SURVEY ON DETECTION OF PHISHING WEBSITE USING MACHINE LEARNING

Yawalkar Prasad Pramod[1], Dr. Prashant N. Chatur[2], Dr. Kamlesh A. Waghmare[3]

[1] Student, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India
[2] Professor, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India
[3] Assistant Professor, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India

ABSTRACT

Phishing is a form of social engineering attack aimed at stealing personal information from individuals through websites or email. This information can include passwords, credit card details, bank account information, and other personal data. Phishing takes place when a malicious actor attempts to impersonate a reputable service provider. Various traditional methods and new techniques are utilized to enhance the detection accuracy and reduce the success rate of phishing websites in stealing information. Several anti-phishing techniques are there such as blacklist, heuristic, visual similarity and machine learning. Machine Learning is efficient technique to detect phishing. It also removes drawback of existing approach. Different algorithms are used with the assistance of Machine Learning techniques to identify Phishing Websites. The main goal of this research is to identify phishing websites through the utilization of Machine Learning techniques. Machine Learning Algorithms like Random Forest, Decision Tree, K-Nearest Neighbors, Gradient Boosting, XGBoost, Support Vector Machine, Naïve Bayes, and more are used to detect phishing websites.

Keyword : - *Phishing, URL, Cybercrime, Personal Information.*

1. INTRODUCTION

The rapid advancement of technology has made internet use an essential part of our daily lives. People around the world rely on the internet to carry out various activities such as commercial transactions, business management, communication, and shopping. However, the internet does not offer robust security, creating opportunities for malicious individuals to exploit it for financial gain or data theft. As a result, phishing has become a significant incentive for many attackers, as cybercrime carries both risks and substantial rewards.

Security researchers are greatly concerned about phishing today, as scammers are making fake websites that look very similar to real ones. Phishing is a fraudulent tactic used by attackers to deceive internet users into disclosing their private information or log-in details in order to make money. Modern phishing attacks have become more complex, posing a greater challenge for detection. In a study conducted by Intel, it was found that 97% of security professionals have difficulty differentiating between authentic emails and phishing emails[1]. The majority of phishing attempts involve sending emails to users online that seem to be from a respected organization or individual. These messages frequently lead the recipient to verify their personal information, like passwords or other sensitive details, resulting in the victim being caught in the phishing scam and enabling the attacker to accomplish their objective. The APWG stated that there were 877,536 phishing attacks in the months of April, May, and June in the second quarter of 2024[2]. Phishing was first introduced on January 2, 1996, on American Online (AOL), a company that offers internet services. The scammer created AOL accounts by generating credit card numbers at random. Afterwards, they utilized AOL's instant messaging or email platform to contact customers, requesting them to confirm their account information by clicking on a link included in the email. Upon clicking the link and inputting their login information, users unknowingly had their data immediately transmitted to the hacker, who subsequently misused it for fraudulent activities.

Machine learning uses algorithms to identify phishing websites by learning the traits of phishing sites and then forecasting new ones. Multiple algorithms such as Random Forest (RF), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), Support Vector Machines (SVM), and Gradient Boosting (GB) are utilized in this procedure. Each of these algorithms provides varying degrees of accuracy when it comes to detecting phishing attempts.

Using machine learning algorithms to identify phishing websites offers an advantage over traditional techniques such as blacklists or heuristic-based systems because they are more precise and efficacious. In contrast to those techniques, machine learning algorithms can be taught to recognize phishing websites by analyzing their characteristics instead of depending on preset rules or signatures. This increases their strength and decreases the chances of incorrect results.

2. LITERATURE SURVEY

In the year 2020, R. M. Kapoor, S. K. Singh, and R. K. Singh published a paper in which Gradient Boosting Classifier (GBC) approach for detecting phishing websites is proposed. The authors evaluate the performance of GBC using a dataset of 5,000 phishing and legitimate websites, 17 features (URL length, domain age, HTTPS usage, etc.) and got accuracy 97.5%.The study demonstrates the effectiveness of GBC in detecting phishing websites. The proposed approach achieves high accuracy and outperforms existing machine learning algorithms[3].

In 2019, S. S. Iqbal, M. A. Khan, and A. A. Abbasi proposed a model using Gradient Boosting Classifier in which 10,000 phishing and legitimate websites data are collected and they achieved 96.2% accuracy[4].

In 2017, Zhao et al. explored the efficacy of various machine learning algorithms, including gradient boosting, for phishing detection. They proposed a feature extraction method based on URL characteristics and evaluated the model's performance on a dataset of known phishing and legitimate websites. Their results indicated that gradient boosting outperformed traditional methods, achieving high accuracy and precision[5].

In the year 2020, Alazab et al. developed a hybrid phishing detection model combining gradient boosting with deep learning techniques. Their approach utilized a comprehensive set of features, including social media links and visual similarity metrics, leading to improved detection performance. The study demonstrated how gradient boosting can effectively complement other machine learning strategies in phishing detection[6].

In 2022, Choudhary et al. compared various machine learning algorithms, including gradient boosting classifiers, for phishing detection. Their findings indicated that gradient boosting achieved higher F1 scores compared to other classifiers[7].

In the year 2023, Singh et al. developed a gradient boosting-based model for detecting phishing URLs. The authors emphasized the importance of incorporating ensemble techniques to enhance detection capabilities[8].

In the year 2022, Sultana et al. evaluated the performance of gradient boosting against several other ensemble methods for phishing detection. They employed an extensive dataset and highlighted the superior performance of gradient boosting, particularly in terms of recall and F1-score, which are critical metrics for cyber security applications[9].

3. BACKGROUND THEORY

Phishing refers to the deceitful effort to acquire sensitive information such as usernames, passwords, and credit card details, typically for malicious purposes, by masquerading as a reputable entity in digital communications. Various types of phishing attacks can occur, including email phishing, website phishing, spear phishing, whaling, tab napping, and evil twin phishing. To prevent these phishing attacks, different anti-phishing measures should be utilized. There are several anti-phishing technologies available, including blacklists, heuristic approaches, visual similarity detection, and machine learning techniques. They are as follows :

- Blacklist Method -
  Blacklist Method is the most commonly used technique where a database stores a list of known phishing URLs. If a URL is found in the database, it is identified as a phishing URL and a warning is issued; otherwise, it is considered legitimate. This method is straightforward and quick to implement since it simply checks whether the URL exists in the database. However, a significant drawback is that even a minor alteration to the URL can allow it to evade detection by this list-based method, and frequent updates to the list are essential to address new threats.
- Heuristic-Based Method -
  Heuristic-Based Method builds upon the blacklist approach and can identify novel attacks by utilizing features extracted from phishing websites to recognize phishing attempts. However, its limitations include an inability to detect every new attack, and it can be easily circumvented once an attacker is aware of the

algorithm or features employed. Furthermore, its detection rate is inadequate because a site might lack certain common characteristics.

- Visual Similarity –
  Visual Similarity approach deceive user by extracting image of legitimate site. But limitation of this is image comparison takes more time as well as more space to store image .produces high false negative rate and fail to detect when visual appearance slightly changes.
- Machine Learning –
  Machine Learning approach works efficiently in large dataset. This also removes drawback of existing approach and able to detect zero day attack .Machine Learning based classifiers are efficient classifiers which achieved accuracy more than 99% .Performance depends on size of training data, feature set, and type of classifier. Limitation of this is it fails to detect when attacker use compromised domain for hosting their site. Various performance measure used for analysis of best algorithm are F-measure, precision, recall, accuracy, AUC, ROC curve etc.

4. MACHINE LEARNING ALGORITHMS

Machine learning provides efficient methods for analyzing data and has displayed encouraging outcomes in immediate classification responsibilities. The main benefit is its ability to create custom models tailored for specific tasks, like identifying phishing attempts. Identifying phishing emails is a type of challenge that can be effectively tackled using machine learning models. These models are able to rapidly adapt to fresh data, spotting trends in fraudulent behavior and aiding in the creation of a flexible, knowledge-driven detection system. The majority of the discussed machine learning techniques fall into the category of supervised learning, in which an algorithm learns to associate inputs with outputs using example pairs of input-output. This procedure includes creating a function using labeled training data. Here, we detail the machine learning techniques employed in our research.

- Decision Tree Algorithm : The Decision Tree algorithm is a commonly used and simple machine learning method. It is appreciated for its straightforwardness and simple application. The first step is to choose the top attribute for dividing the data, which becomes the root of the tree. The algorithm proceeds to construct the tree incrementally by choosing more attributes for splitting until it reaches the leaf nodes. Every branch node in the tree symbolizes a characteristic, while every leaf node corresponds to a class label. The algorithm utilizes metrics like the Gini index and information gain to find the best splits.
- Random Forest Algorithm : The Random Forest algorithm is an effective machine learning technique that expands upon the idea of Decision Trees. It creates a collection of decision trees forming a "forest". Having more trees typically leads to increased accuracy in detection. Trees are built through the bootstrap approach, where characteristics and instances from the data set are chosen randomly and with replacement to form each tree. The algorithm selects the best split for classification from a set of randomly chosen features. Just as with the Decision Tree algorithm, Random Forest also utilizes the Gini index and information gain to identify these divisions. This cycle continues until the forest reaches a set number of trees. Every tree predicts, and the Random Forest algorithm combines these predictions through voting, ultimately selecting the most common prediction as the final result.
- Support Vector Machine Algorithm : The Support Vector Machine (SVM) is a powerful algorithm in machine learning that is utilized for classification and regression purposes. In Support Vector Machines, every data point is represented in a space with n dimensions, and the goal is to identify a hyperplane that effectively divides the two classes. The SVM detects support vectors, which are the points closest to the hyperplane, and then connects them with a line. It creates a perpendicular separating hyperplane that maximizes the margin, the distance between the hyperplane and the support vectors. SVM uses the kernel trick to process intricate and non-linear data by transforming it into a higher-dimensional space for improved distinction
- K-Nearest Neighbors Algorithm : KNN is a non-parametric classification method that predicts based on the proximity of the target instance to its closest neighbors. Distance calculations for continuous data usually involve Euclidean distance, while Hamming distance is commonly used for categorical data.
- Naive Bayes Classifier : Naive Bayes is a simple classification method that assigns class labels to instances shown as feature vectors. It is efficient in tasks such as document classification, aiming to categorize documents using word frequencies. Even though it is simple, Naïve Bayes can compete well with more complicated techniques such as support vector machines, especially when the appropriate pre-processing is done. Efficient training can be achieved through supervision in a learning context. For example, websites containing many outbound links and areas to enter passwords could be marked as questionable.

- Logistic Regression : Logistic regression is a type of supervised learning algorithm used for tasks involving binary classification. It calculates the likelihood of a binary result (such as yes/no, success/failure) by examining single or multiple predictor factors (attributes).
- Gradient Boosting Classifier : Gradient Boosting is a machine learning technique that combines the predictions of multiple weak models, usually decision trees, in a progressive manner. The aim is to enhance the overall predictive accuracy by adjusting the model's weights using the errors from previous rounds. This step-by-step method minimizes forecast mistakes and improves the model's precision. It is frequently utilized for tasks involving linear regression.

5. CONCLUSIONS

Phishing is a method used to acquire a user's sensitive information through email or websites. With the extensive usage of the internet today, nearly everything can be found online, whether it's shopping for clothing, electronic devices, household items, or paying bills for mobile services, television, and electricity. Instead of waiting in long lines for hours, people are increasingly choosing digital methods. This presents phishers with ample opportunities to carry out phishing scams. Although significant research has been conducted in this field, no single technique suffices to identify all forms of phishing attacks. As technology evolves, phishing attackers are developing new strategies each day. This underscores the need for effective classifiers for the detection of phishing. Many of research have been performed in this area of phishing detection. Most research has worked on improving accuracy of phishing website detection using different classifiers. Various Classifiers used are KNN, SVM, Gradient Boosting, Decision tree, ANN, Naïve Bayes, PART, ELM and Random forest. Among all of this Gradient Boosting classifiers is best as in terms of accuracy as per my literature survey.

In this paper, we performed detailed literature survey about phishing website detection. According to this, we can say Gradient Boosting Classifier in Machine Learning approach is best suitable than other.

6. REFERENCES

[1]. Intel, Phishing Attack Survey, 2018

[2]. APWG. APWG phishing trends report 2nd quarter 2024, 2024.

[3]. R. M. Kapoor, S. K. Singh, and R. K. Singh , "Phishing Website Detection Using Gradient Boosting Classifier",2020 International Journal of Advanced Research in Computer Science

[4]. Iqbal, S. S., Khan, M. A., & Abbasi, A. A., "A Gradient Boosting Classifier Model for Phishing Website Detection." Journal of Cybersecurity, 5(3), pp 112-118, 2019.

[5]. Zhao, X., Li, Y., & Wang, Z., "Exploring Machine Learning Algorithms for Phishing Detection: A Gradient Boosting Approach." Journal of Information Security and Applications, 34, pp 52-58, 2017.

[6]. Alazab, M., Tang, M., & Arora, A., "A Hybrid Phishing Detection Model Combining Gradient Boosting and Deep Learning Techniques." Journal of Cybersecurity and Privacy, 6(2), pp 102-109, 2020.

[7]. Choudhary, P., Sharma, R., & Gupta, S., "Comparative Analysis of Machine Learning Algorithms for Phishing Detection: A Focus on Gradient Boosting Classifiers." Journal of Applied Machine Learning, 10(4), pp 233-240, 2022.

[8]. Singh, R., Kumar, P., & Mehta, S., "A Gradient Boosting-Based Model for Phishing URL Detection: Enhancing Detection Capabilities with Ensemble Techniques." International Journal of Cybersecurity and Digital Forensics, 15(1), pp 45-52, 2023.

[9]. Sultana, S., Alam, M., & Rahman, M.,"Evaluation of Gradient Boosting and Other Ensemble Methods for Phishing Detection: A Comprehensive Performance Analysis." Journal of Cybersecurity Research, 18(3), pp 210-217, 2022.