Detection of Phishing Sites Using Machine Learning

Pooja Maroti Jadhav[1], Tejashree Devendra More[2,] Vaishnavi Hemant Bagul[3,] Prof. S.V. Shardul[4]

**1,2,3 Department of Computer Engineering, Shatabdi Institute of Engineering and Research, Nashik, Maharashtra, India.*

**4 Professor, Department of Computer Engineering, Shatabdi Institute of Engineering and Research, Nashik, Maharashtra, India.*

Abstract - The content of phishing websites and online-based material provides a variety of indicators. One of the criminal attacks that are successful in the online world is phishing sites that direct users to a phishing website that looks and acts like a legitimate website in order to steal the victim's personal and sensitive information. The Extreme Learning Machine-based version that was suggested was excellent at spotting phishing websites. Interne page types vary greatly in terms of their characteristics. So, in order to protect against any phishing attack, we must use a set of web page features. To defend against these threats, a machine learning strategy is used.

The phishing dataset that is planned to be imported, authentic URLs from the database, and also the data that is collected are all pre-processed. Four groups of URL characteristics are used to detect phishing websites: domain, address, anomalous based, HTML, and JavaScript features. Data that has been analyzed is used to extract URL characteristics and produce URL attribute values. ML approaches are used to analyze URLs and determine the threshold value and range value for URL properties. This project's goal is to develop an ELM categorization for a number of database characteristics and a few phishing sites.

Keywords: Support Vector Machine (SVM), Extreme Learning Machine (ELM), URL Phishing Websites, Browser add-ons, Random Forest.

1. INTRODUCTION

Phishing is the tactic of pretending to be a trustworthy entity during an internet encounter in order to get sensitive information such as usernames, passwords, and credit card numbers, usually for malicious purposes [1]. Nowadays, safety researchers are concerned about the idea of phishing's in cent's simple to create a fake website that mimics age nine website. Although experts can spot fake websites with ease, most people struggle to tell them apart, and as a result, they become targets of phishing assaults. The attacker's primary goal is to steal login information for financial institution accounts. Customers who fall victim to phishing cost American businesses $2 billion annually [3].

2. Prior Work

In this section, salient features of existing approaches are summarized. Abu-Nimeh et al. [5] compared the predictive accuracy of several machine learning methods, including LR, CART, RF, NB, SVM, and BART. They used 43 features to capture phishing emails. They analyzed 1,718 legitimate emails and 1,117 phishing emails. For the case of Random Forests, their results showed a 7.72% as the lowest error rate. Basnet et al. [6] evaluated six diverse detection methods that are based on machine learning. Using 12 features, they analyzed 3,027 legitimate emails and 973 phishing emails. Their results showed the lowest error rate at

2.01%. For [7] and [8], even though they used different experimental parameters, their detection of phishing emails, based on machine learning, led to high accuracy.

In [9], the solution to the phishing emails problem was provided. It checks the email for 17 elements of phishing emails. Some of them are not suitable in modern life because they don't follow the current trends. For instance, if the URL has '@' sign, it will be interpreted as a phishing attack, or if the text inside the anchor HTML tag is not the same as that in its attribute, it will also be considered as phishing. Still, now to register on some sites, you need to follow the link with the text "Click here to finish the registration." [10] provided the solution of the detection of a phishing attack using neural networks, features of MIME, and checking the context of the email on 6 elements, one of them being the total number of links in the message. Some of the new ad sites send emails with 10 and more links on their products, and they will be considered as phishers even though they are not.

3. Methodology

Using implicit programming, this branch of artificial intelligence has the capacity to learn. Unsupervised learning, reinforcement learning, and other machine learning methods are all supervised. The many machine learning methods include:

A. Supervised learning
B. Unsupervised learning
C. Reinforcement learning

Machine Learning Algorithms:

Extreme learning machine (ELM):

The extreme learning machine (ELM) is a one-hidden-layer Artificial Neural Network (ANN) model. The threshold value, weight, and activation parameters of an ANN must have values appropriate to the data system to be modeled if advanced learning is to be ensured. All of those parameters are repeatedly adjusted to pertinent values in gradient-based learning systems.

Random forest algorithm:

A set learn in regression and category approach called random forest (RF) may be used to solve issues with data classification. Decision trees are used in RF to make predictions. A few decision trees are

Constructed during the training phase (specified by the programmer) and then utilized for class

Prediction; this is done by taking into account the graded classes of all individual trees, with the class with the highest grade being the output [10]

SVM (Support Vector Machine):

This method is used in the medical industry for illness detection, text content recognition, picture classification, and other purposes. With the use of a fixed rule, a quadratic equation, and statistics, the data will be divided into classes in this way. For the binary classification of the data, a separating hyper plane is utilized, which reduces the space of the margin based on kernel characteristics. This method is employed to identify the ideal response to the issue. This method does not analyze large amounts of data [2].

4. Literature Survey

1. N Megha, K R Ramesh Babu Elizabeth Sherly (2020) **An Intelligent System for Phishing Attack Detection and Prevention** worked on Implementing a multi agent-based architecture and ML classifier for detecting and rectifying web phishing attacks. (Focus on Accuracy)
2. Suleiman Y. Yerima, Mohammed K. Alizaylaee (2020) **High Accuracy Phishing Detection Based on Convolutional Neural Networks** worked on This paper proposes a

deep learning model based on ID CNN for the detection of phishing websites.

3. Suleiman Y. Yerima, Mohammed K. Alizaylaee (2021) **Phishing and non- non phishing website classification.** Worked on It Proposes feature selection method are also used to increase the accuracy of classification model by selecting best feature & result.
4. SK Hasane Ahammad, Sunil D. Kale, Amol V. Dhumane (2022) **Phishing URL detection using machine learning methods worked** on ML Algorithms such as Random Forest, Decision Tree, Light Lgm and SVM is used. (Focus on Multiple Algorithm)

5. Algorithm and Sequence Flow:

The most popular machine learning classifier for drawing the best connection between two lessons is the SVM classifier. The categorization challenges are addressed by the predictive analytic algorithm known as logistic regression. Random forest performs well in classifications in ceit only looks with in randomly chosen predictors for the best split.

The following Steps to be followed are:

1. Choose random samples from a predetermined dataset.
2. For each pattern, this formulation will then construct a decision tree. After then, it will receive the results of each decision tree's forecast.
3. Voting is conducted for each expected result at some point in this step.
4. Finally, choose the top-rated for cast result to serve as the final forecast outcome.

In contrast to ANN, which updates its parameters as supported gradients, ELM researching techniques hand-picks the input weights in every way while analytically calculating the o/p weights.

The following Steps to be followed are:

1. Go to a website or webpage on the internet
2. Review the policies for the thirty entry attributes supported characteristics.
3. Grouping samples into the dataset is step three.
4. Randomly select 90% of the dataset's training samples and 10% of its test samples.
5. Classification using ELM.
   5.1. Randomly assign hidden nodes and generate the parameters for hidden nodes.
   5.2. Determine the hidden layer's output matrix.
   5.3. Compute the output weight matrix in section
6. Phishing or real predictions

The intended method for importing genuine computer addresses and phishing information sets from the data is preprocessed, as are the imported facts. Phishing internet web page detection is carried out with the aid of four types of URL functionality training: domain-based, address-based, abnormal-based, HTML, and JavaScript features. With the use of processed data, these URL traits are extracted, and values are created for each URL trait. A machine learning algorithm that calculates the range cost and consequently the edge value for the properties of the URL completes the analysis of the URL. Then it is divided into phishing and official URL categories. The characteristic values are used to determine the variety value and the edge co stand are calculated by extracting features from phishing websites.

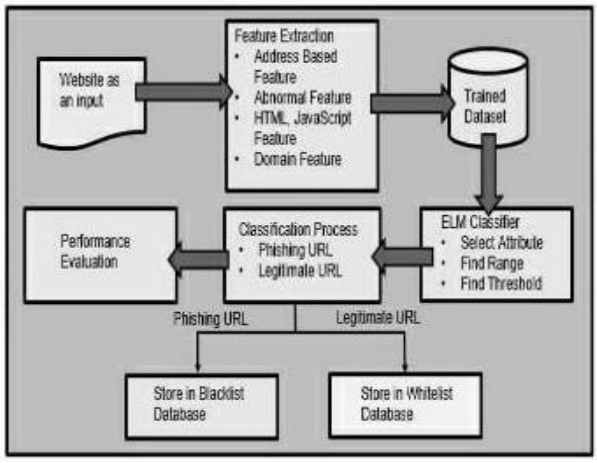

Fig -1: System Architecture

CONCLUSIONS

Websites can be used to establish a system for everything from record entry to scientific applications. The statistics used as input may be processed, and the output of the processing may be information. Websites are now used in a wide range of disciplines, including medicine, technology, business, education, and economics. Due to its wide spread use, hackers may also utilize it as a tool for evil ends. A rogue object makes an apparent phishing attack. Many research contributions provide completely original methods, approaches for locating phishing URLs, and applications of those methodologies. The equipment's goal is to create a category for the classification of phishing as one of the attack types that cyber threats use.

REFERENCES

1. Operantly, Deepak Upadhyay, Deepak Upadhyay, IJERT, ISSN:2278-0181,04, April-2020
2. Meenu, Sunila Gadara, Phishing Detection using Machine Learning Techniques, IJEAT, ISSN: 2249– 8958, 2, December, 2019.
3. Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick,
4. Lavanya Basilicata, Ravali Reddy Guder, Siri Chandana Gotha, IJITEE, ISSN: 2278- 3075, June 2019.
5. Ankit Kumar Jainand Gotha EURASIP Journal on Information Security (2016).
6. 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.
7. E. Jakobsson, and E. Myers, Phishing and Counter-Measures: Understanding the Increasing Problem of Electronic Identity Theft. Wiley, 2006, pp.2–3.
8. L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen," Detecting phishing web sites: A heuristic URL-based approach," in 2013 International Conference on Advanced Technologies for Communications (ATC 2013), 2013, pp. 597-602.
9. Z. Zhang, Q. He, and B. Wang," A Novel Multi-Layer Heuristic Model for Anti-Phishing," New York, NY, USA, 2017, p. 21:1-21:6.
10. N.," A and A. Rungs Awang," Web Phishing Detection Using Classifier Ensemble," New York, NY, USA, 2010, pp. 210-215.
11. Basnet, R., Mukkamala, S., Sung, A.H. "Detection of phishing attacks: A machine learning approach." Studies in Fuzziness and Soft Computing, 226:373–383, 2014.
12. Lakshmi, V. Santhana, and M. S. Vijaya. "Efficient prediction of phishing websites using supervised learning algorithms." Procedia Engineering 30 (2012): 798-805.