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

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11318

Detection Of Phishing Websites Using Gradient Boosting Classifier Based On URL

D.Urlamma[1], M. Supriya[2], D. Lavanya[3], A. Hari Priya[4]

M.Tech, Computer science & Engineering, Bapatla women's Engineering College, Bapatla, INDIA[1]

B.Tech, Computer science & Engineering, Bapatla Women's Engineering College, Bapatla, INDIA[2-4]

Abstract: Phishing websites pose a severe risk to internet security because they try to obtain private information from gullible visitors. Researchers have created a number of methods, including machine learning algorithms, for identifying phishing websites in order to counter this problem. Large datasets of reputable and phishing websites can be used to train machine learning algorithms to find patterns and traits that differentiate the two. Subsequently, these algorithms can be employed to detect and prevent phishing websites from exploiting users. Feature extraction is one method of machine learning-based phishing website detection in which several aspects of a website, like URL structure, domain age, and content, are examined to detect phishing websites. These methods have the potential to be a valuable weapon in the fight against online phishing assaults with additional study and refinement.

Keywords: **:** SVM, Xgboost, Gradient boosting, Adaboost, Machine learning techniques

I. INTRODUCTION

The act of spotting and reporting phishing websites involves impersonating trustworthy websites in an effort to get sensitive data, including credit card details, login credentials, and personal identification numbers. Over time, phishing assaults have grown more complex, and attackers frequently employ strategies like social engineering and phony login screens to fool victims into divulging personal information. Phishing websites can also impersonate reputable websites by using URL spoofing. Identifying and blocking phishing websites is one of the best strategies to defend against phishing attempts. Phishing website detection is the process of identifying and reporting websites that are intended to trick consumers by using a variety of methods and tools. These phishing websites frequently use domain names that are similar to legitimate websites or are hosted on compromised servers. Using machine learning algorithms that can examine metadata, content, and other aspects of a website to find possible phishing sites is one method of identifying phishing websites. Large datasets of well-known phishing websites can be used to train these algorithms to find recurring themes and traits. Real-time data feeds may also be used by some machine learning models in order to recognize and detect newly generated phishing websites. Using reputation-based systems that keep lists of websites that are known to be dangerous is another method for detecting phishing websites. These systems can recognize and prevent phishing websites using a variety of information sources, including threat intelligence feeds, user reports, and blacklists.All things considered, identifying and blocking phishing websites is a crucial part of any successful cyber-security plan. As a result, it is essential to maintain vigilance and stay current with phishing threat trends and detection techniques.

II. LITERATURE SURVEY

1.J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology''.

A lot of fraudulent websites have appeared on the Internet in recent years with the intention of harming consumers by obtaining their personal data, including passwords, user names, and account IDs. Phishing is a type of social engineering assault that mostly targets mobile devices. One possible outcome of that would be financial losses. It provides a thoughtful analysis of the phishing problem, an up-to-date machine learning solution, and a future research agenda on machine learning-based phishing threats.

2.Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, "Phishing web sites features classification based on extreme learning machine," 6th Int. Symp. Digit. Forensic Secure. ISDFS - Proceeding.

A lot of fraudulent websites have appeared on the Internet in recent years with the intention of harming consumers by obtaining their personal data, including passwords, user names, and account IDs. Phishing is a type of social engineering assault that mostly targets mobile devices. One possible outcome of that would be financial losses. It Level-1 Heading: A level-1 heading must be in Small Caps, centered and numbered using uppercase Roman numerals.

provides a thoughtful analysis of the phishing problem, an up-to-date machine learning solution, and a future research agenda on machine learning-based phishing threats.

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

International Advanced Research Journal in Science, Engineering and Technology

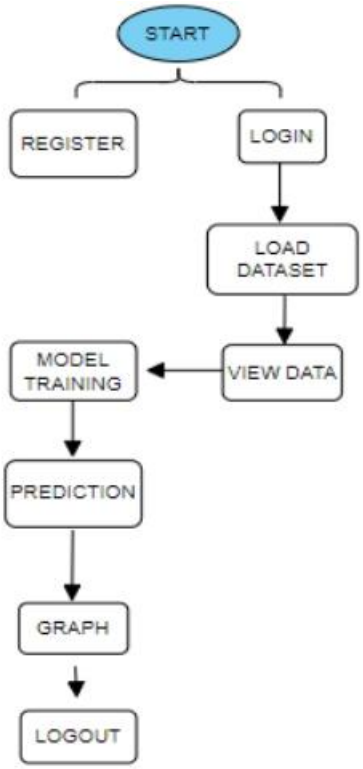
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11318

III. PROPOSED SYSTEM

The proposed system for phishing website detection using machine learning algorithms aims to overcome the limitations of existing systems. One approach is to use Gradient boosting classifier This can reduce the time and cost of data labelling and improve scalability. It is used for classification tasks and works by combining multiple weak classifiers into a strong classifier. This algorithm works by iteratively adding new decision trees to the model, where each new tree tries to correct the errors of the previous tree.

3.1 work Flow of Proposed system



3.2 DATA INFORMATION

Domain	Have_IP	Have_At	URL_Length	URL_Depth	Redirection	https_Domain
graphicriver.net	0	0	1	1	0	0
ecnavi.jp	0	0	1	1	1	0
hubpages.com	0	0	1	1	0	0
extratorrent.cc	0	0	1	3	0	0
icicibank.com	0	0	1	3	0	0
nypost.com	0	0	1	4	0	0
kienthuc.net.vn	0	0	1	2	0	0
thenextweb.com	0	0	1	6	0	0
tobogo.net	0	0	1	2	0	0
akhbarelyom.com	0	0	1	5	0	0
tunein.com	0	0	1	5	0	0

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

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11318

IV. METHODOLOGIES & ALGORITHMS

4.1 GRADIENT BOOSTING CLASSIFIER

As is well known, bias error and variance error are the two main categories into which errors in machine learning systems fall. Since gradient boosting is one of the boosting methods, it is employed to reduce the model's bias error.We are unable to mention the base estimator in the gradient boosting algorithm, in contrast to the Adaboosting approach. For the Gradient Boost approach, the base estimator is fixed and is called Decision Stump. We can adjust the gradient boosting algorithm's n_estimator, much like with AdaBoost. On the other hand, the default value of n_estimator for this algorithm is 100 if the value is left out. As a classifier or regression, the gradient boosting approach can be used to predict both continuous and categorical target variables.

Mean Square Error (MSE) is the cost function when it is used as a regressor, and log loss is the cost function when it is used as a classifier. With the aid of one example, let's now comprehend how the Gradient Boosting Algorithm functions. In the example below, Likes Exercising, GoToGym, and Drives Car are independent variables, and Age is the Target variable. Gradient Boosting Regressor is utilized here since the target variable is continuous, just like in this example. Now let's determine the estimator-2. As demonstrated below, the residues (age i – mu) of the first estimator are used as root nodes in the Gradient Boosting algorithm, in contrast to AdaBoost.

4.2 XGBOOST

The acronym for "Extreme Gradient Boosting" is XGBoost. A distributed gradient boosting library optimized for maximum efficiency, versatility, and portability is called XGBoost. It provides parallel tree boosting to address a range of data science tasks with speed and accuracy.

V. IMPLEMENTATION

5.1 Module

1: User

1.1View Home Page: The user is viewing the phishing website prediction web application's home page here.

1.2View Upload page: Users can find out additional information about the phishing prediction on the about page.

View Page: The user can view the dataset on the view page.

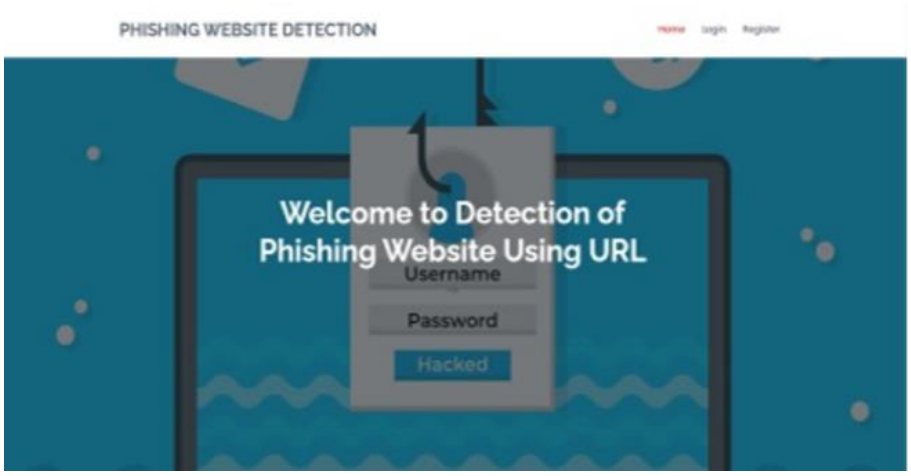
1.3 Input Model: In order to obtain results, the user must enter values for specific fields.

1.4View Results: The model's generated results are viewed by the user.

1.5View score: The user can see the score in percentage here.

VI. RESULT

Home Page:



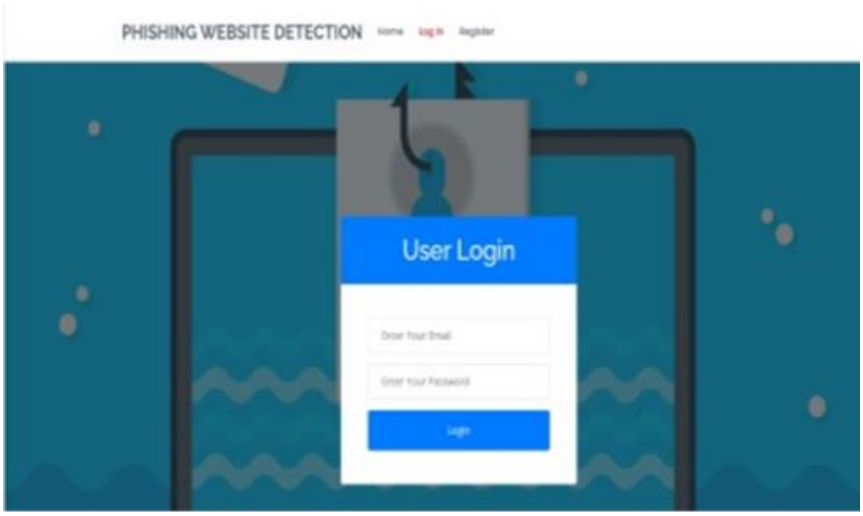
IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11318

Login Page:



Prediction Page:

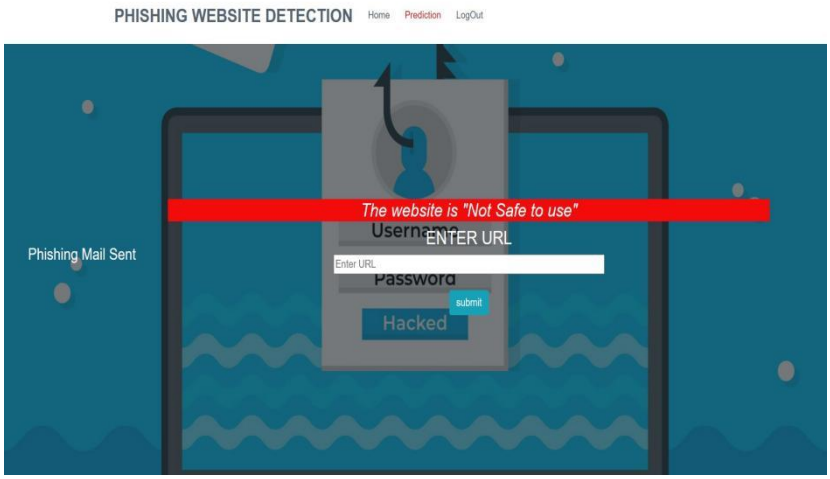
Result Page:

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

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11318



VII . CONCLUSION

Using machine learning to detect phishing websites is a promising way to counter the growing issue of online fraud. By training algorithms to recognize trends in the behavior and features of phishing websites, machine learning can detect and prevent dangerous websites before they cause harm. According to recent research, machine learning algorithms are highly accurate at identifying phishing websites. These algorithms can assess a website's likelihood of being a phishing site by looking at a number of characteristics, including the content, user interface, and URL structure.

VIII. FEATURE ENHANCEMENT

In the future, even with efficient AI and machine learning algorithms, the website won't be safe or secure. It will be better to show the explanation and the website name at the same time, as this would benefit the project's visitors. Furthermore, systems that combine machine learning and artificial intelligence may be able to forecast threats ahead of time and identify phishing schemes before they are implemented. Sharing of collaborative cloud-based threat intelligence can potentially be a frontier, allowing threat trends to spread more quickly.

REFERENCES

1.J. Shad and S. Sharma, “A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology,” pp. 425–430, 2018.

2.Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, “Phishing web sites features classification based on extreme learning machine,” 6th Int. Symp. Digit. Forensic Secure. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018

3.T. Peng, I. Harris, and Y. Sawa, "Using Natural Language Processing and Machine Learning to Detect Phishing Attacks," Proceedings of the 12th IEEE International Conference on Semantical Computing, ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

4."Performance comparison of classifiers on reduced phishing website dataset," by M. Karabatak and T. Mustafa, Proceedings of the 6th International Symposium on Digital Forensic Security, ISDFS 2018-Vol. 2018–Janua, pp. 1–5, 2018.

5. "A Novel Approach for Phishing Website Identification: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952. S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe.

6.The paper "Classification of URL bitstreams using bag of bytes" was presented by K. Shima and colleagues at the 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.

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

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11318

7. "Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks," Mazhayil, Vinayakumar, R., and Soman, K., 2018's 9th International Conference on Computing, Communication, and Networking Technologies, ICCCNT 2018, pp. 1–6.

8."On Feature Selection for the Prediction of Phishing Websites," IEEE 15th International Conference on Dependable, Auton. Secur., Comput., 15th International Conference on Pervasive Intelligence, Comput., 3rd International Conference on Big Data Intelligence, Comput., Cyber Sci. Technol. Congr., pp. 871–876, 2017.

9."Boosting the Phishing Detection Performance by Semantic Analysis," X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, 2017.

10.Phishing site detection using the C4.5 decision tree algorithm was presented by L. MacHado and J. Gadge at the 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5.

BIOGRAPHY

Mrs. D.Urlamma working as Assistant Professor in Department of CSE, Bapatla Women's Engineering College, Bapatla. She completed her B.tech in Computer Science Engineering from VNR College of Engineering Ponnur, and completed her M.tech in CSE from VNR College of Engineering, Ponnur.She has 6 years of Teaching experience in various Engineering Colleges.

M. Supriya B.Tech with Specialization of Computer Science and Engineering in Bapatla Women's Engineering College, Bapatla

D.Lavanya B.Tech with Specialization of Computer Science and Engineering in Bapatla Women's Engineering College, Bapatla

A.Hari Priya B.Tech with Specialization of Computer Science and Engineering in Bapatla Women's Engineering College, Bapatla