IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 3, March 2022

DOI: 10.17148/IARJSET.2022.9340

SURVEY ON ENHANCING THE PERFORMANCE OF ANTIPHISHING MECHANISM USING MACHINE LEARNING

Roopesh Kumar BN[1], R Soumya[2], Sri Chandana P[3], Vijetha[4], Sushmitha S[5]

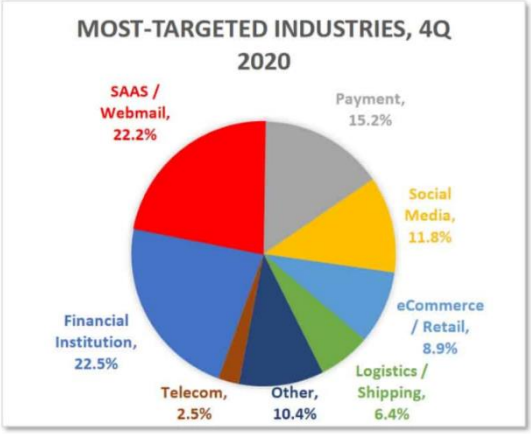
Department of Computer Science and Communication Engineering, K.S Institute of Technology, Bangalore, India[1-5]

Abstract: A Phishing is one of the most potentially disruptive actions that can be performed on the Internet. Phishing sends malicious links or attachments through emails that can perform various functions, including capturing the victim's login credentials or account information. It is a form of identity theft, in which criminals build replicas of target websites and lure unsuspecting victims to disclose their sensitive information like passwords, PIN, etc. It is one of the social engineering methods that gathers personal information through malicious websites and deceptive e-mail to canvass personal information from a company or an individual [5]. In terms of website interface and uniform resource locator (URL), most phishing webpages look identical to the actual webpages**.** Various strategies for detecting phishing websites, such as blacklist, heuristic, Etc., have been suggested. There is a demand for an intelligent technique to protect users from the cyber-attacks. In this study, we are trying to propose a URL detection technique based on Machine learning approaches. Boosting method is employed to detect phishing URL.

Keywords: Anti-phishing, Phishing types, Phishing websites, Phishing detection techniques, Cyber security, Machine learning classifiers.

I. INTRODUCTION

Phishing is considered as one of the malicious use of internet resources where the user are tricked into revealing their personal information, username and password and other personal information to the attacker. Phishing can appear through a variety of communication forms such as instant messaging, SMS, VOIP, online messenger [5]. The fake webpage contains input forms requesting personal critical information such as credit card, social security numbers, mother's maiden name, etc. Although existing spam filtering techniques can be employed to combat phishing emails, these measures are not entirely scalable [2] According to the RSA's online fraud report, the year 2013 has been confirmed to be a record year where many phishing attacks have been launched globally. Additional1y, RSA estimates that over USD $5.9 billion was lost by global organizations due to phishing attacks at the same period. The Internet Security Threat Report 2014 reports that cybercrimes are prevailing and damaging threats from cybercriminals still emerge over businesses and customers [4]. There's been a marked change from previous years, though, with Software as a Service (Saas) and webmail attacks dropping from 31.4% to 22.2% in a single quarter. As such, financial institutions are now the most common target,



accounting for 22.5%. Meanwhile, attacks on eCommerce platforms and payment platforms have both risen by a few percent. The success of phishing website detection techniques mainly depends on recognizing phishing websites accurately and within an acceptable timescale. The conventional URL detection approach is based on a blacklist (set of malicious URLs) obtained by user reports or manual opinions. However, these techniques are not efficient enough, since

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 3, March 2022

DOI: 10.17148/IARJSET.2022.9340

a new website can be launched within few seconds. For instance, cybercriminals can use a Domain Generation Algorithm (DGA) to circumvent the blacklist by creating new malicious URLs. Thus, an exhaustive blacklist of malicious URLs is almost impossible to identify the malicious URLs. Thus new malicious URLs cannot be identified with the existing approaches[11]. This paper develops an anti-web spoofing solution based on inspecting the URLs and content of fake web pages. This solution developed takes series of steps to check characteristics of websites URLs.

Here we propose Enhancing the performance of anti-phishing mechanism scheme using a CNN-LSTM model based on best selected features, to detect phishing websites with high accuracy[1]. In addition, Boosting method is implemented using XG-Boost technique in detection of phishing websites. The CNN-LSTM module, which contains data pre-processing, feature extraction and classification. In feature extraction, the CNN is used to extract local features and LSTM is used to extract context dependency. The result of the CNN-LSTM is then used by XG-Boost for classification.

II. TYPES OF PHISHING

1] Spear phishing: Spear phishing is known as the phishing of phishing tentative targeting individuals or businesses. Phishers generally got the information of individuals through social media sites such as Linkedln, Facebook and use of fake addresses for sending emails that similarly happens to be the mail that was received from anyone of our co-workers [12].Example: The attacker is encouraging the target to sign an updated employee handbook. This is an example of a spear phishing email where the attacker is pretending to work in HR and is encouraging the target to sign a new employee handbook.

2]Whaling: Whaling refers to attacks on spear phishing targeted directly senior management and other high-profile objectives [12]. In many whaling phishing attacks, the attacker's goal is to manipulate the victim into authorizing high-value wire transfers to the attacker. The use of Whaling emails and malware-infected websites are the most notorious whaling methods used to perform the necessary actions.
Example: In November 2020, the co-founder of Australian hedge fund Levitas Capital followed a fake zoom link that installed malware on its network.
The attackers attempted to steal $8.7 million using fraudulent invoices. In the event, they only got away with $800,000. But the reputational damage was enough to lose Levitas its biggest client, **forcing the hedge fund to close.**

3**]** Catphishing and catfishing: Catphishing (spoiled with a 'ph') is something of an online disappointment, involving knowing someone closely to gain access or control over the behaviour of the user in the use of information or services. In some cases, a catfisher steals other individual's complete identity including their date of birth, image and location and pretends that it is their own.

4] Tab nabbing: Opening multiple tabs at a time is an advantage of tab nabbing. Redirecting the user to affected site and other types. Reverse technique is method loaded here that is copying the affected sites into the original site happens here. The objective of tab-nabbing is similar to phishing. In which attacker send their similar links to victims as victims would not get the difference and it leads to loss of their sensitive data.

5] Clone Phishing: Clone phishing is little different than a phishing attempt. A clone phishing attack uses a legitimate or previously sent email that contains attachments or links. A legitimate web-application is cloned to make the user to believe he is signing in a genuine form. The preventive measure is that user has to follow the two-step verification by this clone phishing can avoidable.

6] Pharming: Pharming, a type of attack being used where stems from domain name system (DNS) cache poisoning is done. Pharming can be done either by changing the host file on a user's computer or by exploitation of a vulnerability in DNS server software [13]. Prevention for this is to ensure whether user is using secure web connections or not. And it can be prevented by avoiding suspicious websites.

7] Deceptive phishing: Deceptive phishing is one of the common phishing attack. Attacker impersonates a sender in this fraud and collects all the personal details like login credentials. These emails will enforce them to click on the links and make them to change the secured details like password or ask them to make a payment.

III. METHODOLOGY

Different sorts of literature studies were reviewed in this study, as well as a survey and phishing detection done. And by analysing all these we were able to identify many different types of phishing techniques for solutions. It's critical to employ high-quality datasets in phishing detection. We built a system that incorporates data pre-processing,

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 3, March 2022

DOI: 10.17148/IARJSET.2022.9340

categorization, and feature extraction into one package. In this study, we combine CNN and LSTM to get efficient and more organized datasets. When it comes to feature extraction, The local features are extracted using CNN, while the context-dependency is extracted using LSTM. The CNN-LSTM model's classification output is then delivered to the XG-Boost model.

WORKING FLOW

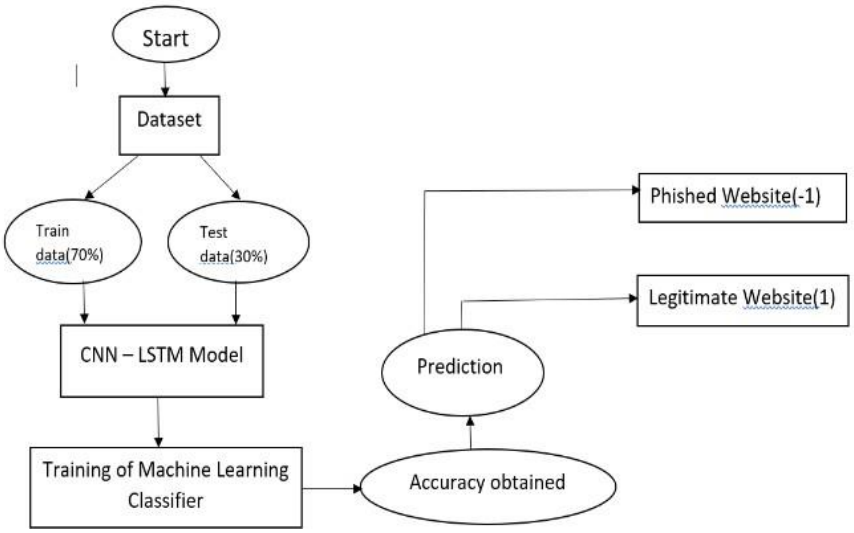


Fig b Flow chart of the methodology

The dataset contains phishing and legitimate URLs records. The URLs were collected from various web sources such as Kaggle and UCI repositories. To get a better accurate value to divide the dataset into two halves, one for data training (70 percent) and the other for data testing (30 percent). The CNN-LSTM model receives the outcomes of the training and testing datasets. Feature extraction and data pre-processing are included in the CNN-LSTM technique. Processed output is sent for classification. The XG-Boost approach is used for classification after the classification accuracy has been established. The trained model predicts whether the given input (URL) is Phished URL or a Legitimate URL.
CNN-LSTM model- For optimum performance, the CNN and LSTM algorithms are blended. This idea of a hybrid model leads to the formation of a new deep learning scheme. In terms of accuracy, a hybrid of CNN and LSTM will give better performance. Here the CNN-LSTM model's operational concept and architecture are described. The output of the extracted features as described, is connected with the CNN layer and the other layer is applied to minimize the dimensionality of CNN output. Next, connect the LSTM layer to it. In the end, we apply the sigmoid activation function and dense output layer to make a better decision as Phished or Legitimate websites.
Feature-Extraction

In CNN-LSTM model feature-extraction is done. It extracts the multiple different types of features of URL using CNN-LSTM. The features are like:

- URL length
- Prefix or suffix
- Having subdomains
- ##
- @
- %
-
- ^
- &
- *
- ?
- /

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 3, March 2022

DOI: 10.17148/IARJSET.2022.9340

- concat
- Console
- https and http
- shortening service
- Internet protocol address
- Domain registration link

Many more features like this are extracted here to find out whether the entered URLs are phished or legitimate websites.
Data pre-processing:
Data set is a collection of the dataset. Data collected from the online sources are in the raw form, it may contain errors so it requires correction for that purpose Data pre-processing has been done. Data pre-processing is frequently used to assess noise, aberrations, and standardize data.
Data pre-processing has been done in the CNN-LSTM model, where it flows in this way.
Data Cleaning: It removes incomplete data and replaces missing values.
Data Integration: It combines multiple sources into a single set.
Data Transformation: It recognizes the raw data into useful dataset.
Data Discretization: It makes data evaluation and management easy.
XG-Boost model:
XG-Boost, a supervised learning system, employs the boosting strategy. Boosting is an ensemble learning method wherein each tree tries to remove the errors and it also takes over the previous errors or weaknesses to the next tree to remove them completely. By this we can get more efficient values. In this study, the result of the CNN-LSTM model is used by the XG-Boost for classification. This algorithm is mostly used as a classifier for mapping the URL's pattern into a specific class. The strengths of this algorithm are better regularization, parallelism, reduces overfitting, Optimization and give better performances. These advantages have made it an excellent tool for classification.
After the classification of datasets, the model predicts whether the entered URL is phished URL(0) or a Legitimate URL(1).

IV. LITRATURE REVIEW

Mohamed Ben Haj Frej, Fathi Amsaad, Abdul Razaque proposed [1] " Detection of Phishing Websites using Machine Learning" Oct 2020. This focuses on identifying phishing websites using website features and blacklist database. Techniques are using blacklist database which contains URLs of all phishing websites. Outcome of this paper identifies specific websites that have a history of spam. And works properly with google chrome extensions.

S.Dilip Kumar, A.Nazarath Fathima, Sifina, U.Jamruth Kani [2], "Preventing phishing attacks using Evolutionary Algorithms". It focuses on selective significant features that discriminate between legitimate and phishing URLs. This detects phishing URL using Binary classification with phishing URLs belonging to positive class and negative class. It prevents phishing attack and to provide high level security based on SVM classification.

Thomas Philip, Jain James, Nisha Mohan P.M [3], APJ Abdul Kalam Technological University, Kerala, India "Anti-Phishing Technology Using Machine Learning Approach". It focuses on using different approaches to enhance the accuracy of the system, providing an efficient protection system. The security issues related to the URL is checked using the machine learning and decision is made accordingly. This new system can be designed to avail better accuracy.

S.Carolin Jeeva and Elijah Blessing Rajsingh [4], "Intelligent phishing URL detection using association rule mining" Uses selected features to differentiate phishing and legitimate. Uses associative ruleapriori and predictive apriori rule generation algorithm. Identifies recurring patterns obtained by frequently occurring features in phishing URLs.

Varun Vyas, Aditya Nair, Allan lopes[5], Information technology department, University college of engineering, Mumbai, India "Heuristic based malicious URL detection" Focuses on phishing prediction based on set of features. This uses heuristic-based approach to classify phishing URLs by using only the information available on URLs and detects new and temporary phishing sites that evade existing blacklist- based technique.

Mohammed Abutaha, Mohammad Ababneh,[6] Princess Sumaya University for Technology Amman, Jordan "URL Phishing Detection using ML Techniques based on URLs lexical analysis" To detect and categorize the malicious URLs according to zero hour phishing attacks. URLs dataset's features mainly depend upon analysing the URLs lexically. High result and effective approach in handling the imbalanced dataset.

DOI: 10.17148/IARJSET.2022.9340

Karim Hashim et al. [7] proposed Mobile Phishing Websites Detection and Prevention Using Data Mining Techniques. The widespread use of smart phones nowadays make users vulnerable to phishing. Mobile devices facilitate phishing attacks due to the following properties. Firstly the rapid increase of mobile users worldwide. Secondly the limited screen sizes makes it difficult for mobile users to determine legitimate web-page from phishing one. To minimize time wastage and system resources consumption, a System Data Base [SDB] have been utilized. Check if the domain name is an IP as a verification of their identities, legitimate websites use their company, institute or services names as a domain name. This work models the prediction of phishing websites on mobile devices as a classification task and demonstrate the machine learning approach to predict the websites status and take the proper action towards it.

Routhu Srinivasa Rao et al. [8] concentrates on URL and Website Content of phishing page. PhishShield takes URL as input and outputs the status of URL as phishing or legitimate website. The heuristics used to detect phishing are footer links with null value, zero links in body of html, copyright content, title content and website identity. PhishShield is able to detect zero hour phishing attacks which blacklists unable to detect and it is faster than visual based assessment techniques that are used in detecting phishing.

Hemali Sampat et al. [9] proposed a system which detects the phishing using features of URLs and WHOIS protocol. They used classification and association Data Mining algorithms to identify and characterize all rules and factors in order to classify the phishing website and relationship that correlate them with each other to detect them by their performance, accuracy, number of rules generated and speed.

Ram B.basnet et al. [10] focusses to study the anatomy of phishing URLs that are created with the specific intent of impersonating a trusted third party to trick users into divulging personal data. Unlike previous work in this area, that only use a number of publicly available features on URL alone. In addition, compares performance of different machine learning techniques and evaluates the efficacy of real-time application of methods used. Applying it on real-world data sets, they demonstrate that the proposed approach is highly effective in detecting phishing URLs. It uses a heuristic-based approach to classify phishing URLs by using the information available only on URLs. It treats the problem of detecting phishing URLs as a binary classification problem with phishing URLs belonging to the positive class and benign URLs belonging to the negative class.

V. CONCLUSION

Any country's success depends on its ability to keep its citizens safe and protected. Over time, the use of internet apps has increased. Hence providing web security is the essential part. When utilizing any application, users want to feel safe and secure. However, in today's world, security has become a difficult issue in all areas, and trust is the institution that should keep us safe.

Phishing websites can lead to huge types of fraud and put the user in a difficult situation. We proposed a system that efficiently identifies phishing threats to handle this human element. It will protect users from fraudulent acts and identity theft while they are online. This will also assist many reputable websites in maintaining positive client Relations.

REFERENCES

[1] Aliya CH1, dr.D.Loganathan, “Deep learning approach for phishing attacks”. International Research Journal of Engineering and Technology (IRJET) ”, Feb 2021 .

[2] Mohamed Ben Haj Frej, Fathi Amsaad, Abdul Razaque, “Detection of phishing websites using machine learning,” 2020 IEEE Cloud Summit Conference Paper.

[3] S.Dilip kumar, A.Nazarath fathima, Sifina, U.Jamruth kani, “Preventing Phishing Attacks using Evolutionary Algorithms,” 2019 International Research Journal of Engineering and Technology (IRJET) ,2019, Volume 06.

[4] S.Carolin Jeeva and Elijah Blessing Rajsingh, “Intelligent Phishing url detection using association rule mining,” 2016 Human-centric Computing and Information Sciences (HCCIS),2019.

[5] Varun Vyas, Aditya Nair, Allan Lopes , “Heuristic based malicious URL detection,” 2020 International Journal for Research in Engineering Application & Management (IJREAM), ISSN:2454- 9150 Vol-06, 2020.

[6] Mohammed Abutaha, Mohammad Ababneh, Khaled Mohmoud, Sherenaz Ai-Haj Baddar, Priness Sumaya , “URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis,” 2021, 12th International Conference on Information and Communication Systems (ICICS), Conference Paper, 2021.

[7] Karim Hashim Al-saedi, Mustafa Dhiaa Al-Hassani Mustansiriyah University, Baghdad, Iraq “This paper explains Mobile Phishing Websites Detection and Prevention Using Data Mining Techniques”.

DOI: 10.17148/IARJSET.2022.9340

[8] Routhu Srinivasa Rao and Syed Taqi Ali "PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach" Department of Computer Engineering, National Institute of Technology, Kurukshetra, Haryana, India.

[9] Hemali Sampat Manisha Saharkar, Ajay Pandey and Hezal Lopes "Detection of Phishing Website Using machine Learning" IRJET,2018.

[10] Ram B.basnet Andrew H.Sung, Quingzhong Liu Colorado "Learning to detect Phishing URLs".

[11] Ashit Kumar Dutta, "Detecting phishing websites using machine learning techniques" Department of Computer Science and Information System, College of Applied Sciences, Almaarefa University Riyadh, Saudi Arabia, Oct 2021.

[12]. A S S V Lakshmi Pooja, Sridhar.M, "Analysis of Phishing Website Detection Using CNN and Bidirectional LSTM", Department of Computer Science and Engineering, GRIET, Hyderabad, Telangana, India. (ICECA-2020).

[13]. Achu Thomas Philip, Jain James, Nisha Mohan P.M. "Anti-Phishing Technology Using Machine Learning Approach". (IRJET) Volume: 07 Issue: 05 | May 2020.

