International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.053 ISSN: 2581-4621

Fraud Attack Detection – Through the Login URL Using Phishing Methodology

P.Bhanu Prakasha[1], D.Likhitha[2], K.Anitha[3], S.Mahammad Ashfaq[4], S.Leeladhar[5]
Assistant Professor, Student, Department of CSE, AITS, Tirupati

Abstract

Phishing is a social engineering cyberattack where criminals deceive users to obtain their credentials through a login form that submits the data to a malicious server. In this paper, we compare machine learning and deep learning techniques to present a method capable of detecting phishing websites through URL analysis. In most current state-of-the-art solutions dealing with phishing detection, the legitimate class is made up of homepages without including login forms. On the contrary, we use URLs from the login page in both classes because we consider it is much more representative of a real case scenario and we demonstrate that existing techniques obtain a high false positive rate when tested with URLs from legitimate login pages. Additionally, we use datasets from different years to show how models decrease their accuracy over time by training a base model with old datasets and testing it with recent URLs. Also, we perform a frequency analysis over current phishing domains to identify different techniques carried out by phishers in their campaigns. To prove these statements, we have created a new dataset named Phishing Index Login URL (PILU-90K), which is composed of 60K legitimate URLs, including index and login websites, and 30K phishing URLs.

Keywords: Phishing, Term Frequency, Machine learning and Deep learning

1. Introduction

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

2. Related Work

Phishing attacks are categorized according to Phisher’s mechanism for trapping alleged users.
Several forms of these attacks are keyloggers, DNS toxicity, Etc., . Th initiation processes in social engineering include online blogs, short message services (SMS), social media platforms that use web 2.0 services, such as Facebook and Twitter, file-sharing services for peers, Voice over IP (VoIP) systems where the attackers use caller spoofing IDs . Each form of phishing has a little difference in how the process is carried out in order to defraud the unsuspecting consumer. E-mail phishing attacks occur when an attacker sends an e-mail with a link to potential users to direct them to phishing websites

3.Classification of phishing attack techniques

Phishing websites are challenging to an organization and individual due to its similarities with the legitimate websites . Technical subterfuge refers to the attacks include Keylogging, DNS poisoning, and Malwares. In these attacks, attacker intends to gain the access through a tool /technique. On the one hand, users believe the network and on the other hand, the network is compromised by the attackers. Social engineering attacks include Spear phishing, Whaling, SMS, Vishing, and mobile applications. In these attacks, attackers focus on the group of people or an organization and trick them to use the phishing URL . Apart from these attacks, many new attacks are emerging exponentially as the technology evolves constantly.

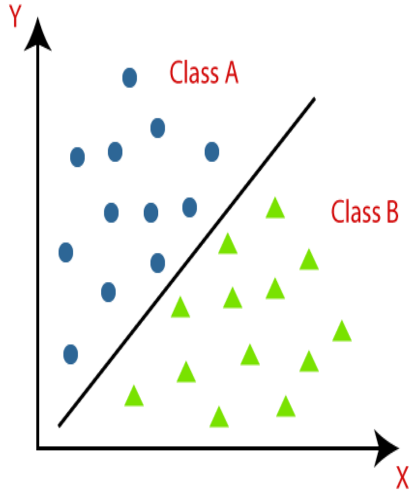


Figure:classification

4.**Phishing detection approaches**

phishing on the server side are better than phishing prevention strategies and user training systems. These systems can be used either via a we browser on the client or through specific host-site software . Heuristic and ML based approach is based on supervised and unsupervised learning techniques. It requires features or labels for learning an environment to make a prediction. Proactive phishing URL detection is similar to ML approach. However, URLs are processed and support a system to predict a URL as a legitimate or malicious. Blacklist and Whitelist approaches are the traditional methods to identify the phishing sites . The exponential growth of web domains reduces the perfor- mance of the traditional method**.** The existing methods rely on new internet users to a minimum.

Once they identify phishing website, the site is not accessible, or the user is informed of the probability that the website is not genuine. This approach requires minimum user training and requires no modifications to existing website authentication systems. The performance of the detection systems is calculated according to the following:

Number of True Positives (TP): The total number of malicious websites. Number of True Negatives (TN): The total number of legitimate websites.

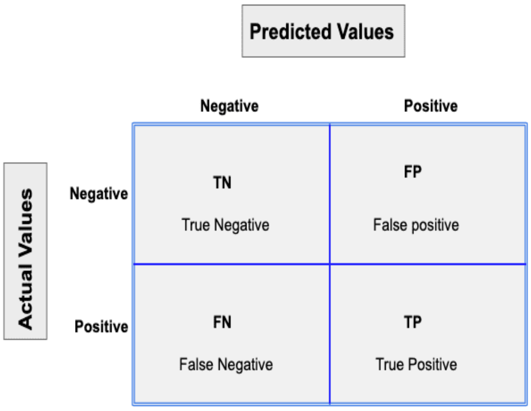
Number of False Positives (FP): The total number of incorrect predictions of legitimate websites as a malicious website. Number of False Negatives (FN): The total number of incorrect predictions of malicious websites as a legitimate website.

Evaluation Metrics

Evaluation metrics are used to measure the performance of a model or algorithm in terms of its accuracy, precision, recall, F1-score, and other relevant measures. It is crucial to evaluate a model to ensure that it is performing well on the given task andto identify areas for improvement.

Confusion Matrix

A confusion matrix is a table used to evaluate theperformance of a classification model by comparing the predicted class labels with the actual class labels ofa set of test data. The predicted values are usually shown in columns, while the actual values are shown in rows.



		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN True Negative	FP False positive
	Positive	FN False Negative	TP True Positive

Fig 2.5.1 : Confusion Matrix

Each prediction can be one of the four outcomes, basedon how it matches up to the actual value:

- **True Positive (TP):** Predicted True and True inreality.
- **True Negative (TN):** Predicted False andFalse in reality.
- **False Positive (FP):** Predicted True and Falsein reality.
- **False Negative (FN):** Predicted False and Truein reality.

There is also a list of rates that are often computedfrom a confusion matrix for a binary classifier:

- **Accuracy:** It is a commonly used evaluation metric for binary classifiers when the classes are roughly equal in size. It measures the proportion of true positives and true negatives among all the samples. It is calculated as follows:

Accuracy = (TP+TN)/Total Predictions

Accuracy can be misleading if the classes are imbalanced, meaning that one class has significantly more samples than the other. In such cases, a classifier that always predicts the majority class will have a high accuracy but may not be useful in practical applications. In such cases, other evaluation metrics such as precision, recall, and F1 score may be more appropriate.

- **Precision:** The precision metric is used to overcome the limitation of Accuracy. The precision determines the proportion of positive prediction that was actually correct. It can be calculated as the True Positive or predictions that are actually true to the total positive predictions

Precision = TP/(TP+FP)

➢ **Recall or Sensitivity:** It is also similar to the Precision metric; however, it aims to calculate the proportion of actual positive that was identified incorrectly.

Recall = TP/(TP+FN)

➢ **F1- Scores**: F1 Score is a metric to evaluate a binary classification model on the basis of predictions that are made for the positive class**.** F1 Score is useful in cases where we have imbalanced classes, where one class is muchmore frequent than the other. F1 Score takes into account both precision and recall, which are important for evaluating the performance of a binary classification model on imbalanced data.

F1 Score = (2*(precision*recall)) / (precision+recall)

3. DESIGN

The module simply means the software components that are been created by dividing the software. The software is divided into various components that work together to form a single functioning item but sometimes they can perform as a complete function if not connected with each other. In this proposed system, there are two modules. They are:

1. Service Provider
2. Remote User

1. Service Provider

The Functionalities of this Modules are as Follows:

1. Login
2. Browse Website URL Datasets
3. Train Datasets
4. Generate Train and Tested Accuracy
5. Provide Prediction of Website URL Type
6. Generate Website URL Type Ratio
7. Download Trained Datasets
8. Logout

2. Remote User

The Functionalities of this Modules are as Follows:

1. Register
2. Login
3. Predict Website URL Type
4. View Profile
5. Logout

3. Activity Diagram for Service Provider

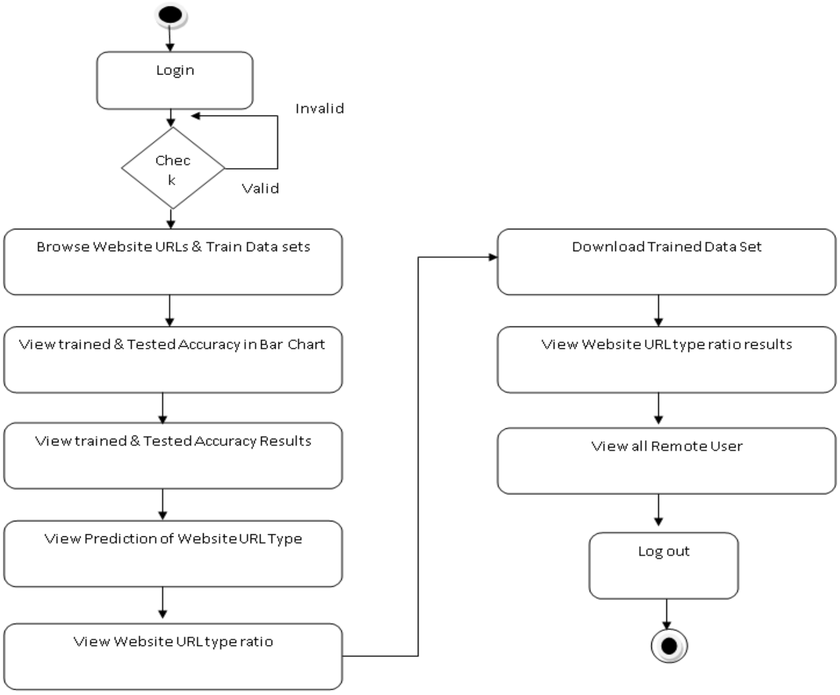

Login
Invalid
Chec
k
Valid
Browse Website URLs & Train Data sets
View trained & Tested Accuracy in Bar Chart
View trained & Tested Accuracy Results
View Prediction of Website URL Type
View Website URL type ratio
Download Trained Data Set
View Website URL type ratio results
View all Remote User
Log out

Fig 3.3.1: Activity Diagram for Service Provider

4. Activity Diagram for Remote User

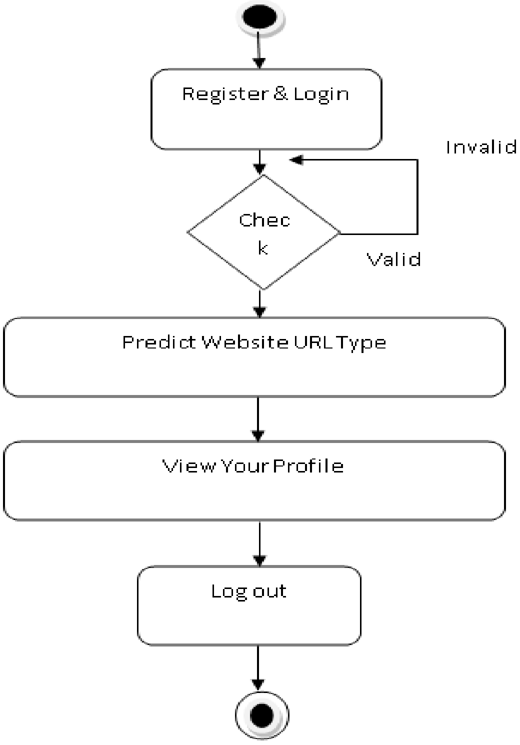


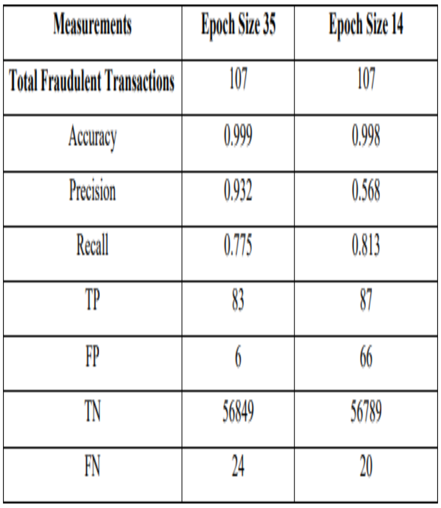
Fig 3.4.1: Activity Diagram for Remote User

4. EXPERIMENTAL RESULT

Therefore, a machine learning model can be created with the help of all the algorithms discussed above, and for testing and training, the machine learning model and 80% of the dataset were used for training and 20% for testing. Light GBM, Random Forest, Decision Tree, Logistic Regression, and SVM are the machine learning methods used to analyze to determine whether such a URL is fraudulent or not. As a result of fitting the dataset to all algorithms, Light GBM produced good results and all.

1. The Accuracy of machine Learning Algorithms

The training and validation accuracy of baseline CNN and proposed CNN algorithms are shown in the Table below.

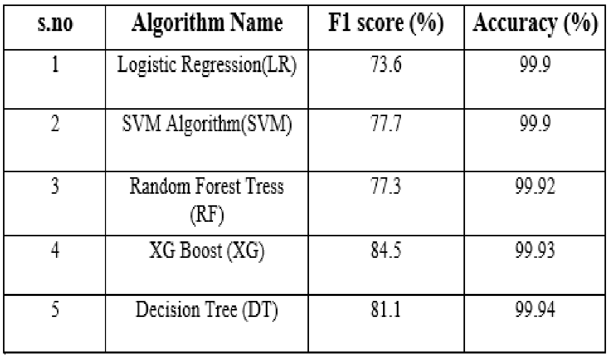


Measurements	Epoch Size 35	Epoch Size 14
Total Fraudulent Transactions	107	107
Accuracy	0.999	0.998
Precision	0.932	0.568
Recall	0.775	0.813
TP	83	87
FP	6	66
TN	56849	56789
FN	24	20

Fig 4.1.1: Deep Learning Accuracy

2. Accuracy and F1 Score of ML algorithms

We construct five different types of classification models during this stage. We could solve categorization issues with a variety of alternative models. The models.The below table represents the accuracy and f1 score ofthe ML algorithms.



s.no	Algorithm Name	F1 score (%)	Accuracy (%)
1	Logistic Regression(LR)	73.6	99.9
2	SVM Algorithm(SVM)	77.7	99.9
3	Random Forest Tress (RF)	77.3	99.92
4	XG Boost (XG)	84.5	99.93
5	Decision Tree (DT)	81.1	99.94

Fig 4.2.1: F1 score and Accuracy of ML models

3. Machine Learning Algorithms Comparative Analysis

The below graph represents the comparative analysis of the f1 score and accuracy of Machine Learning algorithms. The below graph represents the comparative analysis of the f1 score and accuracy of Machine Learning algorithms.

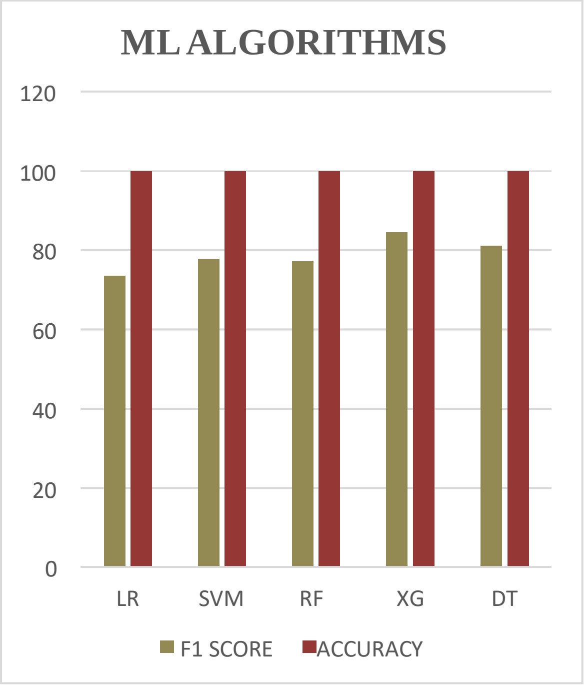


Fig 4.3.1: Graph of comparison of ML Algorithms

5. Conclusion:

Internet users are at serious risk from phishing. Web security faces a significant issue as a result of the rapid development and spread of phishing methods. It is difficult to identify a fraudulent URL, but machine learning methods can help. In this study, using the Random Forest, Decision Tree, Light GBM, Logistic Regression, and Support Vector Machine, we investigated the linguistic and domain-based properties of the URL and created a machine learning model.

6. References:

[1] Statista. (2020). Adoption Rate of Emerging Technologies in Organizations Worldwide as of 2020. Accessed: Sep. 12, 2021. [Online]. Available: https://www.statista.com/statistics/661164/worldwide-cio-surveyoperati%onal-priorities/

[2] R. De', N. Pandey, and A. Pal, ''Impact of digital surge during COVID- 19 Pandemic: A viewpoint on research and practice,'' Int. J. Inf. Manage.,vol. 55, Dec. 2020, Art. No. 102171.

[3] P. Patel, D. M. Sarno, J. E. Lewis, M. Shoss, M. B. Neider, and C. J. Bohil, ''Perceptual representation of spam and phishing emails,'' Appl. Cognit. Psychol., vol. 33, no. 6, pp. 1296–1304, Nov. 2019.

[4] J. A. Chaudhry, S. A. Chaudhry, and R. G. Rittenhouse, ''Phishing attacks And defenses,'' Int. J. Secur. Appl., vol. 10, no. 1, pp. 247–256, 2016.

[5] M. Hijji and G. Alam, ''A multivocal literature review on growing social Engineering based cyber-attacks/threats during the COVID-19 pandemic: Challenges and prospective solutions,'' IEEE Access, vol. 9, pp. 7152– 7169, 2021

[6] A. Alzahrani, ''Coronavirus social engineering attacks: Issues and Recommendations,'' Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 5, pp. 154–161, 2020.

[7] Phishing Activity Trends Report 3Q, Anti-Phishing Working Group, International, 2017. Accessed: Sep. 12, 2021.

[8] Phishing Activity Trends Report 1Q, Anti-Phishing Working Group, International, 2021. Accessed: Sep. 14, 2021.

[9] R. Chen, J. Gaia, and H. R. Rao, ''An examination of the effect of recent Phishing encounters on phishing susceptibility,'' Decis. Support Syst.,vol. 133, Jun. 2020, Art. No. 113287.