IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

# DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING

## Done By

**D.NEERAJA,V.RAMANA MURTHY,D.ROHIT SIVA REDDY,B.RAKESH,V.SWATHIKA,B.UDAY KIRAN**

Students of Visakha Institute of Engineering & Technology
Computer Science Engineering
Visakhapatnam,Andhra Pradesh

Under Guidance of

K RUKMINI DURGA
Visakha Institute of Engineering & Technology
Faculty of Computer Science Engineering
Visakhapatnam,Andhra Pradesh

Under Guidance of

A.S.C.TEJASWINI KONA
Visakha Institute of Engineering & Technology
HOD of Computer Science Engineering
Visakhapatanam,Andhra pradesh

*Abstract*— **Criminals seeking sensitive information construct illegal clones of actual websites and e-mail accounts. The e-mail will be made up of real firm logos and slogans. When a user clicks on a link provided by these hackers, the hackers gain access to all of the user's private information, including bank account information, personal login passwords, and images. Random Forest and Decision Tree algorithms are heavily employed in present systems, and their accuracy has to be enhanced. The existing models have low latency. Existing systems do not have a specific user interface. In the current system, different algorithms are not compared. Consumers are led to a faked website that appears to be from the authentic company when the e-mails or the links provided are opened. The models are used to detect phishing Websites based on URL significance features, as well as to find and implement the optimal machine learning model. Logistic Regression, Multinomial Naive Bayes, and XG Boost are the machine learning methods that are compared. The Logistic Regression algorithm outperforms the other two.**

***Keywords-Logistic Regression,Multinomial Naïve Bayes,XG Boost.***

## I. INTRODUCTION

Consumers have lost billions of dollars each year as a result of phishing operations. Refers to thieves' tricks for obtaining private information from a group of unwitting Internet users. Fraudsters obtain personal and financial account information such as usernames and passwords using fake email and phishing software to steal sensitive information. This research examines strategies for detecting phishing Web sites using machine learning techniques to analyze various aspects of benign and phishing URLs. It investigates how linguistic cues, host features, and page significance attributes are used to identify phishing site.The fine-tuned parameters aid in the selection of the most appropriate machine learning method for distinguishing between phishing and benign sites. Criminals that seek to steal sensitive information first establish illegal duplicates of legitimate websites and e-mail accounts, frequently from financial institutions or other companies that deal with financial data. The e-mail will be made up of real firm logos and slogans. One of the reasons for the rapid growth of the internet as a means of communication is that it allows themisuse of trademarks, brand names, and other corporate identities that consumers rely on as verification processes. "Spoof" e-mails are sent to many people in order make them involved in the criminal deception. Consumers are paid on a fraudulent website that appears to come from the real company when these emails are opened or when a link is clicked on the email.

## II. LITERATURE REVIEW

According to Erzhou Zhu (2018), phishers typically put up a false website where victims were tricked into providing passwords and perceptive information[1]. As a result, it's critical to detect rogue websites before they cause any harm to their victims. This study proposes a new method based on deep reinforcement to model and detects malicious URLs, fueled by the dynamic nature of criminal websites to steal sensitive information[2]. The suggested model may learn the properties related to phishing website identification by accommodating the dynamic behavior of phreaking websites[3]. The use of various types of machine learning algorithms for the problem of classification, particularly security and virus detection, has piqued the research community's interest in recent years[4]. Deep learning algorithms have also opened a new chapter on pattern recognition and artificial intelligence with the growth of processing capacity[5]. As a result, these powerful learning algorithms may now be used to solve a wide range of categorization, decision, and automation challenges[6]. When a high number of characteristics are included in the computation, deep learning-based techniques are very effective. Because algorithms based on reinforcement learning may estimate solutions (i.e., action) based on stochastic transformations and the rewards of selecting that state action, the proposed method is robust and

flexible[7]. By examining the given URLs, this project proposes a deep reinforcement learning-based model to find phishing websites[8]. The model adapts to modify the URL structure on its own. The traditional classification challenge is exemplified by the problem of recognizing phishing websites[9]. To handle this categorization challenge, a reinforcement learning model based on deep neural networks is constructed[10]. The problem of an "agent" performing an action that is entrenched on "trial and error" through interactions with an uncharted "environment" that offers a response in the form of numericalwards" is defined as the adaptive learning paradigm[11]. Other deep learning-based 4 algorithms, such as LSTM, should be investigated for the challenge presented in this research[12]. This classifier can also be used to solve other binary classification problems, such as detecting Webspam and the presence of hostile bots in the network[13]. Because the classifier in the RL-based technique is more flexible, it can be used to address a variety of privacy and security concerns in wearable devices[14].

To find criminal websites and its objective, SeenaThomas(2017) recommended extracting features from URLs and webpage links[17]. The matrix element is made up of basic links to the webpage of a given URL, in addition to the basic URL properties provided, such as length, suspicious characters, and a number of dots. In addition, statistical features such as mean, average, and variance are retrieved from each column of the feature matrix[21]. The given URL, links, and content on its web page, such as title and text content, are also used to extract dictionary features[24]. In order to detect the crime of identity theft, a number of machine learning models were tested, with the Deep Forest model showing competitive performance[25]. In particular, to discover hacker targets, an adequate technique based on search operators through search engines was devised. This method is rapid, however, it does not identify newly constructed phishing URLs[26]. Heuristic-based algorithms extract textual features, which can recognize newly created URLs, to detect phishing websites. Some textual components collected from the content, on the other hand, cannot be employed to determine phishing websites in those other languages[27]. According to some scholars, analogybased methods should be used to compare the similarities between the actually given web pages under attack,
i.e. the target of identity theft, which should be identified in advance. The criminal intentions of stealing sensitive information can now be seen automatically[28]. However, this approach is slow because it requires finding and analyzing a large number of web pages in order to improve the parasite community[29]. Using Deep Forest, as well as a range of contemporary machine learning models, such as GBDT and XG Boost, are used to represent URLs in vector form may be applied to detect sensitive identity theft. [30]. The proposed method works with websites written in a variety of languages. The obtained features can be employed by a variety of classification methods, with DF outperforming the competition[31].

## III. PROPOSED METHODOLOGY

The model is preprocessed in the proposed system, the words are tokenized, and stemming is performed. Data Processing is the process of converting or encoding data for easy machine transfer. In other words, the algorithm can now easily define data features. We must vectorize our URLs now that we have the data. Because some words in URLs are more essential than others, such as "virus," ".exe," and so on, the model employs Count Vectorizer and tokenizer to aggregate words. Let's make a vector representation of the URLs. A tokenizer that separates a string using a regular expression that matches either the tokens or the separators between tokens is known as a regular expression tokenizer. A regex pattern is a particular language for representing general text, numbers, or symbols in order to extract texts that match the pattern. 's+' is a simple example.. The method will sync at least one or more gaps if you add a '+' at the end.. In the world, stemming is crucial. Queries and Internet search engines both use stemming. The Fast Api framework is used for deployment. Fast API is a web framework for constructing APIs withPython 3.6+ and standard Python type hints that is current and fast (high-performance). The following are the main characteristics: Fast: Extremely fast, comparable to NodeJS and Go (thanks to Starlette and Pydantic). One of the quickest Python frameworks on the market. The UI is provided using FastAPI by loading the machine learning model into it. The architectural flow is shown in fig.1..

### *A. Advantages of proposed system*

- User Interface is provided
- Model is trained using many features
- High level of accuracy

### *B. Logistic regression*

A statistical strategy for predicting binary classes is logistic regression. The outcome or target variable is a binary variable. The term dichotomous refers to the fact that there are only two potential classes. It can, for example, be utilized to solve cancer detection issues. It calculates the likelihood of an event occurring.

### *C. XG Boost*

Extreme Gradient Boosting is abbreviated as XG Boost. The word XG Boost, on the other hand, refers to the engineering goal of pushing the computational resources for boosted tree algorithms to their limits. XG Boost is a software library that may be downloaded and installed on a computer and then accessed through a variety of interfaces.

### *D. Multinomial NB (MNB)*

The Multinomial NB (MNB) in Natural Language Processing, an algorithmic is a possible learning method (NLP). Using the Bayes theorem, the software estimates a text's tag, such as an email or a news report.

### *E. Testing*

System testing is based on the logical assumption that, if all components of the system are correct, system testing will be useful as a user-oriented vehicle prior to deployment. System testing finds faults, provides a recommendation to the administrator and alters the alteration, as well as checks the output's reliability. Before going live, the system is checked to see if the necessary software and hardware are in place to complete the project. To guarantee that this project is correct, it has passed the following

testing methods.

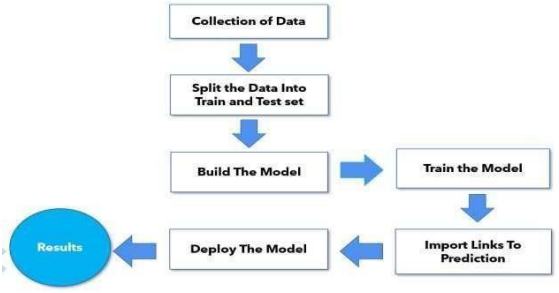


Figure 1. Graphical workflow of proposed models for detection phishing of websites

## IV. Results And Discussion

The current system merely detects phishing websites using multiple machine learning techniques and calculates their accuracy. The best model for detecting phishing websites is generated in the suggested system, and the model is saved and deployed, which takes the URL and predicts whether it is a criminal identity theft website or a real website. When compared to the old approach, the aforementioned statements show that this delivers better accuracy in detecting phishing websites. The accuracy of Logistic Regression is 96.63 percent, and the overall comparison is presented. The overall comparison is given in fig.2.

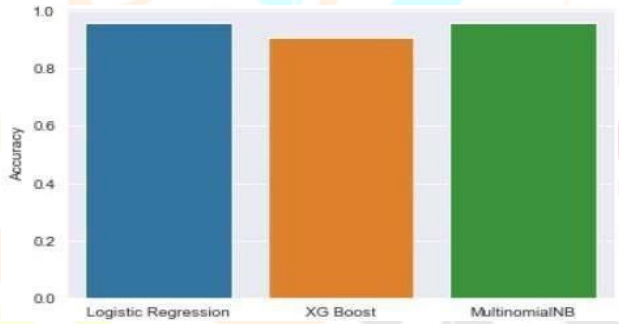


Figure 2. Comparison of the accuracy

## V. Conclusion

It is remarkable that a good anti-phishing system should be able to predict phishing attacks in a reasonable amount of time. Accepting that having a good anti-phishing gadget available at a reasonable time is also necessary for expanding the scope of phishing site detection. The current system merely detects phishing websites using multiple machine learning techniques and calculates their accuracy.

## References

[1] Chengshan Zhang, Steve Sheng, Brad Wardman, Gary Warner, Lorrie Faith Cranor, Jason Hong. Phishing Blacklists: An Empirical Study In: CEAS 2009: Proceedings of the 6th Conference on Email and Anti-Spam, Mountain View, California, USA, July 16-17, 2009.

[2] Andrew Jones, Mahmoud Khonji, Youssef Iraqi, Senior Member A Literature Review on Phishing Detection 2091-2121 in IEEE Communications Surveys and Tutorials, vol. 15, no. 4, 2013. 2013.

[3] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Many Understanding and Assisting Users' Online Choices with Nudges for Privacy and Security 50(3), Article No. 44, ACM Computing Surveys, 2017.

[4] Helena Matute, Mara M. Moreno-Fernández, Fernando Blanco, Pablo Garaizar I'm looking for phishers. To combat electronic fraud, Internet users' sensitivity to visual deception indicators should be improved. pp.421- 436 in Computers in Human Behavior, Vol.69, 2017.

[5] F.J. Overink, M. Junger, L. Montoya. Preventing social engineering assaults with priming and warnings does not work. pp.75-87 in Computers in Human Behavior, Vol.66, 2017. 2017.

[6] M. El-Alfy, El-Sayed M. Probabilistic Neural Networks and K-Medoids Clustering are used to detect phishing websites. The Computer Journal, 60(12), pp.1745-1759, published in 2017.

[7] Shuang Hao, Luca Invernizzi, Yong Fang, Christopher Kruegel, Giovanni Vigna. Cheng Huang, Shuang Hao, Luca Invernizzi, Yong Fang, Christopher Kruegel, Giovanni Vigna. Gossip: Detecting Malicious Domains from Mailing List Discussions Automatically pp. 494-505 in Proceedings of the 2017 ACM

[8] Asia Conference on Computer and Communications Security (ASIA CCS 2017), Abu Dhabi, United Arab Emirates, April 2-6, 2017.

[9] Gonzalo Nápoles, Rafael Falcon, Koen Vanhoof, Mario Köppen. Frank Vanhoenshoven, Gonzalo Nápoles, Rafael Falcon, Koen Vanhoof, Mario Köppen. Machine learning algorithms are used to detect dangerous URLs. The 2016 IEEE Symposium Series on Computational Intelligence (SSCI 2016) was held on December 6-9, 2016.

[10] Hillary Sanders, Joshua Saxe, Richard Harang, Cody Wild A Deep Learning Approach to Detecting Malicious Web Content in a Fast, Format- Independent Way. pp. 8-14 in Proceedings of the 2018 IEEE Symposium on Security and Privacy Workshops (SPW 2018), San Francisco, CA, USA, August 2.

[11] Jie Wu, Longfei Wu, Xiaojiang Du Phishing Attacks on Mobile Computing Platforms: Effective Defense Schemes 6678-6691 in IEEE Transactions on Vehicular Technology, vol. 65, no. 8, 2016.

[12] Ilango Krishnamurthi, R. Gowtham A system for detecting phishing websites that is both thorough and effective. pp. 23-37 in Computers & Security, Vol. 40, 2014.

[13] Lorrie Cranor, Guang Xiang, Jason I. Hong, Carolyn Penstein Rosé CANTINA+: A Phishing Web Site Detection Framework with a Feature-Rich Machine Learning Framework. Article No. 21 in ACM Transactions on Information and System Security, 14(2), 2011.

[14] Chengcheng Ye, Erzhou Zhu, Dong Liu, Feng Liu, Futian Wang, Xuejun Li An Effective Phishing Detection Model Using Neural Networks and Optimal Feature Selection In: Proceedings of the IEEE International Symposium on Parallel and Distributed Processing with Applications, 16th IEEE International Symposium on Parallel and Distributed Processing with Applications 781-787, Melbourne, Australia, December 11-13, 2018. (ISPA 2018).

[15] Systematization of Knowledge (SoK): ASystematic Review of Software- Based Web Phishing Detection, by Zuochao Dou, Issa Khalil, Abdallah Khreishah, Ala Al-Fuqaha, and Mohsen Guizani, IEEE Communications Surveys & Tutorials, 2017.

[16] "Detection and analysis of drive-by-download assaults and malicious javascriptcode," Proceedings of the 19th International Conference on World Wide Web, pp. 281-290, 2010. Marco Cova, Christopher Kruegel, Giovanni Vigna.

[17] "Phishing WebsiteDetection Using URL-Assisted Brand Name Weighting System," 2014 IEEE International Symposium on Intelligent Signal Processingand Communication Systems (ISPACS), December 1-4, 2014. Choon Lin Tan, Kang LengChiew, San Nah Sze.

[18] "Mining the web to detect phishing urls," Proceedings of the International Conference on Machine Learning and Applications, vol. 1, pp. 568-573, December 2012. R. B. Basnet and A. H. Sung.

[19] Mohiuddin Ahmed, AbdunNaser Mahmood, and Jiankun Hu, "A study of network anomaly detection strategies," J. Netw. Comput. Appl., vol.60, no. C, pp. 19-31, 2016.

[20] "A Novel Approach for Phishing DetectionUsing URL-Based Heuristic," LuongAnh Tuan Nguyen, Ba Lam To, HuuKhuong Nguyen1, and Minh Hoang Nguyen, IEEE 2014 International Conference on Computing, Management, and Telecommunications (ComManTel).

[21] "Intelligent phishing urldetection via association rule mining," Human- centric Computingand Information Sciences, vol. 6, no. 1, pp. [10]. "Feature selection for phishing detection: a review of studies," HibaZuhair, Ali Selamat, and MazleenaSalleh, International Journal of Intelligent Systems Technologies and Applications, Vol. 15, No. 2,201610, 2016.

[22] "An Online Backpropagation Algorithmwith Validation Error-Based Adaptive Learning Rate," in ArtificialNeural Networks – ICANN 2007, Porto, Portugal, 2007. S. Duffner and C. Garcia.

[23] "Predictingphishing websites based on self-structuring neural network," NeuralComputing and Applications, vol. 25, no. 2, pp. 443-458, 2013-B.
R. M. Mohammad, F. Thabtah, and L. McCluskey.

[24] Countermeasure techniques for deceptive phishing attack, International Conference on New Trends in Information and Service Science (NISS09), 30 June–02 July, 2009, China, pp.636–641.

[25] A. Mayuri and M. Tech (2012), "Phishing detection based on visualsimilarity," International Journal of Scientific and EngineeringResearch (IJSER), Vol. 3, No. 3, March.

[26] Chandrasekaran, Madhusudhanan, Krishnan Narayanan, and Shambhu Upadhyaya are the authors of this book. "Detection of phishing emails based on structural features." The 2006 NYSCyber Security Conference was held in New York City.

[27] "Discoveringphishing target based on semantic linknetwork," by Wenyin Liu et al. Future Generation Computer Systems, vol. 26.3, no. 3, pp. 381- 388, 2010.

[28] Wenyin Liu et al., "Discoveringphishing target via semantic link networks," Future Generation Computer Systems, vol. 26.3, no. 3, pp. 381- 388, 2010.

8]. "Intelligent phishingwebsite identification and prevention system by employing link guardalgorithm," by M. Madhuri, K. Yeseswini, and U. VidyaSagar. 9-15 in Int. J. Commun. Netw.Secure (2013).

[29] Greenstadt, Afroz, Sadia, and Rachel "Phishzoo: Detecting phishing websites by inspecting them." IEEE International Conference on Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. IEEE, 2011.

[30] "Discovering phishing target based on semantic linknetwork," Wenyin, Liu, and colleagues wrote. Computer Systems, vol. 26.3, no. 3, pp. 381- 388, 2010.

[31] "Improved Phishing Detection Using Model-Based Features," by A. Bergholz, J. H. Chang, G. Paass, F. Reichartz, and S. Strobel. 2008, CEAS.

**DOCUMENTATION LINK :**

https://drive.google.com/file/d/1HJZ0NLjhQZQIlnWgJ_7 cqGsNmw48Wcrx/view?usp=sharing

**CODE LINK:**

**https://drive.google.com/drive/folders/1XuQSPSYS5V SSMwgVedJL3yJbXHGTGm5q?usp=sharing**