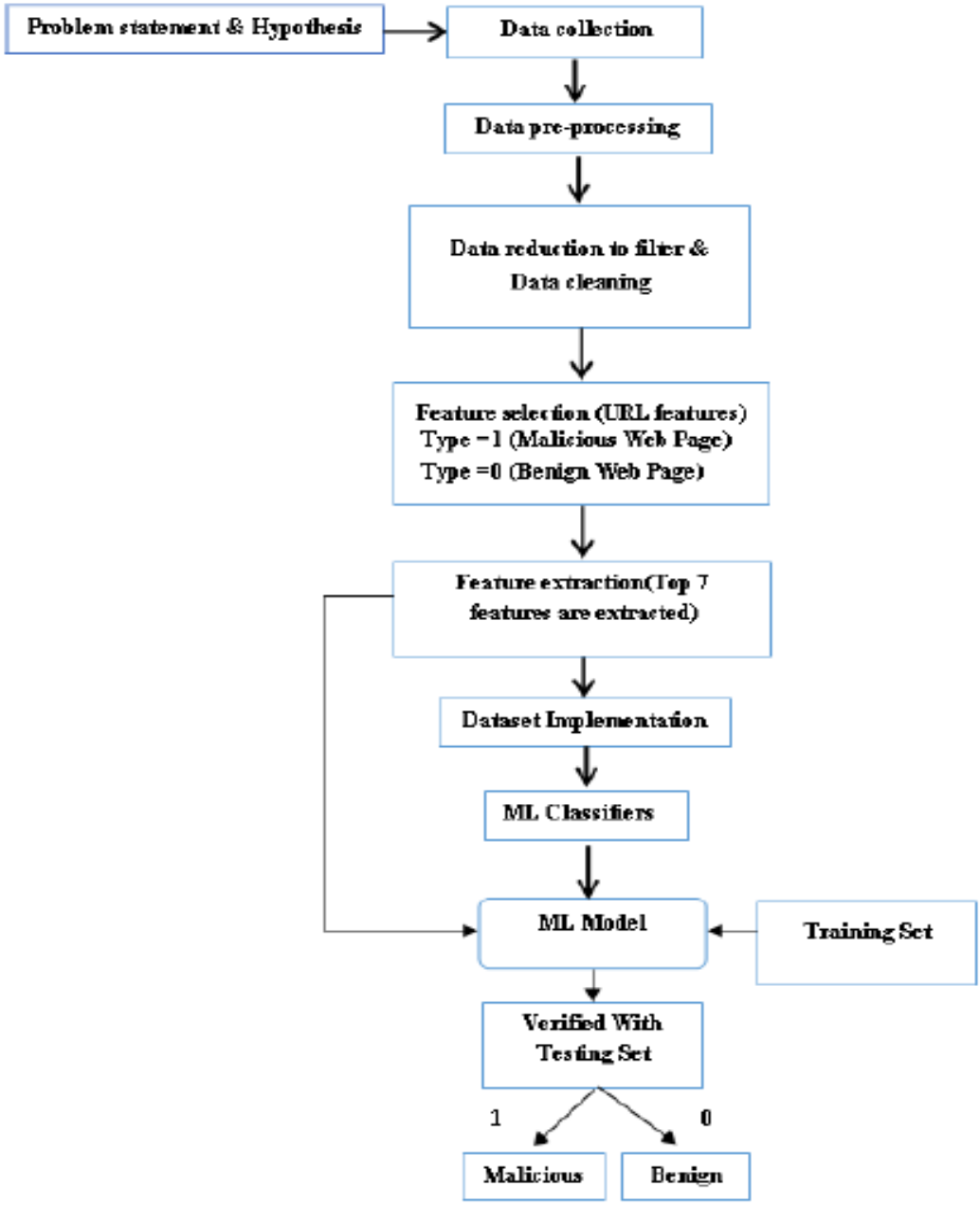
How order is defined is influenced by the datasets that are used. As anticipated, we must choose an appropriate dataset. To remedy this, we extracted a collection of URLs from the Kaggle dataset [14]. This has both harmful and helpful websites, along with 1782 records and 21 features. Only 812 records out of 1782 are actually used.



Feature Extraction

The characteristics are separated using several approaches. Because sometimes we can detect the harmfulness of pages somewhat by analysing the URL or by questioning the informat particle associated with the referred to have, its wellbeing can be distinguished, we have removed the features physically dependent on the URL in our examination work. Using URL characteristics has many advantages, such as avoiding downloading the actual page content and adaptability to various contexts, such as site pages and messages. Out of the twenty-one variables in the dataset, we extract seven that are essential to the language and the facilitation of URL. Fig. 14 addresses the two generally basic qualities, SOURCE-APP-PACKETS and Distant APP-PACKETS, that are part of our feature list. These properties significantly differ from other attributes when it comes to detecting malicious site pages. Furthermore, none of the current approaches make use of these two attributes. As a result, our method of placing orders is ultimately more effective than the alternatives. We keep track of the characteristics of the selected URL. The URL-based attributes are essential to our suggested discovery method, which use machine learning algorithms to distinguish between malicious and benign web sites.
Classifiers in Machine Learning There are several approaches to learning about classifiers. In order to construct our classifiers, we choose four machine learning computations.

https://ijgst.com.2024.v13.i2.pp1067-1074



Experimental Results Analysis

Classificat particle computations, including calculated relapse, arbitrary backwoods, Gaussian Guileless Bayes, and backing vector machine, have been used to finish many research. An easy python environment for information research, Jupyter Notebook [13] was used to develop and test all the studies. It allows for a more interesting and reasonable display of the code's development with its integrated support for Pandas, Scikit-Learn, Matplotlib, markup language, graphs, and tables. We then examine the four machine learning classifiers' displays. Since this is not a web

page, we have evaluated its performance using the exhibition metric, exactness. Therefore, the precision execution metric is crucial for optimal outcomes. We find that RF, a machine learning classifier, outperforms other classifiers on malignant page identification and achieves a greater accuracy of 95%. The results of the experiments show that our method achieves unparalleled performance with only a small configuration of URL-based characteristics.

Classifiers	Evaluation Criteria(Accuracy)
Gaussian NB	47%
SVM	89%
LR	91%
RF	95%

Conclusion

Emerging in the field of network security is the concept of malicious website page ID. While several studies have attempted to address the problems of malicious website age detection, they have been very resource- and time-intensive. In this research paper, we used a different site layout system that relies on URL characteristics to use machine learning algorithms to predict whether the website pages will be generous or malicious. One example of a machine learning classifier, Random Forest (RF), has a 95% accuracy rate. Our approach can successfully detect the malicious website page, according to the testing findings. It has been desired to expand feature sets and conduct analyses using other sources of data in order to enhance the classifier's performance in future work.

REFERENCE

[1] Tao, Wang, Yu Shunzheng, and Xie Bailin. "A novel framework for learning to detect malicious web pages." In 2010 Internat ional Forum on Informat ion Technology and Applicat ions, vol. 2, pp. 353-357. Ieee, 2010.

[2] Eshete, Birhanu, Adolfo Villafiorita, and KomministWeldemariam. "Malicious website detect ion: Effect iveness and efficiency issues." In 2011 First SysSec Workshop, pp.123-126. IEEE, 2011..

[3] Aldwairi, Monther, and Rami Alsalman. "Malurls: Alightweight malicious website classificat ion based on url features." Journal of Emerging Technologies in Web Intelligence 4, no. 2 (2012): 128-133.

[4] Yoo, Suyeon, Sehun Kim, Anil Choudhary, O. P. Roy, and T. Tuithung. "Two-phase malicious web page detection scheme using misuse and anomaly detect ion." International Journal of Reliable Informat ion and Assurance 2, no. 1 (2014): 1-9.

[5] Hwang, Young Sup, Jin Baek Kwon, Jae Chan Moon, and Seong Je Cho. "Classifying malicious web pages by using an adapt ive support vector machine." Journal of Information Processing Systems 9, no. 3 (2013): 395-404.

[6] Yue, Tao, Jianhua Sun, and Hao Chen. "Fine-grained mining and classification of malicious Web pages." In 2013 Fourth Internat ional Conference on Digital Manufacturing & Automation, pp. 616-619. IEEE, 2013..
[7] Krishnaveni, S., and K. Sathiyakumari. "SpiderNet : An interact ion tool for predict ing malicious web pages." In Internat ional Conference on Information Communication and Embedded Systems (ICICES2014), pp. 1-6. IEEE, 2014.
[8] Sun, Bo, Mit suaki Akiyama, Takeshi Yagi, Mitsuhiro Hatada, and Tatsuya Mori. "Automat ing URL blacklist generat ion with similarity search approach." IEICE TRANSACTIONS on Informat ion and Systems 99, no. 4 (2016): 873-882. Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2
[9] Urcuqui, Christ ian, Andres Navarro, Jose Osorio, and Melisa García. "Machine Learning Classifiers to Detect Malicious Websites." In SSN, pp. 14-17. 2017.).
[10] Wang, Rong, Yan Zhu, Jiefan Tan, and Binbin Zhou. "Detect ion of malicious web pages based on hybrid analysis." Journal of Informat ion Security and Applications 35 (2017): 68-74.74.
[11] Kim, Sungjin, Jinkook Kim, Seokwoo Nam, and Dohoon Kim. "WebMon: ML-and YARA-based malicious webpage detect ion." Computer Networks 137 (2018): 119-131.
[12] Altay, Betul, Tansel Dokeroglu, and Ahmet Cosar. "Context -sensit ive and keyword density-based supervised machine learning techniques for malicious webpage detect ion." Soft Comput ing 23, no. 12 (2019): 4177-4191.
13] website: ht tp://jupyter.org/
[14] ht tps://archieve.ics.uci.edu/ml/dataset/
[15] Vivek, Kolla, et al. "An Efficient Triple-Layered and Double Secured Cryptography Technique in Wireless Sensor Networks." *2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*. IEEE, 2021.
[16] Anusha, Pureti, T. Sunitha, and Mastan Rao Kale. "Detecting and Analyzing Emotions using Text stream messages." *ECS Transactions* 107.1 (2022): 16913.