Phishing Websites Detection Using Machine Learning

P. Amba Bhavani ASST.PROFESSOR, Department of Information Technology, Maturi Venkat Subba Rao (MVSR) Engineering College, Email: bhavani_it@mvsrec.edu.in

Chalamala Madhumitha, Department of Information technology, Maturi Venkata Subba Rao (MVSR) Engineering, Hyderabad, India. Email: 245118737011@mvsrec.edu.in

Pinnam Sree Likhitha, Department of Information technology, Maturi Venkata Subba Rao (MVSR) Engineering, Hyderabad, India. Email: 245118737029@mvsrec.edu.in

Chanda Pranav Sai, Department of Information technology, Maturi Venkata Subba Rao (MVSR) Engineering, Hyderabad, India. Email: 245118737004@mvsrec.edu.in

Abstract: ***The availability of multiple services such as online banking, entertainment, education, software downloading, and social networking has accelerated the Web's evolution in recent years. As a result, a massive amount of data is constantly downloaded and transferred to the Internet. This allows attackers to access sensitive personal or financial data such as usernames, passwords, account numbers, and social security numbers. This is known as a Web phishing attack, and it is one of the most serious issues in web security. Web phishing attempts have become much more common in recent years, and phishing is now considered one of the most hazardous Web crimes, with potentially disastrous consequences for online businesses. The phisher constructs a fake or phishing website to deceive Web users and get their sensitive financial and personal information in a Web phishing attack. Phishing attacks are the most straightforward method of obtaining sensitive information from unsuspecting consumers. The goal of phishers is to obtain sensitive information such as usernames, passwords, and bank account numbers. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. By extracting and evaluating numerous aspects of authentic and phishing URLs, this project uses machine learning technology to detect phishing URLs. Phishing websites are detected using CNN LSTM & CNN Bi-LSTM, Logistic regression and XGBoost algorithms. The project's goal is to detect phishing URLs and narrow down the best machine learning algorithm by evaluating each algorithm's accuracy rate, false positive rate, and false negative rate***.

Keywords* : *phishing, legitimate, trust worthy, false positive, false negative.

I. INTRODUCTION+

Phishing is an attempt to steal personal information such as usernames, passwords, and credit card numbers (and, indirectly, money) by impersonating a trustworthy entity in an electronic contact, usually for harmful purposes. Because the use of bait in an attempt to catch a victim is comparable, this word was coined as a homophone of fishing. Phishing is most commonly carried out through email spoofing or instant messaging, and it frequently urges people to submit personal information on a false website that looks and feels very identical to the authentic one. Victims are frequently lured by communications appearing to be from social media sites, auction sites, banks, online payment processors, or IT administrators. Links to malware-infected websites may be included in phishing emails. Phishing is a type of social engineering method that takes advantage of flaws in current web security to deceive consumers. Legislation, user training, public awareness, and technical security measures are all being used to combat the rising number of reported phishing instances. Many websites have developed additional tools for applications, such as game maps, however they should be clearly labelled as to who built them, and users should not use the same passwords across the internet.

II. LITERATURE SURVEY

[1] P.A. BARRACLOUGH, G. SEXTON, N. ASLAM (2015)

Phishing assaults are on the rise, resulting in millions of dollars in losses each year, particularly in online transactions. Toolbars and filters that display user warnings against phishing websites have been used in the past to combat phishing attempts. Despite current solutions, there is still an inadequacy in online transactions due to a lack of accuracy in real-time solutions. This study builds on our prior work by creating an online toolbar that runs in the background of the Internet Explorer web browser and checks all websites users request in real-time against a set of data. To detect phishing websites and notify users from phishing assaults, the proposed approach is a feature-based online toolbar with six sets of inputs that includes a voice-generating user warning interface with text directives and color status. The new toolbar system was thoroughly tested on a wide range of websites, including 200 Phishing websites, 200 Suspicious Websites, and 200 Legitimate Websites, and it showed the best performance (96 percent) when compared to prior field results. The research presents a novel voice-generating user warning interface technique that has not previously been studied in the field of phishing website identification.

[2] ABDULGHANI ALI AHMED, NURUL AMIRAH ABDULLAH (2016)

Web spoofing entices users to connect with bogus websites instead of the actual ones. The primary goal of this assault is to steal confidential information from users. The attacker develops a 'shadow' website that appears to be identical to the original site. This deception allows the attacker to view and edit any information the victim provides. This study presents a phishing website detection technique based on inspecting web page Uniform Resource Locators (URLs). The proposed approach checks the Uniform Resources Locators (URLs) of suspected online pages to distinguish between real and fraudulent web pages. To detect phishing web pages, URLs are examined based on specific features. The discovered assaults are reported to help avoid future attacks. The suggested solution's performance is assessed using the Phistank and Yahoo directory datasets. The acquired findings demonstrate that the detection method is deployable and capable of detecting various forms of phishing attempts while minimizing false alarms.

[3] G KUMARI, M NAVEEN KUMAR, A MARY SOWJANYA (2017)

A large number of people buy things online and pay for them using numerous websites. Multiple websites frequently request sensitive information such as a user's username, password, or credit card information for authentication. However, there are certain phishing websites. Which then uses that information for nefarious purposes. We developed a flexible and successful solution based on data mining algorithms to detect and anticipate phishing websites. To classify their authenticity, we used the Logistic Regression algorithm and approaches. In the final phishing detection rate, various significant features such as URL, domain identity, and security can be used to detect the phishing website. Many internet users can use this tool to protect themselves from an ocean of phishing sites. This system's data mining algorithm gives higher performance. This technique also allows users to purchase things online without fear of being scammed. Admi scan can also add phishing website URLs or phoney website URLs to the system, which the system can browse and scan. When a user submits it, new suspicious URLs can be added.

[4] AKANSHA PRIYA, ER. MEENAKSHI (2017)

Phishing sites are imitations of legitimate websites created by dishonest individuals. These websites resemble the official websites of any corporation, such as a bank or an educational institution. Phishing's main goal is to steal sensitive information from users, such as passwords, usernames, and pin numbers. Phishing victims may reveal sensitive financial information to attackers, who may use this information for budgetary and criminal purposes. To identify phishing sites, various technical and non-technical ways have been presented. Non-technical approaches have no defense against phishing websites' ability to vanish quickly. One of the classes of technical approach is data mining, which has demonstrated good results in detecting phishing websites. Data mining techniques, as opposed to non-technical approaches, can develop classification models that can provide real-time predictions on phishing websites. The WEKA tool was used to analyze the C4.5 (J48) data mining technique in this work. C4.5 is a data mining benchmark technique that can accurately detect phishing websites. The method J48, which is a WEKA version of the C4.5 algorithm, was trained using a dataset of 750 URLs. The classifier developed during J48's training is utilized to make predictions using a testing dataset of 300 URLs. After the testing process, the true positive rate, true negative rate, false positive rate, false negative rate, success rate, error rate, and accuracy are calculated. C4.5 has an accuracy of 82.6 percent, according to the results.

[5] VAIBHAV PATIL, TUSHAR BHAT (2018)

Annually, phishing costs Internet users a lot of money. It refers to attacks that take advantage of weaknesses on the user's side. Because there is no single method to effectively mitigate all vulnerabilities in the phishing problem, numerous techniques are used. We cover three methods for identifying phishing websites in this research. The first strategy examines various elements of the URL; the second examines the authenticity of the website by learning where it is hosted and who manages it; and the third way examines the website's visual look. For evaluating these diverse aspects of URLs and webpages, we use Machine Learning techniques and algorithms. In this paper, an overview about these approaches is presented.

[6] AMANI ALSWAILEM, BASHAYR ALABULLAH (2019)

Phishing websites are one of the internet security issues that focus on human vulnerabilities rather than program flaws. It's the technique of luring online users in order to get sensitive information like usernames and passwords. We present an intelligent technique for detecting phishing websites in this research. The technology works as an addon to an internet browser, notifying the user when it finds a phishing website. The system is based on supervised learning, which is a type of machine learning. We chose the Random Forest approach because of its high categorization performance. Our goal is to develop a higher-performing classifier by analysing the characteristics of phishing websites and selecting the best combination of them to train the classifier. As a result, we conclude our work with a 98.8% accuracy rate and a total of 26 characteristics.

III. METHODOLOGY

1.Logistic regression

Logistic regression is a regression model where the dependent variable (DV) is categorical. Based on one or more independent variables, logistic regression is a mathematical approach for predicting the probability of a binary response. Given certain variables, logistic regression is used to predict a result with two values, such as 0 or 1, pass or fail, yes or no, and so on. The logistic regression, like the rest of the regression models, is a prognostic study. It's commonly used to visualize data and highlight the relationship between one

dependent binary and one or more nominal, ordinal, interval, or ratio-level independent variables. A more complicated cost function is also required. Instead of a linear function, this cost function is called the 'Sigmoid function' or 'logistic function.' As demonstrated in Equation, the algorithm's hypothesis swings toward the cost function's limit between 0 and 1. This Logistic regression covers the case of a binary dependent variable--that is, where it can take only two values, "0" and "1", which represent outcomes such as Yes/No ,True/False ,High/Low etc.

$$0<h(x)<1$$

2.XGboost algorithm

XG Boost stands for eXtreme Gradient Boosting. It's a gradient boosted decision tree application that's designed for speed and efficiency. Boosting is an ensemble learning strategy that incorporates additional techniques to correct faults in previously presented models. Models are added in order until no further improvement is possible. To reduce the loss when adding new models, it employs a gradient descent method. This approach is used to give fast computing time and memory. The goal of this approach was to get the most out of the available resources to train the model. The two main reasons to work with XG Boost are execution speed and model performance.

3.CNN-LSTM algorithm

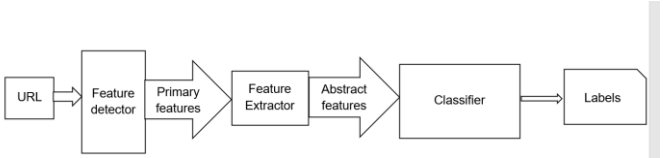
CNN and LSTM integration is a typical notion for integrating benefits due to the accessibility of CNN and LSTM. The notion for a novel deep learning scheme was proposed in this work by integrating CNN and LSTM. Two layers of CNN were utilized to ensure that the multidimensional data was properly correlated and captured. CNN layer feature series were used as an input for the LSTM algorithm. Time dependencies were extracted further in the layer LSTM. FC1, FC2, and FC3 were all connected in the architecture. The characteristics retrieved from the CNN layer are obtained using FC1 and FC2, and the final prediction of results is performed using FC3. The information on the phishing website cannot be adequately represented by the URL input matrix. In this section, the CNNLSTM URL, a web page code, a text function, and a rapid grading result are integrated to generate multidimensional features that explain the overall flow in detail. In order to perplex users, Phishers usually generate phishing URLs by mimicking the URL of your website. For example, a phishing URL appears to contain a PayPal imitator in its subsidiary domain name, which is a mess. PayPal rips off PayPal.

4.CNN BI-LSTM algorithm

A Recurrent Neural Network is a type of Bidirectional Long Short-Term Memory. It consists of two hidden layers that process data in both forward and backward directions, allowing the structure to retain knowledge of past input. It is the second layer in our suggested architecture, and it is used to remember prior transactions that are helpful in predicting the output y, which may be expressed as follows.

$$y^t = g(w_y [h^t, c^t] + b_y)$$

where t = transaction, w is the weight value assigned to the concatenation of the hidden and current state generated by the Bi-LSTM , h and c are the hidden and current state.



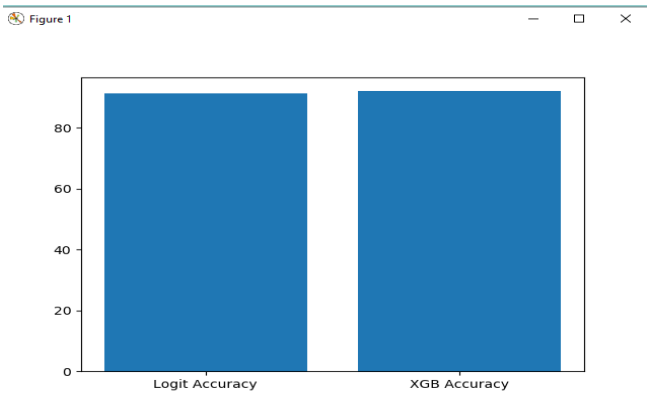
5.DATASET

Dataset used to train the model is taken from Kaggle.com. This consists of 50 attributes and more than 25 thousand data entries. Among them 80 percent is considered as training data and 20 percent as test data.

IV. RESULT

For the model, dataset is splitted in the ratio of 80:20 for the training and testing. The summary of the Machine learning models applied is shown in fig

Algorithms	Accuracy
CNN LSTM	57.85%
CNN BI-LSTM	56.36%
Logistic regression	91.89%
XGBoost	92%



V. DISCUSSION

SL.NO	Algorithm	Accuracy	Limitations
1	Online-toolbar algorithm	96%	Limited to Banking URL's
2	Goldphish	90%	Less secure

	algorithm		
3	C4.5 data mining algorithm	82.6%	Time consuming
4	Data mining algorithm using python and Django	-	Short term phishing websites
5	matching algorithm using Blacklist & Whitelist Approach	96.23%	detection of some minimal false positive and false negative results
6	Random forest Algorithm	96%	Consider only few features

The previous works of this project detects the phishing websites using data mining techniques, toolbars etc. The features considered are more, and the accuracy with Machine learning algorithms is high compared to the previous approaches. The number of iterations in gradient decision trees of xgboost algorithm and the use of sigmoid function and categorical dependency variables of logistic regression are the reason for the best model with high accuracy.

VI. CONCLUSION

In this paper, the issue of phishing attacks are considered and thus proposed a constructive model using CNN LSTM , CNN-Bi-LSTM, Logistic regression and XGBoost algorithms which combined machine learning mechanism and deep neural networks in data science to detect and classify the illegitimate URL's. Compared with the most extensively used LSTM model, the Logistic regression and XGBoost algorithm model achieve a good accuracy in detecting the phishing URL's. Analysis results show the adequacy of the model, and results into 92% accuracy. We can further develop this application as a website so that it can be accessed by anyone.

VI. REFERENCES

[1] AO Kaspersky lab. (2017). The Dangers of Phishing: Help employees avoid the lure of cybercrime. [Online] Available: https://go.kaspersky.com/Dangers-Phishing-Landing-Page- Soc.html [Oct 30, 2017].

[2] A Machine Learning Approach for Detection of Phished Websites Using Neural Networks by Charmi J. Chandan, Hiral P. Chheda, Disha M. Gosar, Hetal R. Shah.

[3] Phishing Detection System Using Machine Learning and Hadoop-MapReduce Kaustubh A. Hiwarekar, Dr. R. C. Thool.

[4] Y. Zhang, J. I. Hong, and L. F. Cranor. Cantina: a content-based approach to detecting phishing web sites. In WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 639– 648, New York, NY, USA, 2007. ACM.

[5] Haijun Zhang, Gang Liu, Tommy W. S. Chow, and Wenyin Liu, "Textual and Visual Content-Based Anti-Phishing A Bayesian Approach", IEEE 2011

[6] J. James, L. Sandhya and C. Thomas, "Detection of phishing URLs using machine learning techniques," in 2013 International Conference on Control Communication and Computing (ICCC), 2013.

and Mobile Communication Conference (IEMCON), 7th Annual. IEEE, 2016.

[7] Jain, Ankit Kumar, and B. B. Gupta." Comparative analysis of features-based machine learning approaches for phishing detection." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on.IEEE, 2016,

[8] Hawanna, Varsharani Ramdas, V. Y. Kulkarni, and R. A. Rane." A novel algorithm to detect phishing URLs." Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on. IEEE, 2016.