A SYSTEMATIC APPROACH FOR FALLACIOUS URL DETECTION IN MACHINE LEARNING

P. Saraswathi[1], S. Harini[2], G. M. Premika[3], V. Shrinithi[4]

Article History: Received: 29.02.2023 **Revised:** 13.04.2023 **Accepted:** 06.06.2023

Abstract

Innocent Internet users are paying the price for the increasing prevalence of fraudulent websites, which generate billions of dollars in illegal revenue. There is a need for intelligent technologies to recognise harmful websites as online criminal activity rises. It has been demonstrated that URL analysis is a useful method for identifying phishing, malware, benign, and defacement. For URL categorization, previous studies have used lexical aspects, network traffic, hosting data, and other techniques. These methods necessitate time-consuming searches that cause real-time systems to experience severe delays. This paper represents a simple method for classifying dangerous websites that relies just on lexical URL analysis. To examine the accuracy of web URLs, machine learning models like Random Forest Classifier, XG Boost, and Light Gradient Boosting Machine are utilised.

Keywords: Malicious Websites, Lexical URL analysis, Defacement, Benign Malware and Phishing.

[1]Faculty, Department of Information Technology Velammal College of Engineering and Technology Madurai, Tamilnadu, India.
[2]IV yr Student, Department of Information Technology Velammal College of Engineering and Technology Madurai, Tamilnadu, India.
[3]IV yr Student, Department of Information Technology Velammal College of Engineering and Technology Madurai, Tamilnadu, India.
[4]IV yr Student, Department of Information Technology Velammal College of Engineering and Technology Madurai, Tamilnadu, India.

Email ID: [1]psw@vcet.ac.in, [2]harinisjan01@gmail.com, [3]premikamahesvaran@gmail.com, [4]srinithi30072002@gmail.com

DOI: 10.31838/ecb/2023.12.si6.330

1. INTRODUCTION

A scam website is any unreliable internet page that deceives users into committing fraud or malicious actions. Scammers utilise the anonymity of the internet to hide their true identities and motivations behind a variety of masks. These could consist of phone security alerts, gifts, and other dishonest formats that seem legitimate. Even while there are many valuable things you can do online, not everything is as it seems. Among the millions of legitimate websites vying for visitors' attention are websites made for a variety of illegal goals. These websites try everything, including credit card fraud and identity theft. Internet usage has ingrained itself into our daily lives as a result of the quickly developing technologies. Numerous aspects of our daily lives are decided after using the internet. The number of social networking sites has grown significantly in recent years. Internet users face numerous hazards as a result of their frequent use. Scam websites, like many other scam types, operate under different premises despite sharing similar mechanics. You'll be better equipped to recognise subsequent attempts as we specifically outline the types of premises a scam website may make use of. Here are some common formats of scam sites:

a) Phishing Scam Websites
b) Online Shopping Scam Websites
c) Scareware Scam Websites
d)Sweepstakes Scam Websites.

Phishing is a significant issue in today's environment. Phishing assaults, which use malicious websites linked to emails, SMS messages, or other forms of communication to trick people are based on social engineering and malware. Spam email is a tool used in cybercrime or fraud. Phishing was done via instant messaging or email spoofing. These texts and emails include a URL. Online shopping scam sites use a phoney or subpar online business to gather victims' credit card information, making it one of the most common methods. These frauds are problematic because they occasionally offer the goods or services to give the appearance of legitimacy. However, there is unavoidably poor quality. More importantly, it serves as an unregulated route for the unauthorised and excessive use of your credit card information. False security alert pop-ups are used by malicious websites to trick you into installing malware that poses as an actual antivirus product. They accomplish this by saying that your device has a virus or malware infection, which may cause you to download a fix out of fear or urgency. Users that don't have an actual internet security package may become victims of malware downloads, although having one would assist prevent this. Large rewards are given away in sweepstakes scams to encourage consumers to participate and eventually give their financial information to pay a fake fee. Due to the rise of e-commerce, many people are now vulnerable to phoney websites where criminals offer fake items or commodities that never materialise. These websites not only collect money from consumers, but they can also steal their credit card number or identity. Current user protection apps are based on blacklists and rules that have a high false-positive rate and require constant updating. To identify these domains based on newly published methodologies and current web page attributes, we constructed and made publicly available a suspicious of being fraudulent website dataset based on distinctive features, including seven novel features. On a dataset of 282 samples, our model achieved up to 75% F1- Score using the Random Forest technique and 11 custom features. The threat posed by malicious websites or URLs to cybersecurity is considerable. Each year, malicious URLs cause billions of dollars in damages by hosting unsolicited content (spam, phishing, drive-by downloads, etc.) and tricking users into falling for scams (including financial loss, identity theft, and malware installation). We have compiled this dataset to contain a large number of samples of malicious URLs in order to develop a machine learning-based model to identify dangerous URLs and stop them from corrupting computer systems or spreading over the internet. The Distributed Machine Learning Community (DMLC) has included it in a larger toolkit since Tianqi Chen first released it. The technique may be used for both classification and regression applications and has been developed to work with large and complex datasets. Microsoft's LightGBM is a distributed high-performance framework for

regression, classification, and ranking applications, much like XGBoost. It employs decision trees.

2. LITERATURE SURVEY

"Anti Phishing Simulator" is used for the prevention of infringements and enhances the security of information. It allows the user to create his own spam list if the malicious site is not available in the database[1]. The author proposed "Anti Phishing Simulator" for examining the website whether it contains malicious software and links. As a result, Support Vector Machine and Artificial Neural Networks provide the same outcome [2]. The author uses various techniques to categorise phishing emails by incorporating important structural elements. In[3], Naïve Bayes algorithm is implemented to identify the accuracy. Accuracy 89% is achieved by modeling Naïve Bayes algorithm with bagging method. Boosting approach is also used. The best accuracy is achieved using Naïve Bayes with bagging method which gives about 89% accuracy. Boosting and stacking methods used with the Naïve Bayes algorithm provides an accuracy of 85% and 51% respectively. Filtering email contents that assist distinguish between phishing scams and other deceptive attempts can improve the process. The goal of this strategy is to target the ad emails that phishers employ to successfully trick people into providing sensitive information or login passwords. In [4] The application blocks any harmful emails from reaching the system-integrated email address. This system also makes use of the current database and accesses keywords that can be utilised to read emails' text. In[5],The proposed system extracts the source code features, URL features and image features from the phishing website. The features that are extracted are given to the ant colony optimization algorithm to acquire the reduced features. To determine whether the webpage is real or phished, the reduced features are once more provided to the Nave Bayes classifier.

In[6], The advice of the author Education and Training Approach The battle against fishing laws relies heavily on education and training, by far its most important components. End users of business systems should be trained in phishing email detection. In addition to making it easier to recognise phishing emails, this will help the secure knowledge-sharing platform for information security, whose establishment we strongly recommend, get priority feed. In[7], Based on "Request URL" and "Website Forwarding", Phishing websites features are classified. In [8], the authors proposed the TDLBA method (Tuning Deep Learning using Bat Algorithm). By setting the parameters of Deep Learning networks, this technique combines swarm intelligence methods. In[9] At CANTINA more features are applied and machine learning techniques are also applied in which 92% True Positive (TP) and0.4% (FP).In[10]The author proposed a solution that calculates higher accuracy than the SVM classifier by using RBF. Using CS-SVM classifier accuracy of 99.52% is obtained. In[11],Phishing is defined as the practise of seducing users in order to get their personal information, such as user names and passwords. In this research, we develop a smart system that can recognise phishing websites. The foundation of this intelligent system is a machine learning model. By analysing the attributes of the phishing site and selecting the best system combinations for the classifier's training, we hope to improve classifier performance. In[12], authentication method is modeled to enhance the security of mobile cloud services for protecting the password attack in a cloud environment. A protocol verification tool named Scyther would be used to assess the performance. In[13],In this research, the author proposed the algorithm LinkGuard which uses the general properties of the hyperlinks in phishing assaults. These traits are ascertained through examination of the phishing data repository made available by the anti-phishing working group (APWG). LinkGuard can recognise both known and unidentified phishing assaults based on the general traits of phishing attacks. In[14],This study discussed a feature set that combines aspects of social networking and conventional heuristics. A suspect URL identification method based on Bayesian classification is also provided for use in social network settings. The suggested method achieved a high detection rate. This paper will outline a potential defence against such assaults by determining the supplied URLs are genuine or phishing. We have provided

datasets for 2000 phishing and 2000 legal URLs. Because of its effectiveness and accuracy, the author took the Random Forest Algorithm into consideration. By considering nine parameters, it evaluates whether a URL is safe to browse or a phishing URL. [15]. The new phishing detection method is modeled based on URL characteristics. The suggested strategy concentrates on how similar the URLs of authentic websites and phishing websites are. Additionally, a key consideration in determining if a website is a phishing site is its PageRank. With a sample of 11,660 phishing sites and 5,000 real sites, the suggested approach is assessed. The findings indicate that the method can identify over 97% of phishing sites[16].The authors of this study provide their results on how to identify phishing websites in order to do this. To get a good outcome, data mining techniques and classifier algorithms are combined[17].This essay makes an effort to categorise existing works in light of the information sources. The classification would not only enable our knowledge of the limits of the present approaches, but also vital information to design new anti-phishing strategies or enhance existing techniques[18]. Traditional classification procedures like blacklisting, regular expression, and signature matching approach are challenged by the large amount of data, changing trends and technology, as well as the complicated interactions between attributes. Binary classification issue is addressed which identifies the malicious URL. The author assesses the performance of many well-known machine learning classifiers. The author utilised a 450000 URL public dataset from Kaggle to train the model. The best classifier was used to identify dangerous URLs on the openphish website. It was found to generate better results. [19]. The tests were conducted on a dataset that originally had 1056937 labelled URLs (phishing and legal). Using several feature reduction techniques, this dataset's 22 distinct features were further condensed into a smaller collection. With a rate of 99.89% accuracy in identifying the tested URLs, Support Vector Machine (SVM) classifiers outscored Random Forest, Gradient Boosting, Neural Network, and other classifiers in the evaluation. By integrating this method with add-on/middleware features in web browsers, it may be used to alert online users whenever they attempt to reach a phishing website using nothing more than its URL [20]. The fraudulent alteration of a website, sometimes with the installation of new pages at URLs where none should exist, is a common source of internet security problems. The author of this study suggests a method for identifying these URLs only based on lexical characteristics, allowing the user to be warned before actually retrieving the website[21].

Through web spoofing, users are persuaded to engage with bogus websites rather than legitimate ones. Taking sensitive information from the users is the assault's main objective. An imitation of the real website known as a "shadow" website is made by the attacker. By verifying the Uniform Resources Locators (URLs) of suspicious online sites, the author's suggested technique may tell a real web page from a phoney one. URLs are scrutinised according to a set of criteria in order to identify phishing websites [22]. The author used a machine learning technique to categorise URLs as hazardous or benign. After addressing the class imbalance issue, a number of classification models that were developed using a range of classification techniques are fed the URL dataset. In order to rank the qualities according to their importance and reduce the number of features required for categorization, a feature selection strategy is also used. Additionally, Apriori, FP-Growth, and Decision Tree Rules are used to build IF-THEN rules, which help establish links between the attributes. [23].

Despite its great efficiency, conducting business online is particularly conducive to fraud and other sorts of dishonest behavior since it is a "remote" activity. In order to distinguish between URLs that are legitimate and those that are not, the author used a neural network as a binary classifier of machine learning. We evaluated the model's performance using binary classification accuracy [24]. The dynamic nature of phishing websites inspired this work to provide a novel method based on deep reinforcement learning to model and recognise bogus URLs. The suggested model may learn the characteristics related with phishing website detection by

adapting to the dynamic behavior of phishing websites [25].

Existing Work

There are currently two main trends in the problem of malicious URL detection: Characteristics-based or rule-based harmful URL detection and behavioral analysis-based malicious URL detection. Precise identification of malicious URLs is made possible by malicious URL detection techniques based on a collection of flags or criteria. This approach, however, cannot detect new dangerous URLs that do not match established characters or restrictions. Malicious URL detection techniques based on behavioural analysis approaches use machine learning or deep learning algorithms to classify URLs according to their activity.

Proposed Work

Based on the features and behaviour of URLs, it employs machine learning techniques to categorise them. Features are newly developed in the literature and are derived from the static and dynamic behaviour of URLs. The primary contributions of the research are these recently proposed features. The system's overall approach to detecting fraudulent URLs includes machine learning methods.

1. Upload Dataset

URL data sets are created, consisting of over 200 URLs. These URLs came from malicious URL-sharing websites or online public data that were crawled. We scanned this information from Phish Tank and Virus Total for dangerous URLs. Most regularly, public records like DMOZ are used to download regular URLs.

2. Pre-Processing

Since "http://" and "https://" are frequently used, they can be safely excluded from detection. Typically, a URL is made up of letters, numbers, and some symbols. After professional processing, the records can be made more concise in the next step to reduce time and resource consumption. Special characters like "_" and "#"are excluded from our methodsince we don't think they significantly affect the categorization outcomes.

3. Classifying URL

Here, we contrast SVM's performance in classification with that of other well-liked machine learning techniques. We have chosen a number of well-known classification algorithms. We employ various parameter settings for each algorithm in an effort to enhance its performance. Classify URLs as spam, malware, phishing, and tampering using the SVM algorithm.

4. Graphical Representation

The analysis of attack kinds inside network datasets makes up a significant portion of the project. Charts can be used to analyse user data for the data. Here, administrators can look for particular answers to their system's problems. A visual representation of the collected data is displayed in the form of graphics. Different diagrams provide the best analysis of the system.

5. Feature Selection

The below mentioned features are chosen for analyzing the web-related data:

1) Address bar based features
2) Domain based features
3) HTML & Javascript based features

Feature Distribution

The graphical representation for the proposed system is given below:

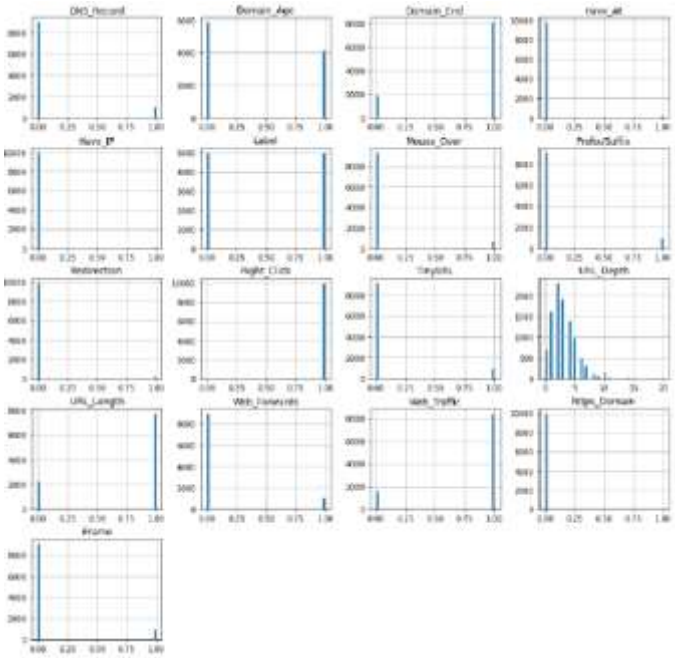

Figure 1: Parameters for classification

Machine Learning Models

In this paper, supervised machine learning models like classification and regression were used to predict the web URL is Benign, Defacement, Phishing and Malware. The dataset has been trained by classification methods Random Forest, XG Boost and LGBM classifiers. Accuracy of web URL is measured using the regression techniques Random Forest and XG Boost. The accuracy of the above models are compared and among them XG Boost provided the more accurate result.

RandomForestClassifier

	precision	recall	f1-score	support
benign	0.99	0.99	0.99	149
defacement	0.88	0.97	0.92	37
phishing	1.00	0.43	0.60	7
malware	0.80	0.80	0.80	10
micro avg	0.96	0.96	0.96	203
macro avg	0.92	0.80	0.83	203
weighted avg	0.96	0.96	0.96	203

accuracy: 0.961

Figure 2: Random Forest Classifier

XGBClassifier

	precision	recall	f1-score	support
benign	1.00	1.00	1.00	149
defacement	0.93	1.00	0.96	37
phishing	1.00	0.43	0.60	7
malware	0.82	0.90	0.86	10
micro avg	0.98	0.98	0.98	203
macro avg	0.94	0.83	0.85	203
weighted avg	0.98	0.98	0.97	203

accuracy: 0.975

Figure 3: XG Boost Classifier

LGBMClassifier

	precision	recall	f1-score	support
benign	1.00	1.00	1.00	149
defacement	0.92	0.97	0.95	37
phishing	0.75	0.43	0.55	7
malware	0.82	0.90	0.86	10
micro avg	0.97	0.97	0.97	203
macro avg	0.87	0.83	0.84	203
weighted avg	0.97	0.97	0.97	203

accuracy: 0.970

Figure 4: LGBM Classifier

S. No	Machine Learning Model	Accuracy
1	Random Forest	0.961
2	**XG Boost**	**0.975**
3	LGBM Classifier	0.970

Figure 5: Comparison between the Classifiers

The above shows that XGBoost classifier gives the most accurate prediction result.

3. IMPLEMENTATION AND RESULTS

A new user could register by providing their Username, Address, e-Mail and Mobile Number. The user must provide the password for security. The registered user information are stored in a database which can be used for login purpose. After the user has been successfully logged in, the home page is displayed. Then, the user can navigate to upload single URL menu and upload the URL which they need to check. After clicking the submit button, the result is displayed. The user could predict the accuracy of three machine learning models and they can able to view the graphical representation of classified URLs in the trained datasets.

Figure 6: Home page

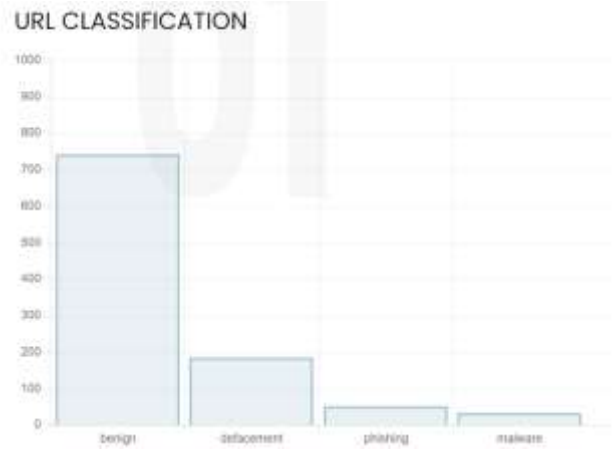

Figure 7: Graphical Representation

4. CONCLUSION AND FUTURE WORK

Here ML models like Random Forest, XGBoost and LGBM (Light Gradient Boosting Machine) are used to predict the web URL as fraudulent or not. The web URLs are identified and trained by the classifiers and predicted the same by using the regression techniques. Accuracy of best model is 96%. Accuracy may be increased by introducing the other machine learning models which will provide good results in future. The more datasets can be included for analysing many number of web URLs.

5. REFERENCES

Y. Sonmez, Turker Tuncer, Huseyin Gokal & Engin Avci (2018). "Phishing web Sites Features Classification Based on Extreme Machine Learning". 6th International Symposium on Digital Forensic and Security.

Basnet, R., Mukkamala, S., & Sung, A. H. (n.d.). Detection of Phishing Attacks: A Machine Learning Approach. Studies in Fuzziness and Soft Computing, 373–383.

Gyan Kamal and Monotosh Manna, Detection of Phishing Websites Using Naive Bayes Algorithm, Proceeding of International Journal of Recent Research and Review, Vol. XI, Issue 4 December 2018, ISSN 2277-8322.

Baykara, M., & Gurel, Z. Z. mm(2018). Detection of phishing attacks. 2018 6th International Symposium on Digital Forensic and Security 355389(ISDFS).

R.Priya (2016), "An Ideal Approach for Detection of Phishing Attacks using Naive Bayes Classifier". International Journal of Computer Trends and Technology(IJCTI). ISSN: 2231-2803.

Singh, P., Maravi, Y. P. S., & Sharma, S. (2015). "Phishing websites detection through supervised learning networks". 2015 International Conference on Computing and Communications Technologies.

M. Kaytan and D. Hanbay "Effective classification of Phishing Webpages Based on New Rules by Using Extreme Machine Learning" Anatolian Journal of Computer Sciences, AJCS 17, pp: 15-36, ISSN: 2548-1304, 2017.

Vrbanˇciˇc, G., Fister, I., Podgorelec, V.: Swarm Intelligence Approaches for Parameter Setting of Deep Learning Neural Network. In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics—WIMS '18, pp. 1–8 (2018)

Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). CANTINA+A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites. ACM Transactions on Information and System Security, 14(2), 1–28.

Niu, W., Zhang, X., Yang, G., Ma, Z., & Zhuo, Z. (2017). Phishing Emails Detection Using CS-SVM. 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications

W. D. Yu, S. Nargundkar and N. Tiruthani, "A phishing vulnerability analysis of web based systems," 2008 IEEE Symposium

on Computers and Communications, Marrakech, 2008

Ellappan, Munivel & A, Kannammal. (2019). New Authentication Scheme to Secure against the Phishing Attack in the Mobile Cloud Computing. Security and Communication Networks. 2019

J. Chen and C. Guo, "Online Detection and Prevention of Phishing Attacks," 2006 First International Conference on Communications and Networking in China, Beijing, 2006

Chia-Mei Chen, D.J. Guan, Qun-Kai Su, "Feature set identification for detecting suspicious URLs using Bayesian classification in social networks," Information Sciences, vol.289, December 2014.

Ma, Justin, et al. "Beyond blacklists: learning to detect malicious web sites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

Nguyen, Luong Anh Tuan, et al. "A novel approach for phishing detection using URLbased heuristic." Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.

Aydin, M., Butun, I., Bicakci, K., & Baykal, N., "Using Attribute-based Feature Selection Approaches and Machine Learning Algorithms for Detecting Fraudulent Website URLs",IEEE 10th Annual Computing and Communication Workshop and Conference, Ankara, Turkey,Goteborg, Sweden, Guzelyurt, Cyprus, 30-May- 2020.

Shahriar, Hossain, and Mohammad Zulkernine. "Information source-based classification of automatic phishing website detectors." In 2011 IEEE/IPSJ International Symposium on Applications and the Internet, IEEE, 2011.

Shantanu, B Janet, R Joshua Arul Kumar, "Malicious URL Detection: A Comparative Study" ,2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS).

Mohammed Abutaha; Mohammad Ababneh; Khaled Mahmoud; Sherenaz Al-Haj Baddar,"URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis",2021 12th International Conference on Information and Communication Systems (ICICS).

Abdulghani Ali Ahmed; Nurul Amirah Abdullah,"Real time detection of phishing websites",2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).

Akshay Sushena Manjeri; Kaushik R.; Ajay M.N.V.; Priyanka C. Nair;"A Machine Learning Approach for Detecting Malicious Websites using URL Features",2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA).

Jasmina Novakovic; Suzana Markovic,"Detection of URL-based Phishing Attacks Using Neural Networks",2022 International Conference on Theoretical and Applied Computer Science and Engineering (ICTASCE).

Moitrayee Chatterjee; Akbar-Siami Namin,"Detecting Phishing Websites through Deep Reinforcement Learning" ,2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC).