RESEARCH ARTICLE OPEN ACCESS

# URL Hoaxing Detection using Machine Learning

**Ms. Apoorva Joshi, Ms. Apoorva Joshi, Manvi Bhardwaj**

Department of Computer Science and Engineering Global Institute of Technology – Jaipur

**ABSTRACT**

This research is all about detecting Phishing website. This project plays a cardinal role in detection of a phishing website and keeping user safe from fraudsters. Phishing is the most dangerous cyber-crime of the world with about 241,324 reported cases in 2020 and 96% of these attacks arrive by emails in form of some masked URL [1]. These attacks result in annual losses amounting to billions. This paper explores a Machine Learning model based on the Random Forest algorithm. It aims to predict whether a website is engaging in phishing activities or not by analyzing its URL and HTML content.

*Keywords— Random Forest Algorithm, Phishing Detection Using URL.*

## I. INTRODUCTION

Phishing is the most common form of cyber-attack all around the world. These attacks can affect a person or an organization by damaging their reputation, or by stealing their data, or direct monetary losses. These attacks are gaining momentum as they are easy to setup. It can be as easy as creating and hosting a fake website and luring users to give away their credentials. The best form of protection against these kinds of attacks is awareness and knowledge of such crimes because these kinds of attacks are very tempting. Attacks are evolving every second of time and attackers are changing their methods to lure users with different techniques. Phishing scams are done to gather data like credit card details, banking details, email passwords, or other personal sensitive data. All these attacks are successful when user clicks or visits the malicious website and enter their sensitive data.[2] There are some known differences between a phishing and a legitimate website and by looking for these differences these attacks can be prevented.

Phishing is carried out through various methods, including the following:

1. Email-to-Email: When an individual gathers an email containing sensitive information to forward it to the sender.
2. Website-to-Website: When someone is directed to a phishing website through a search engine or online advertisement.
3. Email-to-Website: When an individual receives an email containing a phishing link disguised as a legitimate website address.
4. Browser-to-Website: When someone mistypes a URL in the browser and is redirected to a phishing website with an address that closely resembles the legitimate URL..

All these attacks, despite being different from one another, have one thing in common, that is masked URL are used and other than that these URLs contain several suspicious flags like length of the URL, misspelled URL, too many special characters, too many sub domains etc. These URLs are very hard to distinguish from the regular ones. These URL leads to malicious and fake webpages which resembles an authentic website.

Most of the time there are some flags in these URLs for which user can look before clicking on them, but this task requires certain knowledge and can also be very tedious and time consuming. More importantly one can never be sure about the website by just looking for the flags in the URL. There is a more efficient method for distinguishing between Legitimate URL and Phishing URL and that is to look for the differences between the phishing webpage and legitimate webpage as they both are different from each other.[3] But, these methods are not easy to implement for a regular user.

In today's technically advanced world this problem can be addressed using Machine Learning. A Machine Learning algorithm called Random Forest is used to develop a model which takes URL of a website and predict whether that URL leads to a phishing website or a legitimate website. This model is trained using tokenized data (Byte Pair Encoding) of html pages of Legitimate and Phishing websites. When a URL is passed in the model, it grabs the HTML code and tokenize it. After tokenizing the html page, it compares it to the pre-learned data of both legitimate and phishing HTMLs and give the results accordingly.

The model is deployed in form of a website using flask where input URL will be taken from user and check whether the website is legitimate or not.

## II. METHODOLOGY

For training the model data has been collected from several phishing and legitimate websites HTML. Then collected data is tokenized using “Byte Pair Encoding” and derive a pattern using “TFIDF score” and memorize that pattern using “joblib”.

When a user enters a URL to check for its authenticity, the model grabs the html code of the following URL and tokenize it. Then “Random Forest” looks for a pattern in earlier memorized patterns and tokens of this HTML file and predict whether the URL is authentic or not.[4]

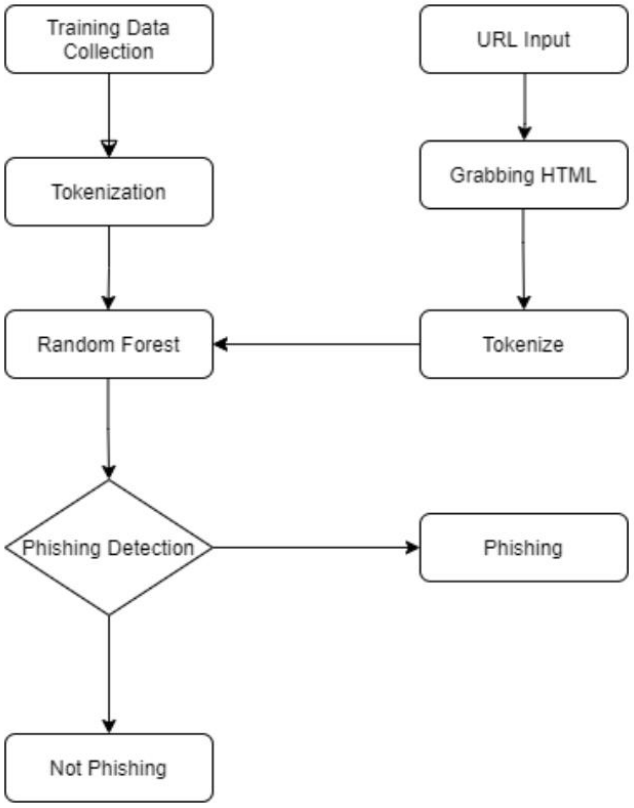



Fig. 1 Methodology

## III. RANDOM FOREST

Random forests, also called as random decision forests, are a type of ensemble learning technique utilized for classification, regression, and various other tasks. During training, they build numerous decision trees.

- In classification tasks, the random forest outputs the class chosen by the majority of trees.
- For regression tasks, it returns the mean or average prediction generated by the individual trees.

The Random Forest pseudocode:

1. Randomly select "k" features from a total of "m" features, where k is significantly less than m.
2. Among the "k" features, determine the optimal split point for the node "d".
3. Use the best split to divide the node into child nodes.
4. Repeat steps 1 to 3 until a certain number of nodes "i" has been reached.
5. Build the forest by repeating steps 1 to 4 "n" times to create "n" trees. In our implementation, classification is used to calculate entropy in the data. [5]

### A. *Formula for Variance / Mean Square Error*

$$\sum_{i=1}^{C} - f_i \log (f_i)$$

$f_i$ represents the frequency of label $i$ at a node.

$C$ is the number of unique labels [6]

## IV. RESULTS AND DISCUSSION

This model does a job of identifying authentic and phishing webpages. So, its efficiency is calculated on four parameters:

1. When a Legitimate website is provided and result is Legitimate.
2. When a Phishing website is provided and result is not Phishing.
3. When a Legitimate website is provided and result is Phishing.
4. When a Phishing website is provided and the result is Legitimate.

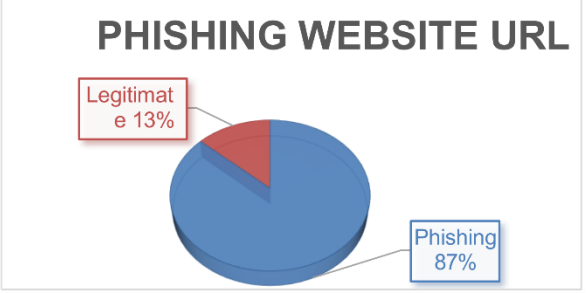
### A. Matrix

| Instance | Classifieds as Phishing | Classifieds as Legitimate |
|---|---|---|
| Phishing | 8776 | 48 |
| Legitimate | 1224 | 952 |
| Total | 10,000 | 1000 |

Table 1 : Confusion Matrix

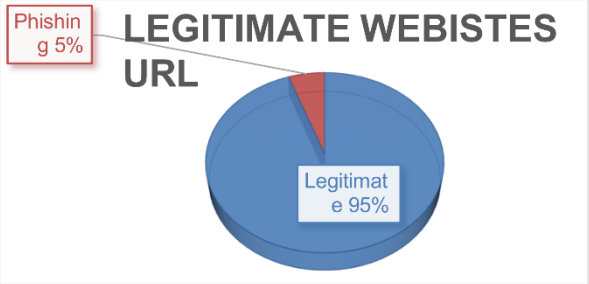
### B. Result when Phishing Websites URL are provided

The model is tested on over 10,000 phishing websites URL where the result was as follows:



C. Result when Legitimate Websites URL are provided

The model is tested on over 1000 legitimate website URL's where the result was as follows:




## V. CONCLUSION AND FUTURE SCOPE

It is always better to take some help of technology to provide accurate results. And from the results its concluded that the model has a successful chance of 87% in detecting a phishing website URL which is actually phishing. The model might tell a legitimate website as phishing but the possibility is as low as 5%. The purpose of the model is to solve a real-world problem that has grown enormously big in the last 2 decades. This research can be further extended in future for a commercially usable program or can be integrated with a browser like an extension. In near future with more training data the efficiency can be improved.

## REFERENCES

[1] M.Mathur, Rahul Jain, "Detection Of Fruit Diseases With Hybrid, Dwt-Glcm Approach", Eur. Chem. Bull. 2023, 12(Special Issue 7), 613-624.
[2] Tessian.| Phishing Statistics (Updated 2021) | 50+ Important Phishing Stats | Tessian. [online] | 2021
[3] Abu Saad Choudhary, Rucha Desai, Lavkush Gupta, Madhuri Gedam | Detection and Prevention of Phishing Attacks | 2021
[4] B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal | Fighting against phishing attacks: state of the art and future challenges | 2017
[5] Choon Lin Tan, Kang Leng Chiew, San Nah Sze | Phishing Webpage Detection Using Weighted URL Tokens for Identity Keywords Retrieval | 2017
[6] Medium | The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. [online] | 2021
[7] Dataaspirant. | How the random forest algorithm works in machine learning. [online] | 2021
[8] Rajesh Kr. Tejwani, Mohit Mishra, Amit Kumar. (2015). New Error Model of Entropy Encoding for Image Compression. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 1(3), 07–11.
[9] Rajesh Kr. Tejwani, Mohit Mishra, Amit Kumar. (2016). Evaluating the Performance of Similarity Measures in Effective Web Information Retrieval. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 2(8), 18–22.
[10] Amit Kumar, Mohit Mishra, Rajesh Kr. Tejwani. (2017). Image Contrast Enhancement with Brightness Preserving Using Feed Forward Network. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 3(9), 266–271.
[11] G.K. Soni, A. Rawat, S. Jain and S.K. Sharma, "A Pixel-Based Digital Medical Images Protection Using Genetic Algorithm with LSB Watermark Technique", Springer Smart Systems and IoT: Innovations in Computing. Smart Innovation Systems and Technologies, vol. 141, pp 483–492, 2020.
[12] Rajesh Kr. Tejwani, Mohit Mishra, Amit Kumar. (2018). Edge Computing in IoT: Vision and Challenges. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 4(8), 88–97.
[13] Mr. Gaurav Kuamr Soni, Mr. Kamlesh Gautam and Mr. Kshitiz Agarwal, "Flipped Voltage Follower Based Operational Transconductance Amplifier For High Frequency Application", International Journal of Advanced Science and Technology, vol. 29, no. 9s, pp. 8104-8111, 2020.
[14] Pradeep Jha, Deepak Dembla & Widhi Dubey , "Implementation of Transfer Learning Based Ensemble Model using Image Processing for Detection of Potato and Bell Pepper Leaf Diseases", International Journal of Intelligent Systems and Applications in Engineering, 12(8s), 69–80, 2024.
[15] Dr. Himanshu Arora, Gaurav Kumar soni, Deepti Arora, "Analysis and Performance Overview of RSA Algorithm", International Journal of Emerging Technology and Advanced Engineering, Vol. 8, Issue. 4, pp. 10-12, 2018.
[16] Pradeep Jha, Deepak Dembla & Widhi Dubey, "Deep learning models for enhancing potato leaf disease prediction: Implementation of transfer learning based stacking ensemble model", Multimedia Tools and Applications, Vol. 83, pp. 37839–37858, 2024.
[17] Vipin Singh, Manish Choubisa and Gaurav Kumar Soni, "Enhanced Image Steganography Technique for Hiding Multiple Images in an Image Using LSB Technique", TEST Engineering Management, vol. 83, pp. 30561-30565, May-June 2020.
[18] K. Gautam, S. K. Yadav, K. Kanhaiya and S. Sharma, "Hybrid Software Development Model Outcomes for In-House IT Team in the Manufacturing Industry" in International Journal of Information Technology Insights & Transformations (Eureka Journals), vol. 6, no. 1, pp. 1-10, May 2022.
[19] J. Dabass, K. Kanhaiya, M. Choubisa and K. Gautam, "Background Intelligence for Games: A Survey" in Global Journal on Innovation, Opportunities and Challenges in AAI and Machine Learning (Eureka Journals), vol. 6, no. 1, pp. 11-22, May 2022.
[20] P. Upadhyay, K. K. Sharma, R. Dwivedi and P. Jha, "A Statistical Machine Learning Approach to Optimize Workload in Cloud Data Centre," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 276-280, doi: 10.1109/ICCMC56507.2023.10083957.
[21] Pradeep Jha, Deepak Dembla & Widhi Dubey , "Crop Disease Detection and Classification Using Deep Learning-Based Classifier Algorithm", Emerging Trends in Expert Applications and Security. ICETEAS 2023. Lecture Notes in Networks and Systems, vol 682, pp. 227-237, 2023.
[22] Survey on Security Implication for the Downtime of VM in Cloud, Shekhawat, D., Ajmera, R.,Proceedings of the 2nd World Conference on Smart Trends in Systems, Security and Sustainability, WorldS4 2018, 2018, pp. 209–214, 8611575

[23] P. Jha, D. Dembla and W. Dubey, "Comparative Analysis of Crop Diseases Detection Using Machine Learning Algorithm," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 569-574, doi: 10.1109/ICAIS56108.2023.10073831.

[24] Gaurav Kumar Soni, Himanshu Arora and Bhavesh Jain, "A Novel Image Encryption Technique Using Arnold Transform and Asymmetric RSA Algorithm", Springer International Conference on Artificial Intelligence: Advances and Applications 2019 Algorithm for Intelligence System, pp. 83-90, 2020. https://doi.org/10.1007/978-981-15-1059-5_10

[25] P. Jha, R. Baranwal, Monika and N. K. Tiwari, "Protection of User's Data in IOT," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2022, pp. 1292-1297, doi: 10.1109/ICAIS53314.2022.9742970.

[26] P. Jha, T. Biswas, U. Sagar and K. Ahuja, "Prediction with ML paradigm in Healthcare System," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 1334-1342, doi: 10.1109/ICESC51422.2021.9532752.

[27] S. Pathak, K. Gautam, M. Regar and Dildar Khan, "A Survey on object recognition using deep learning," in International Journal of Engineering Research and Generic Science (IJERGS), vol. 7, no. 3, pp. 19-23, May-June 2021.

[28] S. Pathak, K. Gautam, A. K. Sharma and G. Kashyap, "A survey on artificial intelligence for Vehicle to everything," in International Journal of Engineering Research and Generic Science (IJERGS), vol. 7, no. 3, pp. 24-28, May-June 2021.

[29] Babita Jain, Gaurav Soni, Shruti Thapar, M Rao, "A Review on Routing Protocol of MANET with its Characteristics, Applications and Issues", International Journal of Early Childhood Special Education, Vol. 14, Issue. 5, pp. 2950-2956, 2022.

[30] K. Gautam, V. K. Jain and S. S. Verma, "A Survey on Neural Network for Vehicular Communication," in *Mody University International Journal of Computing and Engineering Research*, vol. 3, no. 2, 2019

[31] Gaur, P., Vashistha, S., Jha, P. (2023). Twitter Sentiment Analysis Using Naive Bayes-Based Machine Learning Technique. In: Shakya, S., Du, KL., Ntalianis, K. (eds) Sentiment Analysis and Deep Learning. Advances in Intelligent Systems and Computing, vol 1432. Springer, Singapore. https://doi.org/10.1007/978-981-19-5443-6_27

[32] P. Jha, D. Dembla and W. Dubey, "Implementation of Machine Learning Classification Algorithm Based on Ensemble Learning for Detection of Vegetable Crops Disease", International Journal of Advanced Computer Science and Applications, Vol. 15, No. 1, pp. 584-594, 2024.

[33] Unmasking Embedded Text: A Deep Dive into Scene Image Analysis, Maheshwari, A., Ajmera.R., Dharamdasani D.K., 2023 International Conference on Advances in Computation, Communication and Information Technology, ICAICCIT 2023, 2023, pp. 1403–1408

[34] Internet of Things (IoT) Applications, Tools and Security Techniques, Kawatra, R., Dharamdasani, D.K., Ajmera, R,et.al. 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022, 2022, pp. 1633–1639