Phishing URL Detection Using Machine Learning

G Varun
Department of Computer Science and Engineering
Bapatla Engineering College
(Autonomous)
(Affiliated to Acharya Nagarjuna University)
Bapatla, AP, India
varun.techmail@gmail.com

CH. Naveen
Department of Computer Science and Engineering
Bapatla Engineering College
(Autonomous)
(Affiliated to Acharya Nagarjuna University)
Bapatla, AP, India
cnaveen508@gmail.com

E. Venkatesh
Department of Computer Science and Engineering
Bapatla Engineering College
(Autonomous)
(Affiliated to Acharya Nagarjuna University)
Bapatla, AP, India
venkateshedubilli2001@gmail.com

B. Karthik
Department of Computer Science and Engineering
Bapatla Engineering College
(Autonomous)
(Affiliated to Acharya Nagarjuna University)
Bapatla, AP, India
karthikballi0001@gmail.com

Abstract - *Phishing internet sites are one of the internet protections issues that focus on human vulnerabilities rather than software program vulnerabilities. It can be defined because of the process of attracting online users to gain their touchy facts which include usernames and passwords. In this paper, we provide a sensible machine for detecting phishing URLs. The system is based on a device gaining knowledge of approach, particularly supervised mastering. We have decided on the Gradient Boost method because of its true overall performance in classification. Our focus is to pursue a better overall performance classifier by analyzing the features of phishing websites and choose the better aggregate of them to train the classifier.*

***Key Words*: Gradient Boost, Classifier, Features, Phishing, Train, Accuracy.**

1. INTRODUCTION

Phishing Uniform Resource Locator (URL) host unsolicited information and attackers use these URLs as one of a primary tool to carry out cyber-attacks. E-mail and social media resources such as Facebook, Twitter, WhatsApp, Orkut, etc. are the most commonly used applications to spread malicious URLs. They host unsolicited information on the web page. Whenever an unsuspecting user visits that website unknowingly through the URL, the host may get compromised, making them victims of various types of frauds including malware installation, data, and identity theft. Every year, malicious URLs have been causing billions of dollars' worth of losses. These factors force the development of efficient techniques to detect malicious URLs promptly and give an alert to the network administrator. Most of the commercial products exist in markets are based on the blacklisting method. This method relies on a database that contains a list of malicious URLs.

The blacklists are continually updated by the anti-virus group through scanning and crowdsourcing solutions. The blacklisting method can be used to detect the malicious URLs which are already present in the database. But they completely fail to detect the variants of the existing phishing URLs or entirely new phishing URLs. The current innovative approach to classifying phishing activity is the utilization of URL and web content features with a machine learning approach to improve detection accuracy and performance.

2. LITERATURE SURVEY

MAHAJAN MAYURI VILAS, KAKADE PRACHI GHANSHAMSAWANT, PURVA JAYPRALASH and PAWAR SHILA [1] in their paper "Detection of Phishing Website Using Machine Learning Approach", the goal of the study is to carry out ELM employing 30 different primary components that are characterized using ML. To prevent being discovered, most phishing URLs use HTTPS. Website phishing can be identified in three different ways. The first method evaluates several URL components; the second method assesses a website's authority, determines if it has been introduced or not, and determines who is in charge of it; the third method verifies a website's veracity.

In [2] MALAK ALJABRI and SAMIHA MIRZA proposed a paper "Phishing Attacks Detection using Machine Learning and Deep Learning Models" In this study, the highest correlated features from two distinct datasets were chosen. These features combined content-based, URL and domain-based features. Then, a comparison of the performance of a number of ML models was carried out. The results also sought to pinpoint the top characteristics that aid the algorithm in spotting phishing websites. The Random Forest (RF) method produced the best classification results for both datasets.

ADARSH MANDADI and SAIKIRAN BOPPANA in their study [3], the user-received URLs will be entered to the machine learning model, which will then process the input and report the results, indicating whether the URLs are phishing or not. SVM, Neural Networks, Random Forest, Decision Tree, XG boost, and other machine learning algorithms can all be used to categorize these URLs. The suggested method uses the Random Forest and Decision Tree classifiers. With an accuracy of 87.0% and 82.4% for Random Forest and decision tree classifiers, respectively, the suggested technique successfully distinguished between Phishing and Legitimate URLs.

In [4] HEMALI SAMPAT, MANISHA SAHARKAR, AJAY PANDEY AND HEZAL LOPES have proposed a system for Detection of Phishing Websites using Machine learning. Their proposed method uses both Classification and Association algorithms to optimise the system, making it faster and more effective than the current approach. The proposed system's inaccuracy rate is reduced by 30% by combining these two algorithms with the WHOIS protocol, making it an effective technique to identify phishing websites.

SAFA ALREFAAI, GHINA ÖZDEMIR and AFNAN MOHAMED [5] used Machine Learning is being used to detect phishing websites. They used Kaggle data with 86 features and 11,430 total URLs, half of which are phishing and half of which are legitimate. They trained their data using Decision Tree (DT), Random Forest (RF), XGBoost, Multilayer Perceptrons, K-Nearest Neighbors, Naive Bayes, AdaBoost, and Gradient Boosting, with X G Boost.

In [6], SUNDARA PANDIYAN S, PRABHA SELVARAJ, VIJAY KUMAR BURUGARI, JULIAN BENADIT P and KANMANI P employed a wide range of techniques, including Decision Tree, Random Forest, Multi-Layer Perceptron, XG Boost Classifier, SVM, Light BGM Classifier, and Cat Boost Classifier. Our team discovered that Light GBM had the best precision, with an average accuracy of about 85.5%. One class SVM, on the other hand, has the lowest precision, at about 79.6%.

3. PROPOSED SYSTEM

Each type of phishing differs slightly in how the procedure is carried out to deceive the unwary customer. When a hacker sends a potential user an email with a link that takes them to phishing websites, this is known as an email phishing attack.

We use different machine learning models trained over features like if URL contains @, if it has double slash redirecting, page rank of the URL, number of external links embedded on the webpage, etc.

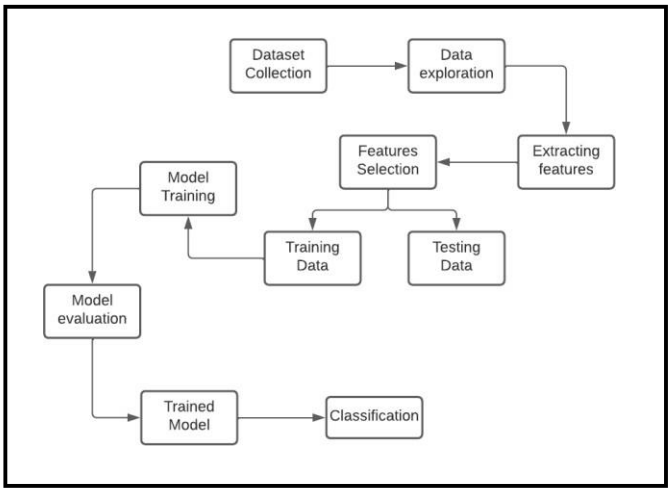

Fig -3.1: Flowchart

4. METHODOLOGY

Data collection, cleaning, and consolidation into a single file or data table are all steps in the process of data preparation, which is done largely for analytical purposes as shown in Fig 3.1. The following are the main activities we utilise for data preparation: data reduction, data transformation, dataintegration, and data discretization.

The crucial libraries, including XGBoost, Numpy, Matplotlib, Pandas, and Numpy, are loaded first. The dataset from Kaggle is then imported after the libraries have been imported. We divided the dataset into training and testing sets after importing it using train test split from sklearn. 20% of the dataset is used for testing, while 80% is used for the training set.

We have set up a model that uses eight distinct algorithms, including Logistic Regression, KNeighborsClassifier, Random Forest, Decision Tree, Support Vector Machine, Naive Bayes Classifier, Gradient Boost and XGBoost, to compare the accuracy of various techniques. We work on model fitting, which makes predictions, to achieve the desired result, and then we work on model evaluation. For this evaluation, test data is utilized. We compare the accuracy of each method using several algorithms, such as confusion matrix, to obtain the best result.

5. DATASET AND FEATURES

The Dataset we used in our project contains 32 columns with 30 optimized features of URL. It contains three different labels where 1 means legitimate, 0 means suspicious and-1 means phishing. The data set borrowed from https://www.kaggle.com/eswarchandt/phishingwebsite-detector. A collection of website URLs for 11000+ websites. Each sample has 30 website parameters and a class label identifying it as a phishing website or not (1 or -1). The

overview of this dataset is, it has 11054 samples with 32 features.

The **Table 5.1** neatly arranges 30 URL features into two columns for efficient presentation.

Table 5.1 ***Features***

UsingIP	LongURL
Symbol@	Redirecting//
SubDomains	HTTPS
Favicon	NonStdPort
RequestURL	AnchorURL
ServerFormHandler	InfoEmail
WebsiteForwarding	StatusBarCust
UsingPopupWindow	IframeRedirection
DNSRecording	WebsiteTraffic
GoogleIndex	LinksPointingToPage
ShortURL	DomainRegLen
PrefixSuffix-	HTTPSDomainURL
LinksInScriptTags	AbnormalURL
DisableRightClick	AgeofDomain
PageRank	StatsReport

6. RESULT

To acquire useful results, we've compared a number of algorithms. There are numerous algorithms that can be used to identify phishing websites; however, after reviewing numerous research articles, we settled on eight algorithms to test the model.

6.1 Model Comparison

Table 6.1 ***Model Accuracies***

	Model	Accuracy
1	Logistic Regression	0.934
2	Decision Tree	0.961
3	Random Forest	0.966
4	KNeighborsClassifier	0.956
5	XGBoost	0.549
6	Naive Bayes Classifier	0.605
7	Support Vector Machine	0.964
8	Gradient Boost	0.974

6.2 Model Output

Classification report of the Gradient Boost Classifier

	precision	recall	f1-score	support
-1	0.99	0.96	0.97	976
1	0.97	0.99	0.98	1235
accuracy			0.97	2211
macro avg	0.98	0.97	0.97	2211
weighted avg	0.97	0.97	0.97	2211

*Figure6. 1-****Classification report***

XGradient Boost is a Machine Learning technique used for both classification and regression tasks. It is an ensemble learning method that builds a strong model by combining the predictions from multiple weaker models. In Gradient Boosting, the model is built in a stage-wise manner, where each new model attempts to correct the errors of the previous models.

One of the key benefits of Gradient Boosting is its ability to capture complex patterns in the data, making it highly effective for a wide range of tasks. Additionally, it offers several hyperparameters that can be fine-tuned to optimize performance, including the learning rate, the depth of the trees, and the number of iterations.

We trained our data using Logistic Regression, KNeighborsClassifier, Random Forest, Decision Tree, Gradient Boost and XGBoost with Gradient Boost achieving the highest accuracy of 97.

7. CONCLUSIONS

To the best of our knowledge, this study is the first analysisto include the findings of all other studies into the detection of phishing websites using machine learning techniques. The suggested research on phishing uses a categorical paradigm, where phishing websites are thought to automatically classify websites into a given range of sophisticated values depending on a variety of factors and the grandeur variable.

The website functionality is used by ML-based phishing approaches to collect information that could be used to classify websites for the purpose of identifying phishing sites. Developing focused anti-phishing approaches and methods as well as minimizing their inconvenience are twoways to prevent phishing.

We achieved 97% detection accuracy using Gradient Boost algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data.

8. REFERENCES

[1] M. M. Vilas, K. P. Ghansham, S. P. Jaypralash and P. Shila, "Detection of Phishing Website Using Machine Learning Approach," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru,India, 2019, pp. 384-389, doi: 10.1109/ICEECCOT46775.2019.9114695.

[2] M. Aljabri and S. Mirza, "Phishing Attacks Detection using Machine Learning and Deep Learning Models," 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA),Riyadh, Saudi Arabia, 2022, pp. 175-180, doi: 10.1109/CDMA54072.2022.00034.

[3] A. Mandadi, S. Boppana, V. Ravella and R. Kavitha, "Phishing Website Detection Using Machine Le arning," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India,2022, pp. 1-4, doi: 10.1109/I2CT54291.2022.9824801.

[4] Hemali Sampat, Manisha Saharkar, Ajay Pandey and Hezal Lopes, “Detection of Phishing Website Using Machine Learning,” 2018 International Research Journal of Engineering and Technology (IRJET),2018, e-ISSN: 2395-0056, p-ISSN: 2395-0072.

[5] S. Alrefaai, G. Özdemir and A. Mohamed, "Detecting Phishing Websites Using Machine Learning," 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2022, pp. 1-6, doi: 10.1109/HORA55278.2022.9799917.

[6] Sundara Pandiyan S, Prabha Selvaraj, Vijay Kumar Burugari, Julian Benadit P, Kanmani P, Phishing attack detection using Machine Learning, Measurement: Sensor Volume 24,2022,100476,ISSN 2665-9174