GREATER NOIDA INSTITUTE OF TECHNOLOGY

Synopsis on Project-
Anonymity and Confidentiality in Website using ML

Department of CSE(AI & ML)
Batch :- 2021-2025

under supervision of
Mr. Umashankar Sharma

Aditya Mishra (2101321530003)
Anshuman Soni (2101321530007)
Anuj Mishra (2101321530008)
Shivam Singh (2101321530045)

Greater Noida Institute of Technology
(ENGINEERING INSTITUTE)
Plot No-7, Knowledge Park-II, Greater Noida

INDEX

Table of contents

Description	Page No.
1. DECLERATION	3
2. ACKNOWLEDGEMENT	4
3. ABSTRACT	5
4. INTRODUCTION	6
5. LITERATURE SURVEY	7-9
6. OBJECTIVE	10
1. PROBLEM STATEMENT	10
2. KEY OBJECTIVES	10
7. METHODOLOGY	11-14
1. TECHONOLOGIES USED	11-12
2. ADVANTAGES & FEATURES OF PROPOSED TOOL	12-14
3. SYSTEM REQUIREMENTS	14
a. Software requirements	14
b. Hardware Requirements	14
8. CONCLUSION	15
9. REFERENCES	16

DECLERATION.

I hereby declare that this project work entitled “Anonymity and Confidentiality in Website using ML” has been prepared by me during the year 2024 – 25 under the guidance of Mr. Umashankar Sharma, Department of CSE (AI & ML), in the partial fulfillment of Btech degree prescribed by the college.

I also declare that this project is the outcome of my own effort, that it has not been submitted to any other university for the award of any degree.

Date: 20/09/2024

Aditya Mishra (2101321530003)
Anshuman Soni (2101321530007)
Anuj Mishra (2101321530008)
Shivam Singh (2101321530045)

ACKNOWLEDGEMENT

I have made efforts in this project. However, it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them. I am highly indebted to Mr. Umashankar Sharma for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. I would like to express my gratitude towards my parents & members of GNIOT for their kind cooperation and encouragement which helped me in the completion of this project. I would like to express my special gratitude and thanks to the industry people for giving me such attention and time. My thanks and appreciations also go to my colleague in developing the project and the people who have willingly helped me out with their abilities.

Date: 20/09/2024

Aditya Mishra (2101321530003)
Anshuman Soni (2101321530007)
Anuj Mishra (2101321530008)
Shivam Singh (2101321530045)

ABSTRACT

A phishing attack is the simplest way to obtain sensitive information from innocent users. The aim of the phishers is to acquire critical information like username, password, and bank account details. Cyber security people are now looking for trustworthy and steady detection techniques for phishing websites detection.

This project deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs.

Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. The aim of the paper is to detect phishing URLs as well as narrow down to the best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

INTRODUCTION

Phishing websites are carefully crafted online traps designed to steal sensitive information from unsuspecting users. They often mimic legitimate sites, luring victims to divulge login credentials, financial data, or other valuable personal details.

Nowadays Phishing has become a main area of concern for security researchers because it is not difficult to create a fake website which looks so close to a legitimate website. Experts can identify fake websites but not all the users can identify the fake website, and such users become the victim of phishing attack.

Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US $2 billion per year because their clients become victim to phishing. In the 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion. Phishing attacks are becoming successful because of lack of user awareness. Since phishing attacks exploit the weaknesses found in users, it is difficult to mitigate them, but it is important to enhance phishing detection techniques.

The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including fast flux, in which proxies are automatically generated to host the webpage; algorithmic generation of new URLs; etc.

A major drawback of this method is that it cannot detect zero-hour phishing attack. Heuristic based detection which includes characteristics that are found to exist in phishing attacks and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high.

To overcome the drawbacks of blacklist and heuristics-based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of many algorithms which require past data to decide or predict future data. Using this technique, algorithms will analyze various blacklisted and legitimate URLs and their features to accurately detect phishing websites including zero- hour phishing websites.

LITERATURE SURVEY

[1] Marwa Abd Al Hussein Qasim, Dr. Nahla Abbas Flayh: Using six criteria based on URL parameters such as the subdomain, principal domain, Page rank, Alexa rank, path domain, and Alexa reputation, this article suggests a novel method for identifying phishing websites. The method focuses on evaluating how closely a phishing site's URL resembles the URL of a reliable website and also takes into account the site's ranking as a crucial component in determining its validity. The approach was tested using data from PhishTank and DMOZ, and the authors showed that it could identify over 97% of phishing sites.

[2] Irfan Siddavatam, Rishikesh Mahajan: Introduced the use of decision trees , SVM, Random Forest to classify phishing websites based on features like URL length, domain age, presence of special characters, and security certificate information. Their approach demonstrated improved detection rates compared to traditional methods but faced challenges in identifying phishing sites.

[3] S. Arvind Anwekar, V. Agrawal : In this study, the authors focused on extracting features from URLs, in addition to other features such as the age of the SSL certificate and the universal resource locator of the anchor, IFRAME, and website rank. They collected URLs of phishing websites from PhishTank and URLs of benign websites from the Alexa website. Using a combination of the random forest (RF), decision tree (DT), and support vector machine (SVM), contributed to improving the detection mechanism for phishing websites and achieved a high noticeable detection accuracy of 97.14%, with a low rate of false positives at 3.14%. The results also showed that the classifier's performance improves with more training data.

[4] N. Choudhary b, K. Jain, S. Jain : This study emphasizes the significance of only using attributes from the URL. Both the Kaggle and Phishtank websites make it easy to get the dataset used in this study. The researchers used a hybrid approach that com-bined Principal Component Analysis (PCA) with Support Vector Machine (SVM) and Random Forest algorithms to reduce the dataset's dimensionality while keeping all im-portant data, and it produced a higher accuracy rate of 96.8% compared to other tech-niques investigated.

[5] A. Lakshmanarao, P. Surya, M Bala Krishna: This thesis collected a dataset of phishing websites from the UCI repository and used various Machine learning techniques, including decision trees, AdaBoost, support vector machines (SVM), and random forests, to analyze selected features (such as web traffic, port, URL length, IP address, and URL_of_Anchor). The most effective model for detecting phishing web-sites was chosen, and two priority-based algorithms (PA1 and PA2) were proposed. The team utilized a new fusion classifier in conjunction with these algorithms and at-tained an accuracy rate of 97%. when compared to previous works in phishing website detection

.

[6] L. Tang, Q. Mahmoud: The proposed approach in the current study uses URLs collected from a variety of platforms, including Kaggle, Phish Storm, Phish Tank, and ISCX-UR, to identify

phishing websites. The researchers made a big contribution since they created a browser plug-in that can quickly recognize phishing risks and offer warn-ings. Various datasets and machine learning techniques were investigated, and the pro-posed RNN-GRU model outperformed SVM, Random Forest (RF), and Logistic Re-gression with a maximum accuracy rate of 99.18%. On the other hand, the suggested method is not always accurate in identifying if short links are phishing risks.

[7] A. Kulkarni & L. Brown: A machine learning system was created to categorize websites based on URLs from the University of California, Irvine Machine Learning Repository. Four classifiers were used: SVM, decision tree, Naive Bayesian, and neural network. The outcome of experiments utilizing the model developed with the support of a training set of data demonstrates that the classifiers were able to successfully differentiate authentic websites from fake ones with an accuracy rate of over 90%. Limitations include a small dataset and all features being discrete, which may not be suitable for some classifiers.

[8] Tyagi; J. Shad; S. Sharma; S. Gaur Gagandeep Kaur: The research taken into account focuses on the use of various machine learning algorithms to identify if a web-site is legitimate or a phishing site based on a URL. This study's most important con-tribution is the creation of the Generalized Linear Model (GLM), a brand-new model. This model combines the results of two various methods. With a 98.4% accuracy rate, the Random Forest and GLM combination produced the best results for detecting phishing websites.

[9] M. Karabatak and T. Mustafa: The objective of this research is to assess the effectiveness of classification algorithms on a condensed dataset of phishing websites obtained from the UCI Machine Learning Repository. The paper investigates how data mining and feature selection algorithms affect reduced datasets through experiments and analysis, finally selecting the methods that perform the best in terms of classifica-tion. According to the results, some classification strategies improve performance while others have the opposite impact. Ineffective classifiers for condensed phishing datasets included Lazy, BayesNet, SGD Multilayer Perceptron, PART, JRip, J48, RandomTree, and RandomForest. However, it was discovered that KStar, LMT, ID3, and R.F.Clas-sifier were efficient. Lazy produced the highest classification accuracy rate of 97.58% on the compressed 27-feature dataset, whereas KStar performed at its best on the same dataset.

[10] X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng: A phishing detection model that applies Bagging, AdaBoost, SMO, and Random Forest algorithms to learn and test phishing detection strategies is offered as a contribution to this work. The model is based on features from URLs and extracts multi-level statistical characteristics, seman-tic features of word embedding, and semantic features from Chinese web content. Legal URLs from DirectIndustry online instructions and phishing data from the Anti-Phishing Alliance of China (APAC) are included in the dataset used to test the algorithm. The study's findings suggest that a fusion model that primarily employed semantic data to identify phishing sites with high detection efficiency had the best performance,

leading to a new contribution with an F-measure of 0.99%. Keep in mind that this approach is specific to Chinese websites and is language-dependent.

[11] W. Fadheel, M. Abusharkh, and I. Abdel-Qader : The present study utilized datasets from the UCI machine learning repository, including Domain, HTML, Address Bar, and URLs, the main contribution was conducting a comparative analysis of the impact of feature selection on detecting phishing websites. The KMO test was applied in the study to evaluate the dataset using (LR) and (SVM) classification algorithms. The test was conducted based on a correlation matrix to analyze the performance. Re-sults showed that LR with the KMO test achieved an accuracy of 91.68%, while SVM with the KMO test yielded an accuracy of 93.59%.

[12] A. Ahmed and N. A. Abdullah: The research team developed a software program known as Phish Checker, which is designed to distinguish between legitimate and phishing websites. The proposed approach focuses on identifying phishing attacks by analyzing the URLs and domain names of suspected phishing websites to determine their authenticity. Data was collected from the Yahoo and PhishTank directories and the results indicate that PhishChecker has an accuracy rate of 96% for identifying phish-ing websites. However, it should be noted that this method is based on heuristics and its effectiveness is reliant on the availability of certain discriminative elements that aid in identifying the type of website. Additionally, the study only examines the validity of URLs in determining website authenticity.

[13] Ankit Kumar Jain & B. B. Gupta: The proposed strategy utilizes an Innovative methodology for defending counteract phishing attempts by incorporating a URL and DNS matching module with a white list of trusted websites that are automatically up-dated based on each user's browsing history. This method offers quick retrieval speeds, high rates of detection, and alerts users to not disclose personal information when at-tempting to access a website, not on the white list. Additionally, hyperlink properties are utilized to verify the validity of a website by retrieving hyperlinks from the source code and applying them to the phishing detection method. The performance of this strategy was evaluated using data from reputable sources such as Stuffgate, Alexa, and PhishTank and achieved an accuracy rate of 89.38 %.

[14] M. Aydin and N. Baykal: Throughout this experiment, phishing websites were detected using subset-based feature selection methods based on URL attributes. A da-taset comprising both legitimate and phishing URLs was obtained from Google and PhishTank, and multiple features were retrieved from URLs. The usefulness of two classification algorithms—Naive Bayes and Sequential Minimal Optimization (SMO)—for identifying phishing websites was investigated in this study. The results showed that SMO performed better than Naive Bayes for phishing detection, with an accuracy rate of 95.39%. The SMO algorithm also had another benefit in that it made use of more chosen features overall. The accuracy rate of the Naive Bayes method, in contrast, was 88.17% while using the same quantity of chosen features.

OBJECTIVE

1. PROBLEM STATEMENT

Phishing websites have become a significant threat in the digital landscape, leading to financial losses and identity theft for individuals and organizations. These fraudulent websites often mimic legitimate ones, tricking users into disclosing sensitive information such as passwords, credit card numbers, and personal details. Traditional methods for detecting phishing sites, such as blacklisting, are limited in scalability and fail to detect new or dynamically generated phishing URLs in real-time.

The goal of this project is to develop a machine learning-based system to detect phishing websites by analysing features such as URL structure, domain age, security certificates, and website content. By leveraging supervised learning techniques, the system should be able to classify websites as phishing or legitimate with high accuracy, based on a set of labelled data.

2. KEY OBJECTIVES

1. Collect and preprocess data, including features from both phishing and legitimate websites.

2. Apply feature engineering to identify key characteristics distinguishing phishing websites.

3. Train multiple machine learning models and compare their performance.

4. Develop an efficient and scalable solution that can identify phishing websites in real-time.

5. Evaluate the model's accuracy, precision, recall, and F1-score using appropriate test datasets.

The successful implementation of this project will result in a robust and efficient system capable of helping users and organizations safeguard against phishing attacks.

METHODOLOGY

1. Technologies Used

Detecting phishing websites using machine learning algorithms typically involves using supervised learning techniques. Some common machine learning algorithms and methods used for phishing website detection include:

1. Logistic Regression: This algorithm is commonly used for binary classification tasks like phishing detection, where the output is either phishing or legitimate.
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable.

2. Decision Trees: Decision trees can be effective for phishing detection as they can capture complex relationships between features. Ensemble methods like Random Forests and Gradient Boosting Machines (GBM) can also be used for improved performance.
A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes, and leaf nodes.

3. Support Vector Machines (SVM): SVMs are good for binary classification tasks and can be trained to distinguish between phishing and legitimate websites based on various features.
Support vector machine is another powerful algorithm in machine learning technology. In support vector machine algorithm, each data item is plotted as a point in n-dimensional space and support vector machine algorithm constructs separating line for classification of two classes, this separating line is well known as hyperplane. The Support vector machine seeks for the closest points called as support vectors and once it finds the closest point it draws a line connecting to them. Support vector machine then construct separating line which bisects and perpendicular to the connecting line. To classify data perfectly the margin should be maximum. Here the margin is a distance between hyperplane and support vectors. In real scenario it is not possible to separate complex and nonlinear data, to solve this problem support vector machine uses kernel trick which transforms lower dimensional space to higher dimensional space.

4. K-Nearest Neighbors (KNN): KNN is a simple and intuitive algorithm that can be used for phishing detection by considering the similarity between features of websites.
The k-nearest neighbors (KNN) algorithm is a machine learning algorithm that uses proximity to classify or predict a data point's grouping. It is a supervised learning algorithm that works by comparing a data point to a set of data it was trained on.

5. Random Forest: Random Forest (RF) is a machine learning algorithm that combines the results

of multiple decision trees to produce a single result.
Random forest algorithm is one of the most powerful algorithms in machine learning technology and it is based on the concept of decision tree algorithm. Random forest algorithm creates the forest with number of decision trees. High number of trees gives high detection accuracy. Creation of trees is based on bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct single tree.
Among randomly selected features, random forest algorithm will choose best splitter for the classification and like decision tree algorithm; Random Forest algorithm also uses Gini index and information gain methods to find the best splitter. This process will continue until random forest creates n number of trees. Each tree in forest predicts the target value and then algorithm will calculate the votes for each predicted target. Finally random forest algorithm considers high voted predicted target as a final prediction.

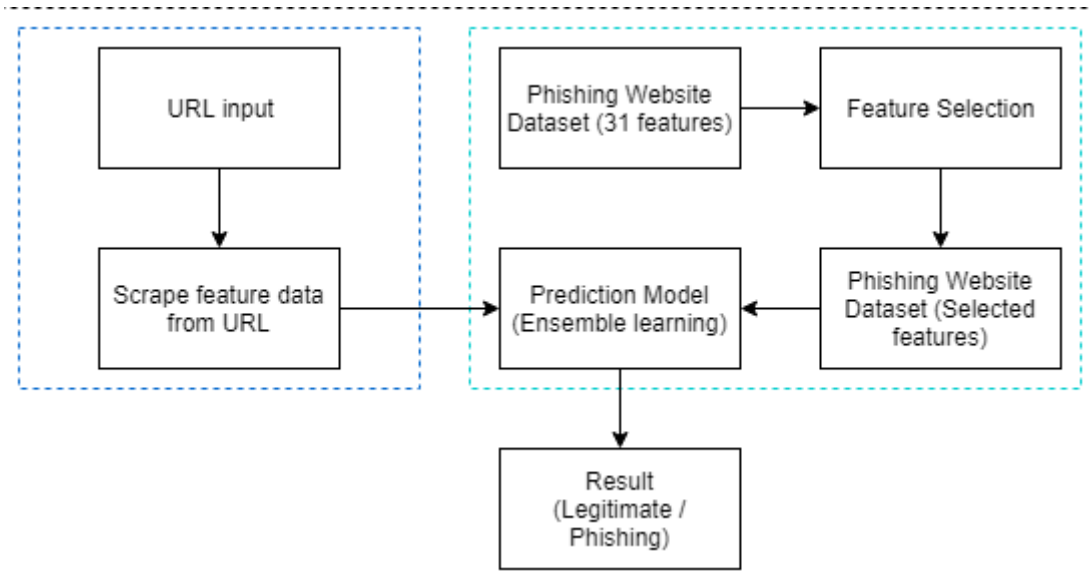

Fig. Flowchart of the Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning

2. Advantages & Features of Proposed Tool

Advantages:

1. **Real-time Detection:**

   The system can analyse websites in real-time, providing instant feedback on whether a website is legitimate or a phishing attempt.

2. **High Accuracy:**

Machine learning models can achieve high accuracy by learning from vast datasets of phishing and legitimate websites, minimizing false positives and negatives.

3. **Adaptive to New Threats:**

   Unlike traditional blacklist-based systems, the tool can detect newly created or dynamically generated phishing websites by recognizing patterns and features associated with phishing attempts.

4. **Scalability:**

   The tool can be easily scaled to analyse many websites simultaneously, making it suitable for both individual users and organizations.

5. **Low Maintenance:**

   Once trained, the system can operate with minimal human intervention, only requiring periodic retraining with new data to stay updated with the latest phishing trends.

6. **Platform Independence:**

   Can be integrated into various platforms such as browsers, antivirus software, or even mobile applications, providing flexible usage across different systems.

7. **Cost-Effective:**

   Reduces the need for expensive manual oversight and cybersecurity teams dedicated to phishing detection, lowering the overall cost for businesses.

8. **Reduced Human Error:**

   Automated detection significantly reduces human errors, which can occur during manual inspection of potentially suspicious websites.

Key Features:

1. **Feature Extraction:**

   Extracts critical attributes such as domain age, URL length, HTTPS status, IP address, and content features like the presence of suspicious scripts or forms.

2. **User-Friendly Interface:**

   The tool provides a simple and intuitive interface, allowing even non-technical users to check the legitimacy of websites with ease.

3. **Continuous Learning:**

   Incorporates feedback loops to improve detection accuracy over time by learning from newly identified phishing websites.

4. **Multi-Model Support:**

   Utilizes multiple machine learning algorithms (e.g., Decision Trees, Random Forest, Support Vector Machines) to ensure the best possible performance, selecting the optimal model based on the dataset.

These advantages and features make the proposed phishing detection tool robust, accurate, and efficient in protecting users and organizations from phishing attacks.

3. System Requirements

a. Software used :-

1. Installed Python.
   Libraries needed - pandas, numpy, urlparse, urlencode, BeautifulSoup, whois, urllib.request, time, socket, HTTPError, datetime, prange, Flask, Pickle

2. Installed IDE to use Python.
   Installed Visual Studio Code - Insiders

b. Hardware used :-

1. A Laptop/Desktop
   Connected to the Internet and has a browser.
   Processor: Preferably 1.0 GHz or Greater.
   RAM: 512 MB or Greater.

2. Single Network Connection
   So that other devices can connect through network URL.

CONCLUSION

This project aims to enhance detection methods to detect phishing websites using machine learning technology. We get very good performance in ensemble classifiers namely Random Forest, XGBoost both on computation duration and accuracy. The main idea behind ensemble algorithms is to combine several weak learners into a stronger one. This is the primary reason ensemble-based learning is used in practice for most of the classification problems.

It is worth mentioning that there is no guarantee that the combination of multiple classifiers will always perform better than the best individual classifier in the ensemble classifiers. The results motivate future works to add more features to the dataset, which could improve the performance of these models, hence it could combine machine learning models with other phishing detection techniques like example List-Base methods to obtain better performance.

REFERENCES

[1] Marwa Abd Al Hussein Qasim, Dr. Nahla Abbas Flayh, “Phishing Website Detection Using Machine Learning: A Review” June 2023Wasit Journal of Pure sciences 2(2):270-2812(2):270-281

[2] Rishikesh Mahajan, Irfan Siddavatam ,”Phishing Website Detection using Machine Learning Algorithms” , International Journal of Computer Applications, Volume 181 - Number 23 Year of Publication: 2018

[3] S. A. Anwekar and V. Agrawal, "PHISHING WEBSITE DETECTION USING MACHINE LEARNING ALGORITHMS."

[4] S. Jain, "Phishing Websites Detection Using Machine Learning," Available at SSRN 4121102.

[5] A. Lakshmanarao, P. S. P. Rao, and M. B. Krishna, "Phishing website detection using novel machine learning fusion approach," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021: IEEE, pp. 1164-1169.

[6] L. Tang and Q. H. Mahmoud, "A Deep Learning-Based Framework for Phishing Web-site Detection," IEEE Access, vol. 10, pp. 1509-1521, 2021.

[7] A. D. Kulkarni and L. L. Brown III, "Phishing websites detection using machine learn-ing," 2019.

[8] I. Tyagi, J. Shad, S. Sharma, S. Gaur, and G. Kaur, "A novel machine learning ap-proach to detect phishing websites," in 2018 5th International conference on signal pro-cessing and integrated networks (SPIN), 2018: IEEE, pp. 425-430.

[9] M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," in 2018 6th International Symposium on Digital Forensic and Security (ISDFS), 2018: IEEE, pp. 1-5.

[10] X. Zhang, Y. Zeng, X.-B. Jin, Z.-W. Yan, and G.-G. Geng, "Boosting the phishing detection performance by semantic analysis," in 2017 IEEE international conference on big data (big data), 2017: IEEE, pp. 1063-1070.

[11] W. Fadheel, M. Abusharkh, and I. Abdel-Qader, "On Feature selection for the prediction of phishing websites," in 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2017: IEEE, pp. 871-876.

[12] A. A. Ahmed and N. A. Abdullah, "Real time detection of phishing websites," in 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Con-ference (IEMCON), 2016: IEEE, pp. 1-6.

[13] A. K. Jain and B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," EURASIP Journal on Information Security, vol. 2016, no. 1, pp. 1-11, 2016.

[14] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," in 2015 IEEE Conference on Communications and Network Security (CNS), 2015: IEEE, pp. 769-770.