*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 618-623*

# Enhanced Techniques in Phishing website detection Using Machine Learning Algorithms

B. Sravani
*PG Research Scholar*
*Department of Computer Applications*
*Madanapalle Institute of Technology & Science, Chittoor Dist. A.P. India*

Mr. S. Balamurugan
*Assistant Professor*
*School of Computers*
*Madanapalle Institute of Technology & Science, Chittoor Dist. A.P. India*

Mrs.S. Savitha
*Assistant Professor*
*Department of Computer Science*
*Saradha Gangadharan College, Pondicherry*

***ABSTRACT:***
*Phishing is a common attack on credulous people by making them disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin personal information like user names, passwords, and online banking transactions. Phishers use websites that are visually and semantically similar to those real websites. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. This paper surveys the features used for detection and detection techniques using machine learning.*
***Keywords:*** *Phishing, Phishing Websites, Detection, Machine Learning, Safe and Security.*

---

Date of Submission: 29-05-2022 Date of acceptance: 10-06-2022

---

## I. INTRODUCTION:

Phishing is the most unsafe criminal exercise in cyberspace. Since most of the users go online to access the services provided by the government and financial institutions, there has been a significant increase in phishing attacks over the past few years. Phishers started to earn money and they are doing this as a successful business. Identity theft is one of the safest online crimes. With more and more users accessing the Internet to access services provided by the government and financial institutions, there has been a dramatic increase in attacks on identity theft over the past few years. The Fishermen started making money and this they did as a successful business. Various methods are used by fraudsters to attack vulnerable users such as text messages, VOIP, poofed links, and fake websites. It is very easy to build fake websites, which look like real websites in terms of structure and content. Or, the content of these websites will be similar to their official websites. The reason for creating these websites is to access private data from users such as account numbers, login IDs, debit card passwords and credit cards, etc. In addition, attackers ask security questions that they must answer in practice as a high-level security measure provided by users. When users answer those questions, they are easily caught in the crossfire of identity theft. Many types of research have been ongoing to prevent the spread of identity theft by various communities around the world. Machine learning algorithms have been one of the powerful techniques in detecting phishing websites. In this study, various methods of detecting phishing websites have been discussed.

## II. LITERATURE SURVEY:

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

In the last few years, many fake websites have developed on the World Wide Web to harm users by stealing their confidential information such as account ID, user name, password, etc. Phishing is a social engineering attack and currently attacks mobile devices. That might result in the form of financial losses. In this

paper, we described many detection techniques using URL, Hyperlinks features that can be used to differentiate between defective and non-defective websites. There are six main approaches such as heuristic, blacklist, Fuzzy Rule, machine learning, image processing, and CANTINA-based approach. It delivers a good consideration of the phishing issue, a present machine learning solution, and future studies about Phishing threats by using the machine learning Approach.

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

Phishing is a common attack on credulous people by making them disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin personal information like user names, passwords, and online banking transactions. Phishers use websites that are visually and semantically similar to those real websites. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. This paper surveys the features used for detection and detection techniques using machine learning.

**3.T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput.**

Phishing attacks are one of the most common and least defended security threats today. We present an approach that uses natural language processing techniques to analyze text and detect inappropriate statements which are indicative of phishing attacks. Our approach is novel compared to previous work because it focuses on the natural language text contained in the attack, performing semantic analysis of the text to detect malicious intent. To demonstrate the effectiveness of our approach, we have evaluated it using a large benchmark set of phishing emails.

**4.M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," 6th Int. Symp. Digit. Forensics Secure. ISDFS - Proceeding.**

These days, numerous enemies of phishing frameworks are being created to recognize phishing substances in online correspondence frameworks. Despite the accessibility of hordes hostile to phishing frameworks, phishing proceeds unabated because of lacking recognition of a zero-day assault, pointless computational overhead, and high bogus rates. Even though Machine Learning approaches have accomplished promising exactness rates, the decision and the exhibition of the component vector limit their successful location. Phishing is a typical assault on guileless individuals by making them unveil their one-of-a-kind data utilizing fake sites. In this work, an upgraded AI-based prescient model is proposed to improve the effectiveness of phishing plans. The prescient model comprises of Feature Selection Module which is utilized for the development of a successful element vector. These highlights are removed from the URL, website page properties, and site page conduct utilizing the gradual segment-based framework to introduce the resultant component vector to the prescient model. The proposed framework utilizes CNN, KNN, AND SVM which have been prepared on a 30-dimensional list of capabilities. AI is an incredible asset used to endeavor against phishing assaults

**5. K. Shima et al., "Classification of URL bitstreams using a bag of bytes," in 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN).**

At present days, websites are mainly responsible for the rapid growth of criminal activities on the internet and corresponding activities which results in many illegal things. So, there are many preventive steps to be taken to stop this kind of activity. Here we propose a model which will classify the given URL into any of the three possible classes, i.e., Benign, spam, and malware. Our model will then detect the classification of the URL without using any website content.

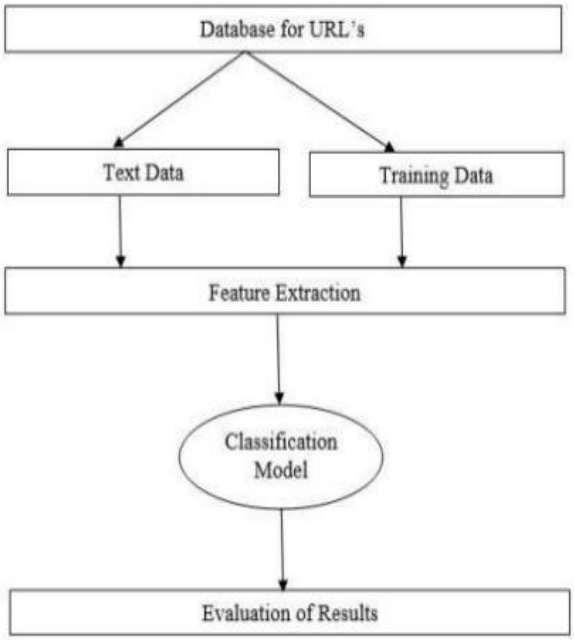
**PROPOSED SYSTEM**

Machine Learning is cutting edge and trending for different kinds of diverse applications in a society where it can deal with tons of data, refined and revised algorithms, and available heavy processing power in terms of GPU. Many types of research have been going on to prevent phishing attacks by different communities around the world. Phishing attacks can be prevented by detecting the websites and creating awareness among users to identify the phishing websites. In this study, various methods of detecting phishing websites have been discussed.

**Advantages:**

1. Fast process
2. Less time

3. Accurate result




The figure shows data flow diagram for the proposed system.

**IMPLEMENTATION**

**A. Feature extraction:**

The feature extraction process is done from the URLs and corresponding binary values are given indicating whether the website is a phishing website or not. Below are the features that We can extract for the detection of fraud URLs.

**1. IP address in URL:** If the IP address is present in the URL, then the feature is set to 1 else set to 0. Most legitimate sites do not use an IP address in an URL to download a webpage. The use of an IP address in the URL indicates that the attacker is trying to collect sensitive information.

**2. '@' symbol in URL:** If @ symbol is present in the URL then the feature is set to 1 else set to 0. Hackers adding the special symbol @ in the URL lead the browser to ignore everything before the rate (@) symbol and the real address often follows the "@" symbol.

**3. Prefix or Suffix separated by (-) to the domain:** If the domain name is separated by the dash (-) symbol then the feature is set to 1 else to 0. This '-' symbol is rarely used in legitimate URLs. Phishers add the hyphen symbol (-) to the domain name so that users feel that they are dealing with a legitimate Webpage.
http://www.onlineamazon.com to trick the innocent users.

**B. Random Forest Algorithm:**

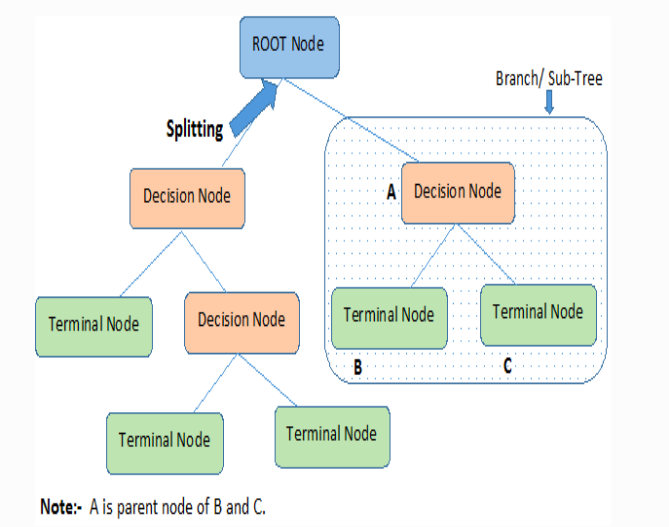
Random Forest is a machine-learning algorithm that belongs to the supervised learning technique; it can be applied for Classification and Regression problems. It is based on Phishing Website Detection using a Machine Learning System Implementation of the concept of Associative learning, which is a process of combining many different classifiers to solve a complex problem and to improve the efficiency of the model. A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions.

**C.Decision tree:**

Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

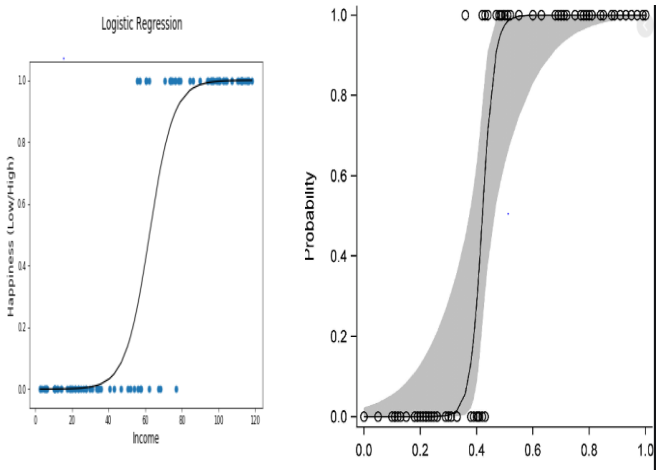
A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn't split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

**Block Diagram for Decision Tree Algorithm:**




**D. LOGISTIC REGRESSION:**

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).




This linear relationship can be written in the following mathematical form (where $\ell$ is the log-odds, is the base of the logarithm, and are parameters of the model):

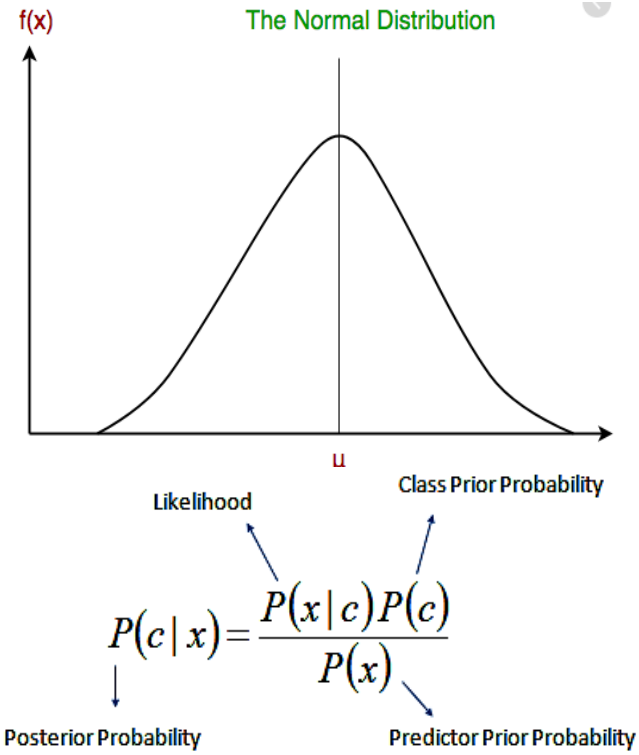
$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

This linear relationship can be written in the following mathematical form (where $\ell$ is the log-odds, is the base of the logarithm, and are parameters of the model.

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

**E. NAIVE BAYES ALGORITHM:**

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a **Naive Bayes classifier** assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.




$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

## III. RESULTS

The proposed system enables people to have safe browsing and safe transactions. It helps users to save their important credentials that should not be leaked. Our proposed system tells whether the given website is legitimate or not to users in the form of an extension that makes the process of finding the truth about the website much easier. The results point to the efficiency with which our proposed system works to achieve the result using the hybrid solution of heuristic features, visual features, and various approaches feeding these distinct features to machine learning algorithms. And thus, we use online learning algorithms. This new system can be designed to make use of maximum accuracy. Using different approaches altogether will improve the precision of the system, providing an efficient protection system. The drawback of this system is detecting some minor false-positive and false-negative results. These disadvantages can be abolished by introducing much-enhanced features to feed to the machine learning algorithm that would result in much higher accuracy.

**FUTURE WORK**

Future work should focus on the direct implementation of the paper to the chrome extension so that as the user clicks on the particular URL and if that URL is a phishing site, then the user gets a pop-up warning message.

## IV. CONCLUSION

This paper presented various algorithms and approaches to detect phishing websites by several researchers in Machine Learning. On reviewing the papers, we concluded that most of the work was done by using familiar machine learning algorithms like Logistic Regression, Decision Tree, and Random Forest algorithm Some authors proposed a new system like Phish Score and Phish Checker for detection. The combinations of features with regards to accuracy, precision, recall, etc. were used. As phishing websites increase day by day, some features may be included or replaced with new ones to detect them.

## REFERENCES


[1]. Balamurugan, S., Ayyasamy, A. & Joseph, K.S.Enhanced Petri nets for traceability of food management using the internet of things. *Peer-to- PeerNetw.Appl.* **14,** 30–43(2021). https://doi.org/10.1007/s12083-020-00943-0

**[2].** Balamurugan, S., Ayyasamy, A. & Joseph, K.S. IoT-Blockchain driven traceability techniques for improved safety measures in food supply chain. *Int. j. inf. tecnol.* **14,** 1087–1098 (2022). https://doi.org/10.1007/s41870-020-00581-y

[3]. J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.

[4]. Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, "Phishing websites features classification based on extreme learning machine," 6th Int. Symp. Digit. Forensics Secure. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

**[5].** T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[6]. M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," 6th Int. Symp. Digit. Forensics Secure. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[7]. S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.

[8]. K. Shima et al., "Classification of URL bitstreams using a bag of bytes," in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.

[9]. A. Vazhayil, R. Vinayakumar, and K. Soman, "Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks," in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 1– 6.

[10]. W. Fadheel, M. Abusharkh, and I. Abdel-Qader, "On Feature Selection for the Prediction of Phishing Websites," 2017 IEEE 15th Intl Conf Dependable, Auton. Secure. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017.

[11]. Srushti Patil, and Sudhir Dhage, "A Methodical Overview on Phishing Detection Along with An Organized Way to Construct an AntiPhishing Framework", 2019 5th International Conference on Advanced Computing & Communication System (ICACCS), pp. 1-6.

[12]. Nikhil K, Dr. Rajesh D S, Dhanush Raghavan, "Phishing Website Detection Using ML", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN: 2456-3307, Volume 7 Issue 4, pp. 194-198, July-August 2021.