



**IJIRCCCE**

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH


IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 4, April 2023

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

**Impact Factor: 8.379**

 9940 572 462

 6381 907 438

 [ijircce@gmail.com](mailto:ijircce@gmail.com)

 [www.ijircce.com](http://www.ijircce.com)

# Survey on Detection of Phishing Sites Using Machine Learning

Divya R. Dhamdhere, Abhishek Sonawane, Omkar Shinde, Amogh Shimpi, Sameer Sonawane

Assistant Professor, Dept. of Computer Engineering, NBNSO, Pune, Maharashtra, India

UG Student, Dept. of Computer Engineering, NBNSO, Pune, Maharashtra, India

UG Student, Dept. of Computer Engineering, NBNSO, Pune, Maharashtra, India

UG Student, Dept. of Computer Engineering, NBNSO, Pune, Maharashtra, India

UG Student, Dept. of Computer Engineering, NBNSO, Pune, Maharashtra, India

**ABSTRACT:** Phishing is a growing concern for security researchers because it is easy to create fake websites that look almost identical to legitimate ones. While experts can typically identify fake websites, many users cannot and become victims of phishing attacks. Attackers aim to steal banking credentials, resulting in an estimated annual loss of \$2 billion for US businesses. One common method for detecting phishing websites is through the use of blacklisted URLs and IP addresses in antivirus databases, known as the "blacklist" method. However, attackers can evade blacklists by using techniques such as obfuscation, fast-flux, and algorithmic generation of new URLs to make their fake websites appear legitimate.

In addition to the blacklist method, there are other techniques that can be used to detect phishing websites. For example, some researchers are developing machine learning algorithms that can analyze website content and detect patterns that indicate a website may be fake. These algorithms can be trained using large datasets of known phishing websites and legitimate websites to improve their accuracy.

Another approach is to educate users about the common signs of phishing attacks and how to avoid them. This can include things like checking the URL for spelling errors or unusual characters, looking for security indicators such as the padlock icon in the address bar, and avoiding clicking on links or downloading attachments from suspicious emails.

**KEYWORDS:** Cybersecurity ; Social Engineering ; Spear Phishing ; SVM ; Decision Tree ; Random Forest

## I. INTRODUCTION

Phishing is a type of fraud where attackers disguise themselves as trustworthy entities in electronic communication to obtain sensitive information, such as usernames, passwords, and credit card details, for malicious purposes. Phishing attacks are a major concern for security researchers because attackers can easily create fake websites that look nearly identical to legitimate ones. While experts can typically identify fake websites, many users cannot and fall victim to these attacks. Attackers aim to steal banking credentials, resulting in significant financial losses for businesses. Phishing attacks are successful because of the lack of user awareness, and it is challenging to mitigate them. Detecting phishing websites typically involves the use of blacklisted URLs and IP addresses in antivirus databases, known as the "blacklist" method. However, attackers can evade blacklists by using techniques such as obfuscation and algorithmic generation of new URLs. Social engineering tactics such as using deceptive email addresses and messages also play a role in phishing attacks. Enhancing phishing detection techniques is crucial to combat these attacks and protect sensitive information from falling into the wrong hands.

To make phishing websites look legitimate, attackers often use social engineering tactics to deceive users. For example, they may create emails that appear to come from a trusted source and ask the user to click on a link that leads to a fake website. The email may contain urgent or important-sounding language to prompt the user to act quickly without thinking carefully.

Attackers can also use spear phishing, which is a targeted form of phishing that involves researching the victim and creating a personalized message that appears to be from someone the victim knows or trusts. By using this technique, attackers can increase the chances of the victim falling for the scam.

To combat phishing attacks, it's important to educate users about how to recognize and avoid them. This includes being cautious about clicking on links or downloading attachments in unsolicited emails, verifying the legitimacy of a website before entering sensitive information, and using two-factor authentication and password protection to make it more difficult for attackers to gain access to accounts.

Encryption can also play a role in protecting sensitive information from phishing attacks. By encrypting data both in transit and at rest, attackers will be unable to read or access the data even if they are successful in intercepting it.

Overall, it's important to take a multi-layered approach to cybersecurity to mitigate the risks of phishing attacks. This includes implementing security measures such as antivirus software, firewalls, and intrusion detection systems, as well as providing regular security awareness training to employees and keeping up to date with the latest security threats and vulnerabilities.

## II. PROPOSED DEFINITION

Phishing attacks can take different forms, including email phishing scams and spear phishing. Users should be cautious and not fully trust common security applications as they can be vulnerable to such attacks. Machine learning can be used as an effective technique to detect phishing and overcome the limitations of existing approaches. Machine learning is a branch of artificial intelligence that has the ability to learn without explicit programming. Some of the commonly used machine learning techniques include supervised learning, unsupervised learning, and reinforcement learning. ML can be used in the development of information security applications, providing optimization, classification, prediction, and decision support systems, which can benefit those responsible for information security. The project aims to explore the use-case of detecting phishing websites using machine learning.

Machine learning is a branch of artificial intelligence that involves the development of algorithms and models which can learn from data and make predictions or decisions without being explicitly programmed.

Machine learning types of machine learning techniques are:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

### 1. Supervised learning

Supervised learning involves training a model on a labeled dataset, where each instance is associated with a target value or output. The goal is to learn a function that can map input data to output data with a high degree of accuracy. Examples of supervised learning algorithms include logistic regression, decision trees, and neural networks.

### 2. Unsupervised learning

Unsupervised learning involves training a model on an unlabeled dataset, where there is no target value or output associated with each instance. The goal is to discover patterns or structure in the data. Examples of unsupervised learning algorithms include clustering, principal component analysis, and anomaly detection.

### 3. Reinforcement learning

Reinforcement learning involves training a model to make decisions based on feedback received from the environment. The model learns by trial and error, adjusting its behavior based on the feedback it receives. Examples of reinforcement learning algorithms include Q-learning and policy gradient methods.

There are also many different machine learning algorithms that can be used with these techniques. One such algorithm is the Extreme Learning Machine (ELM), which is a feed-forward artificial neural network with a single hidden layer. The ELM requires appropriate parameter values such as threshold value, weight, and activation function for the data

system to be modeled. In gradient-based learning approaches, these parameters are changed iteratively for appropriate values.

Another algorithm is the Random Forest Algorithm, which is an ensemble learning classification and regression method suitable for handling problems involving grouping of data into classes. In RF, prediction is achieved using decision trees. During the training phase, a number of decision trees are constructed (as defined by the programmer) which are then used for class prediction.

The Decision Tree algorithm looks like a flowchart, where each non-leaf node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The C4.5 algorithm is used to construct decision trees through learning from class-labeled training tuples. It adopts a greedy, top-down recursive divide-and-conquer approach to construct decision trees. Attribute selection is a key problem in constructing decision trees, which determines which attribute to be split (i.e., as a non-leaf node of the decision tree).

Lastly, the SVM (Support Vector Machine) algorithm is used in medicine for diagnosis of diseases, text recognition, and for classification of images, among other fields. This algorithm partitions the data into two categories using a fixed rule, quadratic equation, and statistics. A separating hyperplane is used for the binary classification of the data and minimizes the space of the margin on the basis of the kernel function. However, this technique may fail in analyzing big data.

### III. CONCLUSION AND FUTURE WORK

Websites are widely used in various fields for data entry and information processing applications. However, their popularity has also made them an attractive target for cyber threats such as phishing attacks. These attacks are typically designed to deceive users into providing their sensitive information or downloading malicious software by presenting a fake website that looks similar to the original one.

To combat phishing attacks, various detection methods and approaches have been developed and implemented. In this context, an application is proposed that uses Extreme Learning Machine to classify and detect phishing URLs. The system aims to inform users about the presence of phishing URLs by identifying them as malicious even before the user visits the website.

The proposed system is based on a dataset obtained from the UCI website. The Extreme Learning Machine algorithm will be used to process the data and classify URLs as either phishing or benign. By using this approach, users can be alerted of potential phishing attacks and take appropriate measures to protect their sensitive information. The system can play a crucial role in mitigating the risks associated with phishing attacks and enhancing the security of web-based applications.

### REFERENCES

1. Oza Pranali P, Deepak Upadhyay, Review on Phishing Sites Detection Techniques, IJERT, ISSN: 2278-0181, 04, April-2020
2. Meenu, Sunila godara, Phishing Detection using Machine Learning Techniques, IJEAT, ISSN: 2249 – 8958, December, 2019
3. Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick, Lavanya Badiginchala, Ravali Reddy Gudur, Siri Chandana Guttha, IJITEE, ISSN: 2278-3075, June 2019
4. Ankit Kumar Jain and B.B. Gupta EURASIP Journal on Information Security (2016) 2016:9
5. Joby James, Sandhya L., Ciza Thomas, Detection of Phishing URLs Using Machine Learning Techniques, 2013 International Conference on Control Communication and Computing (ICCC), December 2013
6. Mohammed Hazim Alkawaz, Stephanie Joanne Steven, Asif Iqbal Hajamydeen, Detecting Phishing Website Using Machine Learning, 2020 16th IEEE International Colloquium on Signal Processing & its Applications, 28-29 Feb. 2020





7. Suleiman Y. Yerima, Mohammed K. Alzaylaee, High Accuracy Phishing Detection Based on Convolutional Neural Networks, IEEE Xplore
8. Megha N, K R Remesh Babu, Elizabeth Sherly, An Intelligent System for Phishing Attack Detection and Prevention, IEEE Xplore ISBN: 978-1-7281-1261-9, 2019 IEEE
9. Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Dr.Aram Alsedrani, Detecting Phishing Websites Using Machine Learning 978-1-7281-0108-8/19/ 2019 IEEE
10. [https://www.hindawi.com/journals/jam/2014/425731/\(randomforest\)](https://www.hindawi.com/journals/jam/2014/425731/(randomforest))
11. <https://pdfs.semanticscholar.org/41ca/257920b5b5e6c1cf4f4417bb85ac5a875935.pdf>
12. <https://archive.ics.uci.edu/ml/index.php>
13. [https://www.google.com/url?q=https://towardsdatascience.com/phishing-domain-detection-withml5be9c99293e5&sa=D&source=hangouts&ust=1604419765459000&usg=AFQjCNEYdPDaUh3C\\_34k2ofrkdDqq\\_D](https://www.google.com/url?q=https://towardsdatascience.com/phishing-domain-detection-withml5be9c99293e5&sa=D&source=hangouts&ust=1604419765459000&usg=AFQjCNEYdPDaUh3C_34k2ofrkdDqq_D)



Impact Factor: 8.379



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  [ijircce@gmail.com](mailto:ijircce@gmail.com)



[www.ijircce.com](http://www.ijircce.com)

Scan to save the contact details