iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45824

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Malicious Websites Classification Using Machine Learning Techniques: A Survey Paper

Shilpa M I[1], Poornima K M[2]
[1, 2]M.Tech, Dept.of CSE, JNN College of Engineering, Shivamogga 577201

Abstract: ***With the rapid rise of online development, Malware detection is critical for determining if a URL is hazardous or not because hackers steal user information such as usernames, passwords, and credit card numbers by impersonating a trustworthy entity via the internet and use it for illegal activities without the user's knowledge. As a result of applying many classifiers to detect URLs and conducting some operations, the best classifier was chosen as having a good performance in detecting URLs as malicious or benign.***
Keywords: URL, malicious, detection, Machine learning, classifiers.

I. INTRODUCTION

The number and variety of online resources, such as e-commerce websites, video sharing sites, and social networking sites, have exploded in recent years. Many studies show that the majority of financial and government businesses have expanded their online offerings to their customers. As a result of these actions, new ways for people to connect as well as new opportunities for other criminals are formed. Such attacks combine websites that mislead customers into giving over sensitive and private information such as passwords, Master card details, and so on, ultimately leading to identity theft, bank fraud, and in some circumstances, the introduction of malware into the customer's system. The attacker's redirection of the target to the perfect page is fundamental to the significant majority of attacks. So, recognizing and preventing these attacks on the user is a major challenge. The system's important information is protected by a malicious threat detection and prevention system, which plays a critical role in repelling these attacks. The URL plays minimal significance in detecting phishing websites in the preceding strategies. The URL, on the other hand, plays a vital part in detecting phishing websites. As a result, this work, will concentrate on URL attributes and ranking in order to detect phishing quickly and find real demand.

II. LITERATURE SURVEY

Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen, and Minh Hoang Nguyen[1]. To detect phishing sites, a new method was proposed that focuses on URL attributes (PrimaryDomain, SubDomain, PathDomain) and site ranking (PageRank, AlexaRank, AlexReputation). A new model for effectively detecting phishing sites. The suggested method, the system model, employs six heuristics to detect phishing sites, and to improve the detection ratio, the proposed technique was integrated with other techniques such as fuzzy, neural networks, and evolutionary algorithms. The system could then be improved further by integrating huge amounts of data and much more heuristic parameters.

Pradeepthi K V and Kannan A[2], Recognize phishing URLs simply by looking at the structure of the URL. As an outcome, the amount of time it takes to evaluate is substantially reduced because it won't be looking at the URL text or page data. The data are generally in the training phase when it is subjected to feature selection and categorization. When unknown data is presented through the user interface during the testing phase, the decision module, which is based on the rule base, classifies the unknown data based on the inferences formed during the testing phase.

Ying Xue, Yang Li, Yuangang Yao, Xianghui Zhao, Jianyi Liu, and Ru Zhang[3], A Vulnerable Sites List is established, which includes domains that are regularly targeted or have a high PageRank. A new method for calculating domain correlation is introduced. Turn multi-objective programming into single-objective programming by using weights to identify the best match URLs for every given site.

Shraddha Parekh, Dhwanil Parikh, Srushti Kotak, and Prof. Smita Sankhe[4], Increase efficiency while using Random forests as our classification algorithm, which can implement with the help of the Rstudio tool. Heuristic classification is applied to the parsed dataset. Other URLs of different websites that the user enters are predicted to use this model. Performance Analysis, which was executed using ROC Curve, was the final phase of the model to be accomplished. Other aspects of performance analysis, such as sensitivity, confusion matrix, Fp Rate, and so on, are included in the ROC Curve and help in good understanding.

Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, and Banu Diri[5], In general, there are two types of phishing detection systems: For identifying legal and phishing web pages, list-based phishing detection systems use two lists: whitelists and blacklists. Phishing detection systems that use whitelists create secure and genuine websites that deliver relevant information. URL records, often known as phishing websites are used to establish blacklists.

Yu Chen, Yajian Zhou, Qingqing Dong, and Qi Li[6], The neural network-based malicious URL detection system has two parts: online training and online testing. In the Training step, the system enters the predefined sandbox with labeled data combined with malicious and normal URLs.

The testing URL samples must be preprocessed during the testing step. The sandbox's testing images are given to the pre-trained CNN model for prediction, and the test results are split into two categories: harmful and benign URLs.

Sunita Choudhary and Anand Sharma[7], Many AI methods have been discovered reasonable for recognizable proof of malware classification and identification, aportion of the systems that have been utilized are Support Vector Machines (SVM), Random Forest, and unsupervised techniques like k-means have been used to cluster or group the malware based on their behavior, but the key challenge with clustering is to implement a consistent clustering strategy.

Jino S Ganesh, Niranjan Swarup V, Madhan Kumar R, and Harinisree A[8], To the algorithm, upload the data set. Take the training data set and divide it into parts, using 75% for training and 25% for testing the algorithm. Using a machine learning algorithm, train the data set.

Following the training, run a test with a 25% dataset. Now that the algorithm is ready to guess new data or unknown data, the outcome can be predicted.

Katherine Haynesa, Hossein Shirazia and Indrakshi Raya[9], Make a comparison of seven state-of-the-art deep learning algorithms show that ANNs can accurately identify phishing using URL and HTML- based features. Demonstrate that learning just URL-based features to detect phishing websites is ineffective.

Demonstrate that NLP models may be used to detect website phishing using only URL strings, indicating that pre-trained transformers perform similarly to other approaches in phishing detection.

Shantanu, Janet B, and Joshua Arul Kumar R[10], A lot of URLs have been used and misused to take advantage of a user's vulnerability.

This study focuses on determining whether or not a URL is benign or harmful. It also compares the outcomes of several machine learning classification approaches. To detect dangerous websites from the OpenPhish domain, the top-performing classifier is chosen.

Shubhankar, Siddhartha Bhaumik, and Prakash Biswagar[11], The approach for training the model is described, in which the seven machine learning algorithms are combined to produce the most adequate model for classifying benign and malicious websites. The most popular models are taken in for prediction at that stage when the models are split and then training is completed for each of the models for the various phases. The preparation is dominated by testing, in which the best-chosen model yields the best result.

Ripon Patgiri, Anupam Biswas, and Sabuzima Nayak[12], DeepBF is a new approach for detecting malicious URLs (deep learning and Bloom Filter). DeepBF is split into two parts. First, a Bloom Filter with self-adjustment using a 2-dimensional Bloom Filter. The best non-cryptography string hash function is determined through experimentation.

Then, by incorporating biases in the hashing procedure, designers derive an altered non-cryptography string hash function from the specified hash function for deepBF and compare the string hash functions. Other types of non-cryptography string hash functions are compared to the modified string hash function. Various test cases are used to compare it to various filters, including counting Bloom Filter and Cuckoo Filter. The test cases reveal the filters' weaknesses and strengths. Second, designers offer a deepBF-based mechanism for detecting malicious URLs.

Muhammad Fakhrur Rozi, Tao Ban, Sangwook Kim, Seiichi Ozawa, Takeshi Takahashi, Daisuke Inoue[13], Finding harmful scripts throughout a website is time-intensive and inefficient in terms of enhancing detection performance. To overcome these issues, designers describe a novel method for detecting fraudulent websites that involves examining the collective representation of a website's JavaScript stack. First, they create a collective graph representation of a website by pooling all JavaScript's abstract syntax trees.

The graph is then encoded into a vectorial form using graph2vec. Finally, for identifying potentially hazardous websites, machine learning-based detection is used.

Table 1 shows the details of the literature review, including the methodology, advantages and disadvantages of the entire article.

Table 1: Details of Survey Paper

Paper	Authors	Methodology	Advantages	Disadvantages
Novel Approach for Phishing Detection Using URL-Based Heuristic. (2014)	Luong Anh Tuan Nguyen, et al.	To detect phishing sites, it focuses on new aspects of URLs and site ranking.	A model for effectively detecting phishing sites.	To achieve more accurate probability.
Performance Study of Classification Techniques for Phishing URL Detection. (2014)	Pradeepthi K V, et al.	Identify phishing URLs only by understanding the structure of the URL.	Tree-based classifiers were shown to be the most suitable for phishing URL classification.	The accuracy of the system can be improved to obtain higher performance if it becomes dynamic.
Phishing Sites Detection Based on URL Correlation. (2016)	Ying Xue, et al.	Turn multi-objective into single-objective programming by using weights to identify the best match URLs.	Implementing new features to Increase Classification accuracy.	It does not Provide Accurate results.
A new method for Detection of Phishing Websites: URL Detection. (2018)	Shraddha Parekh, et al.	Using a Random forest classifier as a model.	The random forest has the highest level of accuracy, estimated to be around 95%.	There hasn't been a single solution to phishing.
Machine learning-based phishing detection from URLs. (2019)	Ozgur Koray Sahingoz, et al.	Use two lists: whitelists and blacklists. URL data, usually known as fake sites, are used to form blacklists.	Language independence, Real-time Execution, Detection of new Websites.	When URLs are shorter, it takes longer to detect them.
A Malicious URL Detection Method Based on CNN. (2020)	Yu Chen, et al.	The neural network-based solution for detecting malicious URLs.	Deep learning techniques have produced excellent results.	With a time delay, the system has precisely predicted.
Malware Detection & Classification using Machine Learning. (2020)	Sunita Choudhary, et al.	The use of AI strategies or methodologies for malware classification and identification has been discovered.	The rapid expansion and advancement of the web, it aids in the detection of malware.	If ML is not properly trained, the algorithm gives limited predictions.
Machine Learning Based Malicious Website Detection. (2020)	Jino S Ganesh, et al.	To the algorithm, upload the data set. Data can be separated into training and testing, and then predictions can be made.	Improve the accuracy of malware detection by building a good, efficient feature.	New malicious URLs that aren't on the list will be extremely difficult to identify.

Lightweight URL based phishing detection using natural language processing transformers for mobile devices. (2021)	Katherine Haynesa, et al.	With excellent efficiency, ANNs can identify phishing via URL and HTML based features.	It takes minimal time to learn and is safer to use.	It's difficult to tell whether the parsed web content is correct for analysis or not.
Malicious URL Detection: A Comparative Study. (2021)	Shantanu, et al.	It analyzes the outcomes of various ML classification methods. From the OpenPhish website, the best is utilized to detect.	The random forest algorithm achieves the best F1 score and accuracy.	The model has Accurately predicted with time delay.
Detection and Classification of Malicious Websites. (2021)	Shubhankar, et al.	To detect URL sites, use seven machine learning classifiers.	The study found the best performance in terms of phishing site classification.	It's difficult to find accurate training data.
deepBF: Malicious URL detection using Self adjusted Bloom Filter and Evolutionary Deep Learning. (2021)	Ripon Patgiri, et al.	To identify malicious URLs, an evolutionary Convolutional neural network was used.	It can be used in real-world projects to Successfully and efficiently filter out all Malicious URLs on a variety of devices.	Adding elements is never a bad idea, but it comes at the cost of a rising false positive rate.
Detecting Malicious Websites Based on JavaScript Content Analysis. (2021)	Muhammad Fakhrur Rozi, et al.	Encode each AST graph into embedding vector Representations using graph2vec.	On a real-world dataset, the suggested approach outperforms Existing approaches in terms of accuracy.	The graph2vec model is unable to handle unseen data.

III.CONCLUSIONS

Many cyber security approaches depend on malicious URL identification, and machine learning algorithms are a promising field. This research, applied machine learning techniques to perform a comprehensive and systematic survey on Malicious URL Detection. They provided a systematic formulation of Malicious URL detection from a machine learning perspective, followed by detailed discussions of existing studies for malicious URL detection, particularly in the form of developing new feature representations and designing new learning algorithms for resolving malicious URL detection tasks. The goal of this paper was to determine the model with the best performance.

REFERENCES

[1] Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen and Minh Hoang Nguyen, “A Novel Approach for Phishing Detection Using URL-Based Heuristic”, IEEE Conference Computing, Management and Telecommunication(ComManTel) , vol. 14, pp. 298-303, 2014.

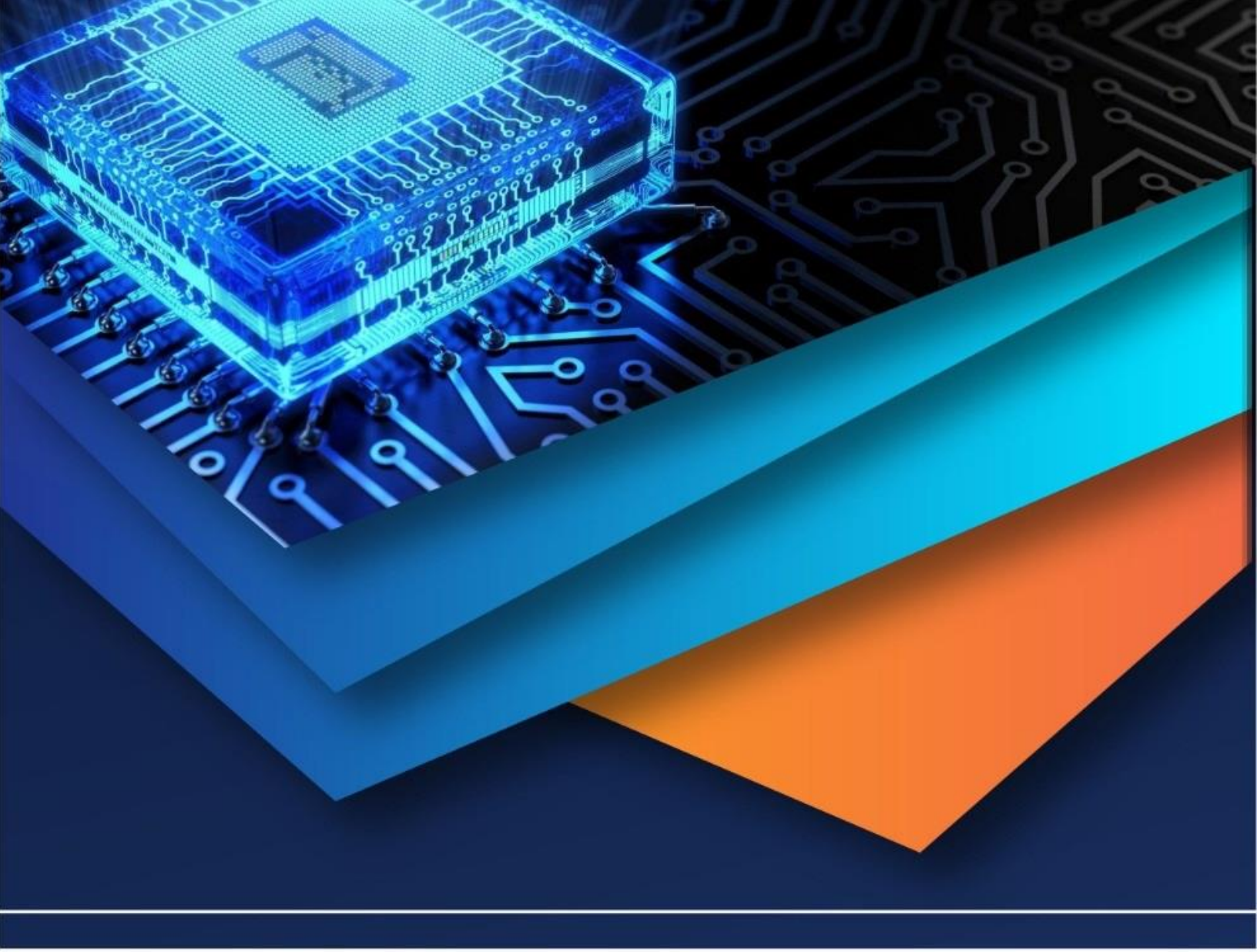
[2] Pradeepthi. K V and Kannan. A , “Performance Study of Classification Techniques for Phishing URL Detection”, Sixth International Conference on Advanced Computing, vol. 14, pp. 135-139, 2014.

[3] Ying Xue, Yang Li, Yuangang Yao, Xianghui Zhao, Jianyi Liu and Ru Zhang, “Phishing Sites Detection Based on URL Correlation”, Proceedings of CCIS, vol. 16, no. 31, pp. 244-248, 2016.

[4] Shraddha Parekh, Dhwanil Parikh, Srushti Kotak and Prof. Smita Sankhe, “A new method for Detection of Phishing Websites: URL Detection”, Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies, vol. 18, pp. 949-952, 2018.

[5] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir and Banu Diri, “Machine learning based phishing detection from URLs”, Expert Systems with Applications, pp. 345–357, 2019.

[6] Yu Chen, Yajian Zhou, Qingqing Dong and Qi Li, "A Malicious URL Detection Method Based on CNN", IEEE Conference on Telecommunications, Optics and Computer Science, vol. 20, pp. 23-28, 2020.
[7] Sunita Choudhary and Anand Sharma, "Malware Detection & Classification using Machine Learning", International Conference on Emerging Trends in Communication, Control and Computing, vol. 20, 2020.
[8] Jino S Ganesh, Niranjan Swarup.V, Madhan Kumar.R and Harinisree, "Machine Learning Based Malicious Website Detection", International Journal of Scientific & Engineering Research, vol. 11, No. 7, pp.113-138, 2020.
[9] Katherine Haynesa, Hossein Shirazia and Indrakshi Raya, "Lightweight URL-based phishing detection using natural language processing transformers for mobile devices", The 18th International Conference on Mobile Systems and Pervasive Computing, pp. 127–134, 2021.
[10] Shantanu, Janet B and Joshua Arul Kumar R, "Malicious URL Detection: A Comparative Study", Proceedings of the International Conference on Artificial Intelligence and Smart Systems, vol. 21, pp. 11147-1151. 2021.
[11] Shubhankar, Siddhartha Bhaumik and Prakash Biswagar, "Detection and Classification of Malicious Websites", Journal of University of Shanghai for Science and Technology, vol. 23, No. 6, pp. 120-131, 2021.
[12] Ripon Patgiri, Anupam Biswas and Sabuzima Nayak, "deepBF: Malicious URL detection using Self-adjusted Bloom Filter and Evolutionary Deep Learning", Transactions on Cybernetics, pp. 1-19, 2021.
[13] Muhammad Fakhrur Rozi, Tao Ban, Sangwook Kim, Seiichi Ozawa, Takeshi Takahashi, Daisuke Inoue, "Detecting Malicious Websites Based on JavaScript Content Analysis", ResearchGate, pp. 727-732, 2021.
[14] Swetha M S, Rakshith Danti, Praveen N M, and Muneshwara M S, "Classification of Malicious Websites using feature based Machine Learning Techniques", 5th Cyber Security in Networking Conference (CSNet), pp. 1-10, 2021



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY