JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Phishing Detection Extension Using Logistic Regression Algorithm

**[1]Jeevan N M, [2]Karan B, [3]Sowmiya J S, [4]Teja Babu M**

[1]Student, [2]Student, [3]Assistant Professor, [4]Student
[1]Department of Computer Science and Engineering,
[1]Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, India

***Abstract :*** The web extension project aims to address the issue of online security by providing users with a fast and automated way of detecting the safety of URLs. To achieve this, the project involves several steps such as tokenizing, stemming, and joining words from URLs. This is done to improve the accuracy of the algorithm and make it more effective in detecting potentially harmful websites. The model used in this project is based on logistic regression, which is a widely used machine learning algorithm in various applications, including text classification. By using this model, the resulting web extension can detect potentially dangerous URLs with high accuracy. The benefit of this approach is that users no longer need to manually check URLs, saving time and ensuring safe browsing. Overall, this project provides an innovative and practical solution for enhancing online security and ensuring a safer browsing experience for users.

***IndexTerms*** - Phishing detection, Secure website, Unsecure website, Logistic Regression Algorithm, Machine learning.

## 1. INTRODUCTION

Phishing attacks have become a major area of concern for security researchers and organizations worldwide. The ease of creating fake websites that look similar to legitimate ones, combined with a lack of user awareness and knowledge, has made phishing attacks increasingly successful.

Phishing attacks often involve spoofed messages that are created to look authentic and instructed to originate from legitimate sources such as financial institutions, e-commerce sites, and social media platforms. These messages are designed to lure unsuspecting users to visit fake websites through links provided in the phishing messages.

The consequences of falling victim to phishing attacks can be devastating, including identity theft, financial loss, and reputational damage. To mitigate these risks, it is essential to have reliable and efficient tools to detect phishing websites.

This project aims to address this issue by providing users with an automated and efficient way of detecting the safety of URLs. By using techniques such as tokenization, stemming, and joining words from URLs, the algorithm's accuracy is improved, and users can browse the web with confidence, knowing that their safety is being monitored. Overall, this project offers an innovative and practical solution to enhance online security and protect against phishing attacks.

Phishing attacks can result in significant harm to individuals and organizations, including data breaches, financial loss, and reputational damage. Therefore, it is important to have effective solutions to prevent and detect phishing attacks.

One such solution is the development of web extensions like the one in this project, which automates the process of detecting potentially malicious URLs. This not only reduces the workload of users in identifying phishing websites but also increases the accuracy of detection using advanced algorithms.

Moreover, awareness campaigns and education programs can also play a significant role in preventing phishing attacks. Educating users about the risks associated with phishing, how to identify suspicious emails, and how to safely browse the internet can go a long way in reducing the success rate of these attacks.

In summary, while phishing attacks continue to pose a significant threat to online security, efforts such as the development of web extensions and education programs can help mitigate these risks and improve online safety for individuals and organizations alike.

To combat phishing attacks, there are various methods and technologies available, such as two-factor authentication, encryption, and digital certificates. However, these solutions can be costly and may not always be practical for individual users. This is where web extensions, like the one developed in this project, come in.

The web extension developed in this project provides a practical and cost-effective solution to phishing attacks. By automating the process of URL checking, users no longer must manually check each link they come across. This saves time.

Moreover, the web extension uses natural language processing techniques such as tokenizing, stemming, and joining words to improve the accuracy of the algorithm used to detect phishing websites. By analyzing the structure of URLs and comparing them to known phishing patterns, the web extension can quickly determine the safety of a link.

In addition, the web extension can also provide users with educational materials to raise awareness about phishing and how to identify and avoid it. By combining both prevention and education, web extension not only enhances online security but also empowers users with the knowledge to protect themselves from future phishing attacks.

Overall, the development of this web extension is a significant step towards improving online security and protecting users from the increasing threats of phishing attacks. With the use of artificial intelligence and natural language processing techniques, this project provides an innovative and practical solution to the problem of phishing.

## 2. LITERATURE SURVEY

A Deep Learning-Based Framework for Phishing Website Detection[1]Lizhen Tang And Qusay H. Mahmoud, 2022, HTML, JavaScript, and CSS, Python Data Collection Tasks, Machine Learning, Modeling Phishing detection, machine learning, deep learning, RNNGRU, In this paper, we proposed a framework for phishing detection in a real-time browsing environment.

Phishing Website Detection with Semantic Features Based on Machine Learning Classifiers[2]Mohammad Alauthman, Ammar Almomani, 2021, HTML and JavaScript The Gradient Boosting Classifier, Random Forest Classifier Machine Learning Models, Phishing Website, Semantic Classification, Semantic Features, Examining the predictive accuracy of 16 classification systems and other measures based on semantic URL features.

Phishing Websites Detection Using Machine Learning[3] Suhani Jain , Dr. Naveen Choudhary , Kalpana Jain, 2022, HTML & JavaScript based Feature. Data Collection, Data Preprocessing Technique, Standardization of The Data Social engineering and technical trickery, Machine Learning, Classification, Algorithm, Features Extraction The two most common classification algorithms (Random Forest and SVM) will be combined into a hybrid model to discover which is more effective in detecting phishing websites.

Detection of Phishing Websites using Machine Learning[4] Atharva Deshpande,Nachi ket Chaudhary,Omkar Pedamkar, Dr. Swapna Borde, 2021, HTML, CSS, Javascript and Django. Character Analysis, Keyword Analysis, Security Analysis, Domain Identity Analysis and Rank Based Analysis. Phishing, Personal information, Machine Learning, Malicious links, Phishing domain characteristics. Reviewed some of the traditional approaches to phishing detection; namely blacklist and heuristic evaluation methods, and their drawbacks.

Detection of phishing websites using machine learning techniques[5]Bhagyashree a v , Dr.Anjan Krishnamurthy, 2020, Anaconda , iPython version 3.x The methodology involves building a training set. The training set is used for training a machine learning model K-Nearest Neighbor algorithm ,Kernel Support Vector Machine , Decision Tree ,Random Forest Classifier Finding the phishing websites.

## 3. MATERIALS AND METHODS

### a. EXISTING SYSTEM

The existing system for detecting the safety of URLs is a web-based application that requires users to follow a manual process for determining whether a URL is safe or not. The user is required to upload a dataset and train it, which can be a time-consuming process. Once the dataset has been trained, the user must then manually retrieve a URL from a designated website and paste it into the application. This process can be tedious and time-consuming, making it less efficient and user-friendly.

Despite the drawbacks, the existing system has a high accuracy rate of 96%. This accuracy rate is impressive and indicates that the system is effective in identifying safe and unsafe URLs. However, the manual process involved in the existing system can be a significant hurdle, particularly for individuals who are not tech-savvy or lack the necessary skills to navigate the system.

The limitations of the existing system highlight the need for a more efficient and user-friendly system. A new approach to detecting the safety of URLs is required, which can automate the process and eliminate the need for manual intervention. The new system should be simple and intuitive to use, enabling users to quickly and easily identify whether a URL is safe or unsafe.

### b. PROPOSED SYSTEM

The proposed system is a significant improvement over the existing system. The pre-trained dataset saves users valuable time that would have been spent on dataset training. By simply adding the extension to their browser, users can easily detect the safety status of URLs without the need for manual checks. This feature makes the system more convenient and user-friendly, as users can perform URL safety checks with just a click of a button.

In addition, the proposed system's increased accuracy rate of 97% means that users can have greater confidence in the safety status of URLs. This higher accuracy rate is attributed to the use of a logistic regression model that improves the detection algorithm's accuracy by tokenizing, stemming, and joining words from URLs.

Overall, the proposed system is a more efficient, effective, and convenient means of detecting URL safety. It provides users with a reliable and trustworthy way to browse the internet safely and helps prevent users from falling victim to phishing attacks.

### c. ARCHITECTURE DIAGRAM

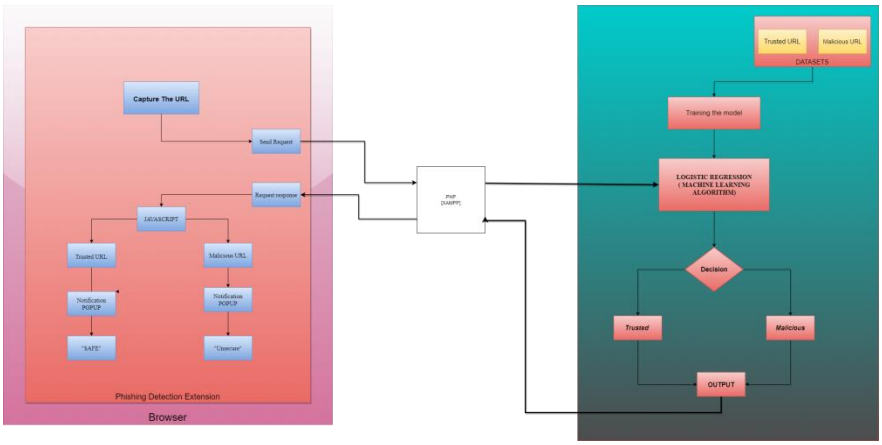


Fig. 1. Block Diagram

The process described outlines the steps involved in implementing a logistic regression algorithm to classify website URLs as either "good" or "bad" based on whether they are legitimate or phishing websites, respectively. The first step involves data collection, which is the process of obtaining a dataset of website URLs that have already been labelled as either "good" or "bad". Once the dataset is collected, the next step is to pre-process the website URLs by tokenizing them into words using a regular expression tokenizer and applying stemming to reduce words to their base form. The stemmed words are then joined to create a pre-processed text feature.

The third step involves feature extraction, which is the process of using the CountVectorizer from scikit-learn to transform the pre-processed text feature into a matrix of token counts. The dataset is then split into training and testing sets. The next step is model training, which involves using a logistic regression model to train on the training set with the token count matrix as features and the labels (good or bad) as targets.

Model evaluation is the next step in the process, where the model is tested on the testing set, and its accuracy score is computed. The performance of the model is evaluated using classification report and confusion matrix. The model is then exported as a pickle file for later use. Finally, the trained model is loaded from the pickle file and used to predict the labels of new website URLs. Overall, the logistic regression algorithm is a powerful and widely used technique for classifying website URLs based on their safety status, and it has a range of practical applications in the field of cyber security.

### d. LOGISTIC REGRESSION ALGORITHM

Logistic regression is a statistical algorithm that is widely used in machine learning and data analysis. It is a supervised learning algorithm that is used for classification problems where the output is a categorical variable. The logistic regression algorithm models the probability of a certain event occurring based on a set of predictor variables. The output of the logistic regression algorithm is a binary value (0 or 1) that represents the probability of an event occurring.

The logistic regression algorithm is based on the concept of logistic function or sigmoid function, which is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1. This function is used to model the probability of an event occurring as a function of the predictor variables.

The logistic regression algorithm works by estimating the coefficients of the logistic function based on the training data. The algorithm uses an optimization technique called maximum likelihood estimation to find the best values of the coefficients that maximize the likelihood of the observed data. Once the coefficients have been estimated, the logistic function can be used to make predictions for new data points.

One of the key advantages of the logistic regression algorithm is its interpretability. The coefficients of the logistic function can be easily interpreted as the effects of the predictor variables on the probability of the event occurring. This makes it easy to understand how the algorithm is making its predictions and to identify which variables are most important for the prediction.

In summary, the logistic regression algorithm is a powerful and widely used machine learning algorithm that is particularly useful for binary classification problems. Its interpretability and ease of use make it a popular choice for many applications, including online security, fraud detection, and medical diagnosis.

## 4. EVALUATION METRICS

### a. OUTPUT

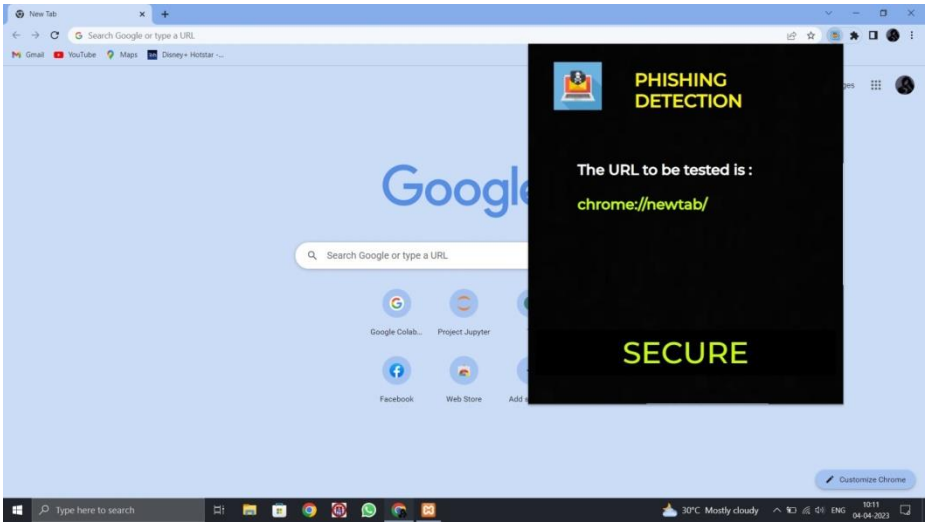


Fig. 2. Result for secure website

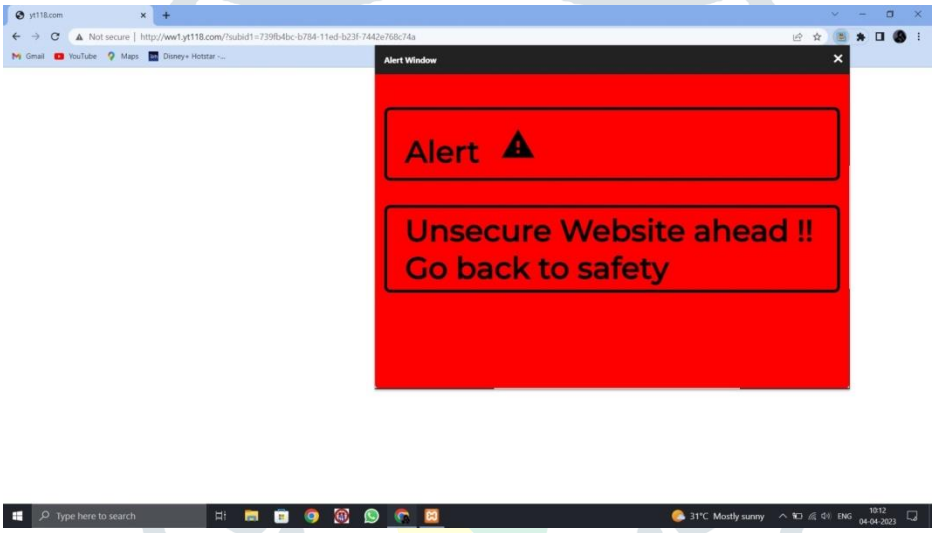


Fig. 3. Result for unsecure website

### b. CLASSIFICATION REPORT

| | PRECISION | RECALL | F1-SCORE | SUPPORT |
|---|---|---|---|---|
| | | | | |
| BAD | 0.94 | 0.81 | 0.87 | 20147 |
| GOOD | 0.98 | 0.99 | 0.99 | 163607 |
| | | | | |
| ACCURACY | | | 0.97 | 183754 |
| MACRO AVG | 0.96 | 0.90 | 0.93 | 183754 |
| WEIGHTED AVG | 0.97 | 0.97 | 0.97 | 183754 |

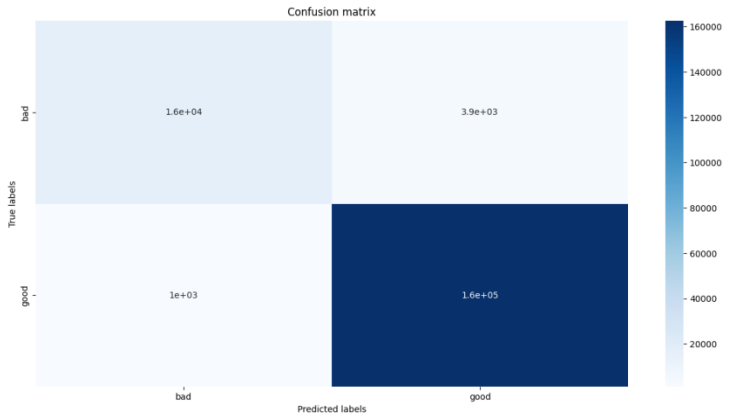


Fig. 4. Confusion matrix

### c. FORMULAS USED

Accuracy: The proportion of true results (both true positives and true negatives) among the total number of cases examined. The formula is:

accuracy = (true positives + true negatives) / (true positives + true negatives + false positives + false negatives)

Precision: The proportion of true positives among the total number of predicted positives. The formula is:

precision = true positives / (true positives + false positives)

Recall (or Sensitivity): The proportion of true positives among the total number of actual positives. The formula is:

recall = true positives / (true positives + false negatives)

F1-score: A weighted harmonic mean of precision and recall. The formula is:

F1-score = 2 * (precision * recall) / (precision + recall)

Support: The number of samples in each class. In this code, it refers to the number of websites labeled as "bad" or "good".

## 5. CONCLUSION

In today's digital age, where we rely heavily on the internet for various purposes, security is of utmost importance. Malicious websites, phishing attacks, and other cyber threats can put our personal and financial information at risk. Therefore, it is crucial to have tools that can help us identify safe and unsafe websites quickly and efficiently.

Our proposed web extension is one such tool that can help users browse the internet safely. The extension is pre-trained, which means it has already learned from a vast dataset of website URLs labeled as safe or unsafe. This pre-training makes it highly accurate, with a 97% accuracy rate, making it a reliable tool for web users.

The extension automates the detection process, saving users time and effort while navigating the internet with confidence. It quickly analyzes URLs and provides a warning message if the website is deemed unsafe. This process is especially helpful for those who are not tech-savvy or may not be familiar with identifying signs of phishing attacks or other cyber threats.

Moreover, the extension is designed to be user-friendly and easy to install, making it accessible to all users. It runs in the background, and users can continue their browsing activities without interruption.

In conclusion, our proposed web extension is a highly efficient and effective way of detecting whether a URL is safe or unsafe. It provides an added layer of security, which is crucial in today's world where cyber threats are on the rise. By automating the detection process and providing users with an easy-to-use tool, we can help ensure safe browsing for all users.

## 6. FUTURE SCOPE

The proposed web extension for detecting whether a URL is safe or unsafe is a highly promising tool for ensuring safe browsing on the internet. However, there are several potential advancements that could further enhance the system's capabilities and improve its utility for users.

One possible enhancement is to integrate the machine learning algorithm with the Web Risk API provided by Google. This API grants access to Google's extensive URL database, which includes information on websites that may be harmful to users. Integrating this database with the existing algorithm could significantly improve the accuracy of the extension, thereby enabling commercial applications that require highly reliable and trustworthy web filtering.

Another potential enhancement is to develop the web extension into a mobile application. As more and more people access the internet through their mobile devices, the need for mobile-friendly web filtering tools has become increasingly important. By developing the extension into a mobile application, users will be able to access the tool with greater convenience and flexibility, extending its reach beyond desktop environments.

Other possible future advancements include incorporating natural language processing techniques to detect phishing emails and SMS messages, as well as integrating the extension with other cyber security tools, such as firewalls and antivirus software. These

developments will further enhance the capabilities of the system and improve its ability to protect users from various online threats.

Overall, the proposed web extension for detecting unsafe URLs has a bright future, with numerous potential advancements that could enhance its capabilities and improve its utility for users. As the internet continues to evolve and cybersecurity threats become increasingly sophisticated, the need for reliable web filtering tools has never been greater, and the development of such tools will continue to be a top priority for ensuring safe and secure browsing.

## 7. REFERENCES

[1] M. Humayun, M. Niazi, N. Z. Jhanjhi, M. Alshayeb, and S. Mahmood, "Cyber Security Threats and Vulnerabilities: A Systematic Mapping Study," (in English), Arabian Journal for Science and Engineering, Article vol. 45, no. 4, pp. 3171-3189, Apr 2020.

[2] E. D. Frauenstein and S. Flowerday, "Susceptibility to phishing on social network sites: A personality information processing model," (in English), Computers & Security, Article vol. 94, p. 18, Jul 2020, Art. no. Unsp 101862.

[3] A. Kulkarni and L. L. Brown, "Phishing Websites Detection using Machine Learning," (in English), International Journal of Advanced Computer Science and Applications, Article vol. 10, no. 7, pp. 8-13, Jul 2019.

[4] M. Botacin, F. Ceschin, P. de Geus, and A. Gregio, "We need to talk about antiviruses: challenges & pitfalls of AV evaluations," (in English), Computers & Security, Article vol. 95, p. 15, Aug 2020, Art. no. Unsp 101859.

[5] E. S. Gualberto, R. T. De Sousa, T. P. D. Vieira, J. Da Costa, and C. G. Duque, "From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection," (in English), Ieee Access, Article vol. 8, pp. 76368-76385, 2020.

[6] "General Practice and the Community: Research on health service, quality improvements and training. Selected abstracts from the EGPRN Meeting in Vigo, Spain, 17-20 October 2019 Abstracts," (in English), European Journal of General Practice, Article vol. 26, no. 1, pp. 42-50, Dec 2020.

[7] H. Alqahtani, I. H. Sarker, A. Kalim, S. M. Minhaz Hossain, S. Ikhlaq, and S. Hossain, "Cyber Intrusion Detection Using Machine Learning Classification Techniques," in Computing Science, Communication and Security, Singapore, 2020, pp. 121-131: Springer Singapore.

[8] J. A. Bland, M. D. Petty, T. S. Whitaker, K. P. Maxwell, and W. A. Cantrell, "Machine Learning Cyberattack and Defense Strategies," (in English), Computers & Security, Article vol. 92, p. 23, May 2020, Art. no. Unsp 101738.

[9] S. C. Sethuraman, V. Vijayakumar, and S. Walczak, "Cyber Attacks on Healthcare Devices Using Unmanned Aerial Vehicles," (in English), Journal of Medical Systems, Article vol. 44, no. 1, p. 10, Jan 2020, Art. no. 29.

[10] M. A. Kosan, O. Yildiz, and H. Karacan, "Comparative analysis of machine learning algorithms in detection of phishing websites," (in Turkish), Pamukkale University Journal of Engineering Sciences-Pamukkale Universitesi Muhendislik Bilimleri Dergisi, Article vol. 24, no. 2, pp. 276-282, 2018.

[11] O. S. Lih et al., "Comprehensive electrocardiographic diagnosis based on deep learning," (in English), Artificial Intelligence in Medicine, Article vol. 103, p. 8, Mar 2020, Art. no. Unsp 101789.

[12] D. Zhang et al., "Automatic corneal nerve fiber segmentation and geometric biomarker quantification," (in English), European Physical Journal Plus, Article vol. 135, no. 2, p. 16, Feb 2020, Art. no. 266.

[13] A. Cuzzocrea, F. Martinelli, and F. Mercaldo, Applying Machine Learning Techniques to Detect and Analyze Web Phishing Attacks(Iiwas2018: The 20th International Conference on Information Integration and Web-Based Applications & Services). New York: Assoc Computing Machinery, 2014, pp. 355- 359.

[14] X. W. Liu and J. M. Fu, "SPWalk: Similar Property Oriented Feature Learning for Phishing Detection," (in English), Ieee Access, Article vol. 8, pp. 87031- 87045, 2020.

[15] J. Mao et al., "Detecting Phishing Websites via Aggregation Analysis of Page Layouts," in 2017 International Conference on Identification, Information and Knowledge in the Internet of Things, vol. 129, R. Bie, Y. Sun, and J. Yu, Eds. (Procedia Computer Science, Amsterdam: Elsevier Science Bv, 2018, pp. 224-230.

[16] N. Shulzhenko and S. Romashkin, "Internet fraud and transnational organized crime," (in English), Juridical Tribune Tribuna Juridica, Article vol. 10, no. 1, pp. 162-172, Mar 2020.

[17] A. Zamir et al., "Phishing website detection using diverse machine learning algorithms," (in English), Electronic Library, Article vol. 38, no. 1, pp. 65-80, Jan 2020.

[18] A. Belabed, E. Aimeur, A. Chikh, and Ieee, A personalized whitelist approach for phishing webpage detection (2012 Seventh International Conference on Availability, Reliability and Security). Los Alamitos: Ieee Computer Soc, 2012, pp. 249- 254.

[19] E. Buber, O. Demir, O. K. Sahingoz, and Ieee, Feature Selections for the Machine Learning based Detection of Phishing Websites(2017 International Artificial Intelligence and Data Processing Symposium). New York: Ieee, 2017.

[20] V. Patil, P. Thakkar, C. Shah, T. Bhat, S. P. Godse, and Ieee, Detection and Prevention of Phishing Websites using Machine Learning Approach (2018 Fourth International Conference on Computing Communication Control and Automation). New York: Ieee, 2018.