International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.104 ISSN: 2581-4621

Comparative Insight Into Machine Learning Strategies For Phishing Website Identification

Shaik Althaf Hussain[1], K.Muddu Swamy[2]
[1]*MCA Student, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*
[2]*Assistant Professor, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*

Abstract
Phishing websites continue to pose a significant threat to individuals and organizations, leading to financial losses, data breaches, and compromised security. This paper presents a comprehensive approach to prevent phishing website attacks through the development of an advanced detection and mitigation system. Leveraging machine learning algorithms, web scraping techniques, and real-time analysis, the proposed system aims to identify and neutralize phishing websites proactively, thereby safeguarding users' sensitive information and preserving the integrity of online transactions. By addressing the shortcomings of existing preventive measures, this system offers a robust solution to combat the growing menace of phishing websites and enhance cybersecurity in the digital landscape.
Keywords: phishing, cyber security, website, ML

Introduction

In the once decades, the operation of internet has been increased extensively and makes our live simple, easy and transforms our lives. It plays a major part in areas of communication, education, business conditioning and commerce. A lot of useful data, information and data can be attained from the internet for particular, organizational, profitable and social development. The internet makes it easy to give numerous services through online and enables us to pierce colorful information at any time, from anywhere around the world. Phishing is the act of transferring a indistinguishable dispatch, dispatches or vicious websites to trick the philanthropist / internet druggies into discovering delicate particular information similar as personal identification number (PIN) and word of bank account, credit card information, date of birth or social security figures. Phishing assaults affect hundreds of thousands of internet druggies across the globe. Individualizes and associations have lost a huge sum of plutocrat and private information through Phishing attacks. Detecting the phishing attack proves to be a challenging task. Tis attack may take a sophisticated form and fool even the savviest users: such as substituting a few characters of the URL with alike unicode characters. By cons, it can come in sloppy forms, as the use of an IP address instead of the domain name. Nonetheless, in the literature, several works tackled the phishing attack detection challenge while using artifcial intelligence and data mining techniques [5–9] achieving some satisfying recognition rate peaking at 99.62%. However those systems are not optimal to smartphones and other embed devices because of their complex computing and their high battery usage, since they require as entry complete HTML pages or at least HTML links, tags and webpage JavaScript elements some of those systems uses image processing to achieve the recognition. Opposite to our recognition system since it is a less greedy in terms of CPU and memory unlike other proposed systems as it needs only six features completely extracted from the URL as input. In this paper, after a summary of this feld key researches, we will detail the characteristics of the URL that our system uses to do the recognition. Otherwise we will describe our recognition system, next in the practical part we will test the proposed system while presenting the results obtained. Last but not least we will enumerate the implications and advantages that our system brings as a solution to the phishing attack.

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.104 ISSN: 2581-4621

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, KNearestNeighbors, 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 Perceptrons, 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%.

[7] Andrei Butnaru et al., used a supervised Machine Learning algorithm to block phishing attacks, based on novel mixture phishing attacks and compare with Google Safe browsers.

[8] Valid Shahrivari et al., proposed a one of the most successful techniques for identifying these malicious works is Machine Learning. It is because of most Phishing attacks have same features which can be noticed by Machine learning techniques. In this many machine learning-based classifiers are used for forecasting the phishing websites. The main advantage of machine learning is the ability to create flexible models for specific tasks like phishing detection. Since phishing is a classification problem, Machine learning models can be used as a forceful tool.

[9] Ammara Zamir et al., proposed a framework for identifying phishing websites using heaping model. Information gain, gain ratio, Relief-F, and recursive feature elimination (RFE) are some of the feature selection algorithms that can be used to analyse Phishing characteristics. The greatest and weakest traits are combined to create two features. Bagging is used in principal component

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.104 ISSN: 2581-4621

analysis using several Machine learning algorithms, including random forest [RF] and neural network [NN]. Two heaping representations heaping1 (RF + NN + Bagging) and heaping2 (kNN + RF + Bagging) are applied by merging highest scoring classifiers to progress classification accuracy.
[10] Nguyet Quang Do, Ali Selamat et al., conducted a study on phishing detection and proposed a four different deep learning technique, includes deep neural network (DNN), convolution neural networks (CNN), Long Short-term memory (LSTM), and gated recurrent unit (GRU). To analysebehaviour of these deep learning architectures, extensive experiments were carried out to examine the impact of parameter tuning on the performance accuracy of the deep learning models. In which each model shows different accuracies from different models.
[11] Ashit Kumar Dutta proposed a URL detection procedure based on Machine Learning methods. An RNN is used for identifying the phishing URL. It is evaluated with 7900 malicious and 5800 genuine sites, respectively. The outcome of this method shows a good concert compare to recent tactics.

Existing System

Phishing is an internet scam in which an attacker sends out fake messages that look to come from a trusted source. A URL or file will be included in the mail, which when clicked will steal personal information or infect a computer with a virus. Traditionally, phishing attempts were carried out through wide-scale spam campaigns that targeted broad groups of people indiscriminately. The goal was to get as many people to click on a link or open an infected file as possible. There are various approaches to detect this type of attack. One of the approaches is machine learning. The URL's received by the user will be given input to the machine learning model then the algorithm will process the input and display the output whether it is phishing or legitimate. There are various ML algorithms like SVM, Neural Networks, Random Forest, Decision Tree, XG boost etc. that can be used to classify these URLs. The proposed approach deals with the Random Forest, Decision Tree classifiers.

Proposed System

Phishing attacks have evolved in terms of sophistication and have increased in sheer number in recent years. This has led to corresponding developments in the methods used to evade the detection of phishing attacks, which pose daunting challenges to the privacy and security of the users of smart systems. This study uses LightGBM and features of the domain name to propose a machine-learning-based method to identify phishing websites and maintain the security of smart systems. Domain name features, often known as symmetry, are the property wherein multiple domain-name-generation algorithms remain constant. The proposed model of detection is first used to extract features of the domain name of the given website, including character-level features and information on the domain name. The features are filtered to improve the model's accuracy and are subsequently used for classification. The results of experimental comparisons showed that the proposed model of detection, which integrates two types of features for training, significantly outperforms the model that uses a single type of feature. The proposed method also has a higher detection accuracy than other methods and is suitable for the real-time detection of many phishing websites.

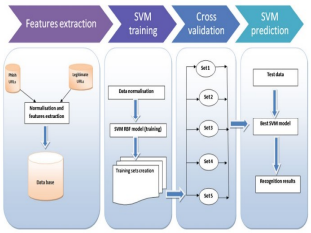


Fig.1. Phishing website process.

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.104 ISSN: 2581-4621

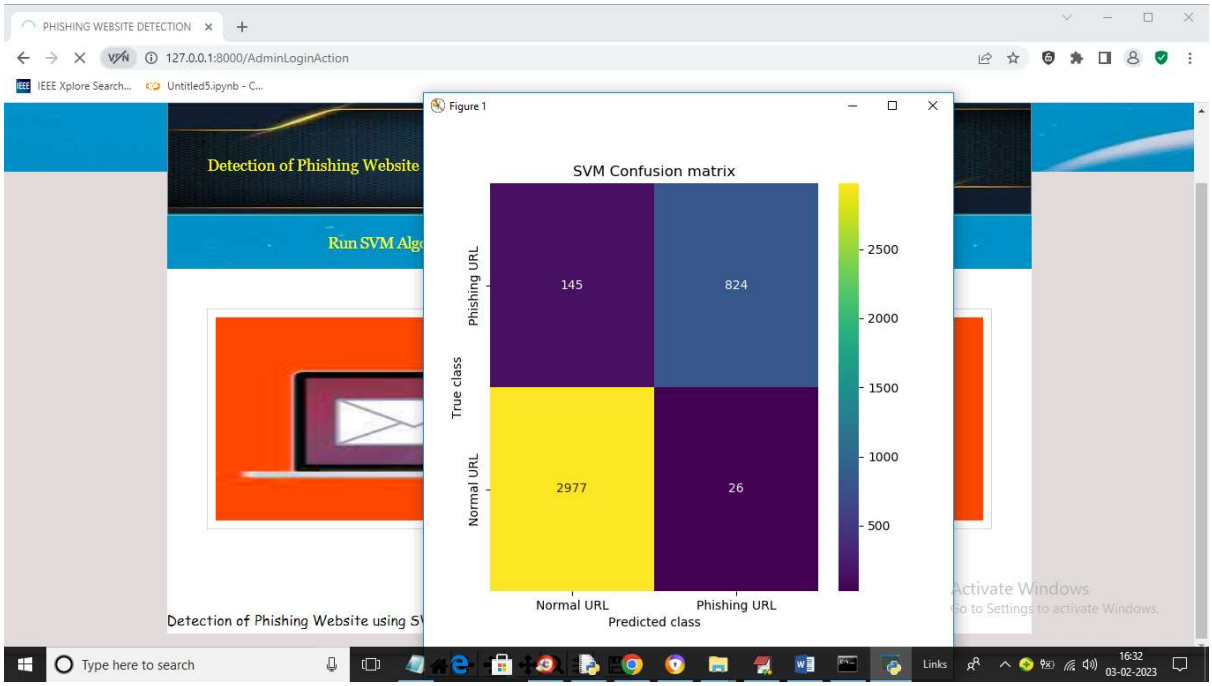
Methodology

In this segment we going to learn about the classifiers used in machine learning to envisage phishing. Here we intend to explain our proposed methodology to detect phishing website. In this we divided into 2 parts one for classifiers and another to explain our proposed system.

Machine learning classifiers and methods to perceive the phishing website Distinguishing and recognizing phishing websites is really an intricate and energetic problem. Machine learning has been extensively used in numerous areas to produce automated results. Phishing attacks can take numerous forms, including dispatch, website, malware, and voice. This paper focuses on detecting website phishing (URL) using the Hybrid Algorithm Approach. It is a mix of different classifiers that work together to improve the system's accuracy and estimate rate. Depending on the application and the nature of the dataset used we can use any classification algorithms. As there are various applications, we cannot discriminate which of the algorithms are superior or not.

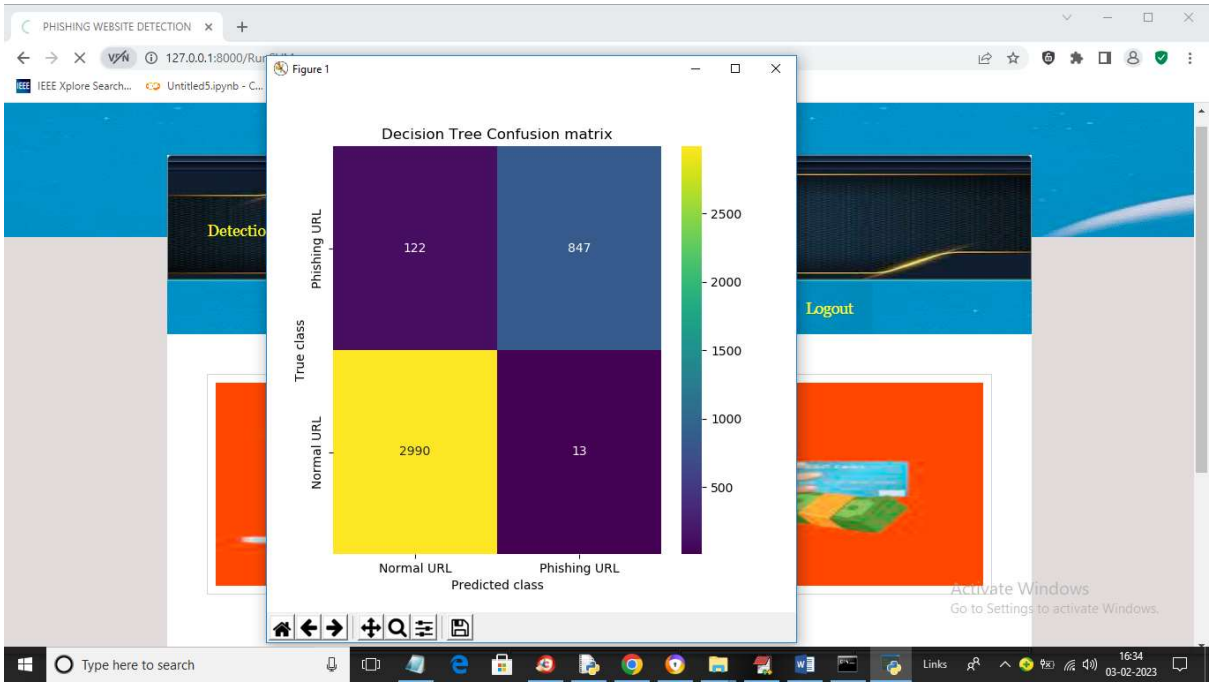
Support Vector Machine (SVM): This is also one of the supervised and simple to use classification algorithms. It can be used in both classification and regression applications; however, classification applications are preferred. SVMs differ from other classification algorithms in that they employ the distance between the nearest data points of all classes to determine the decision boundary. The maximum margin classifier or maximum margin hyper plane is the decision boundary created by SVMs. The classification is based on the differences between the classes, which are data set points in various planes.

Results

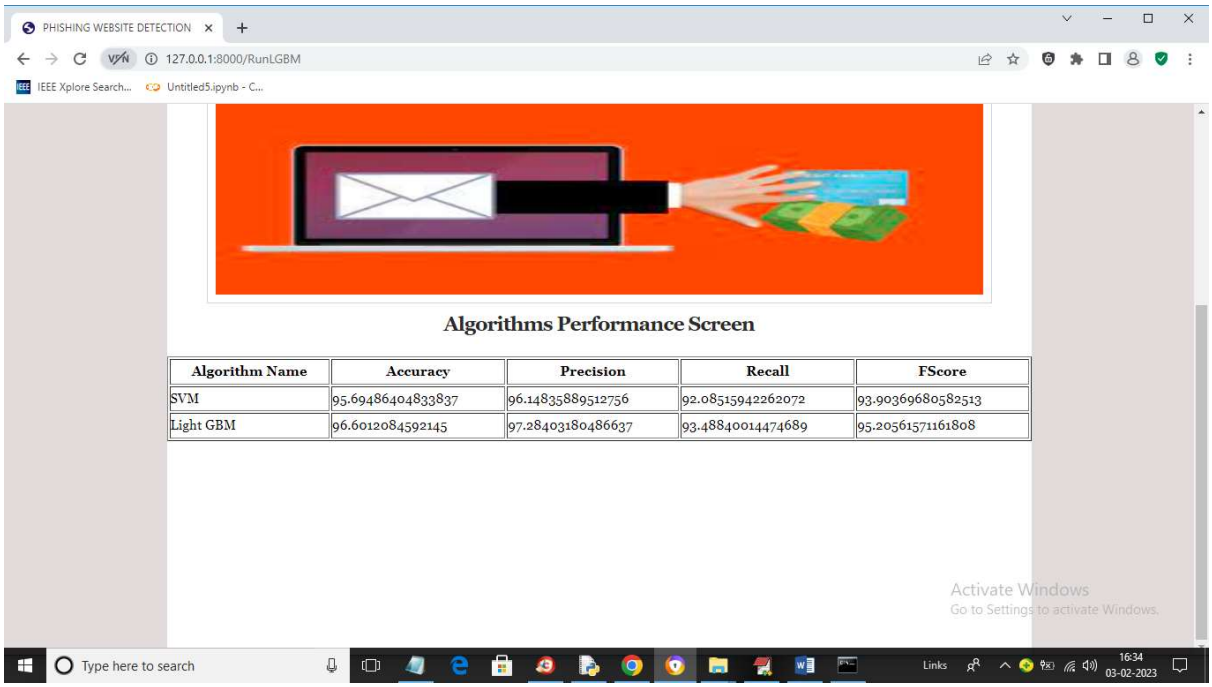


In above screen we can see SVM confusion matrix where x-axis represents predicted class and y-axis represents TRUE class and we can see SVM predict 2977 records correctly as NORMAL and only 145 are incorrect prediction and it predict 824 records as PHISHING URL and only 26 are incorrect prediction and now close above graph to get below output

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.104 ISSN: 2581-4621



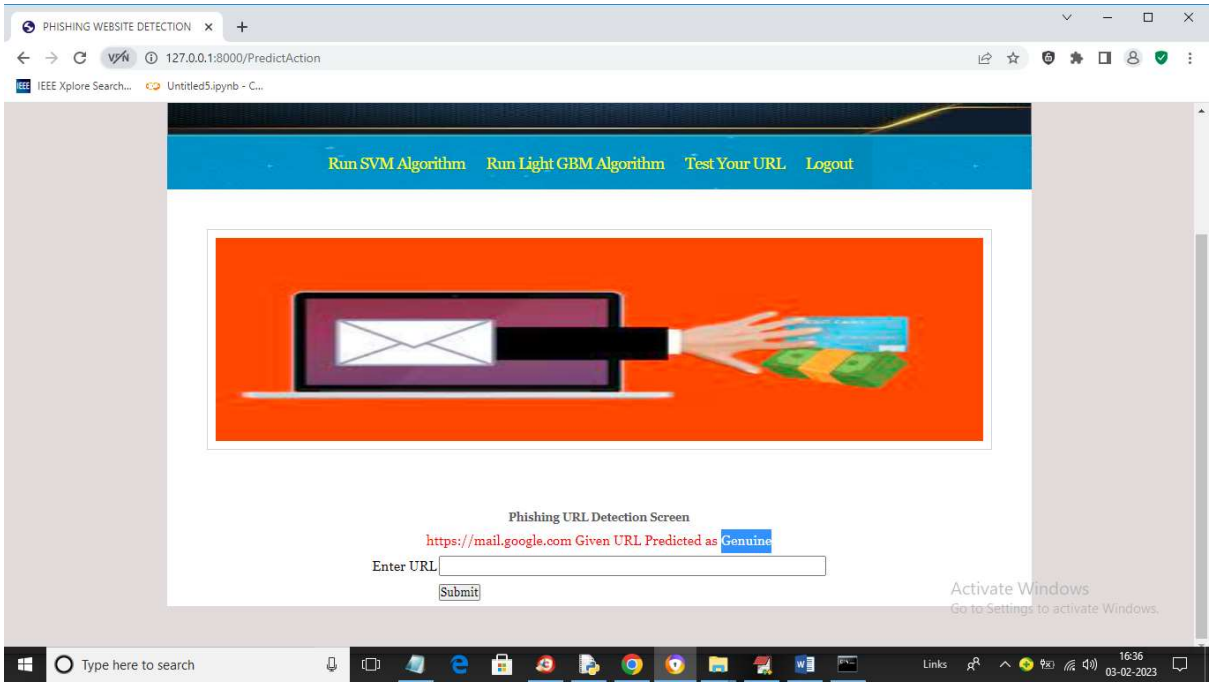
In above screen we can see Decision Tree confusion matrix graph and now close above graph to get below output



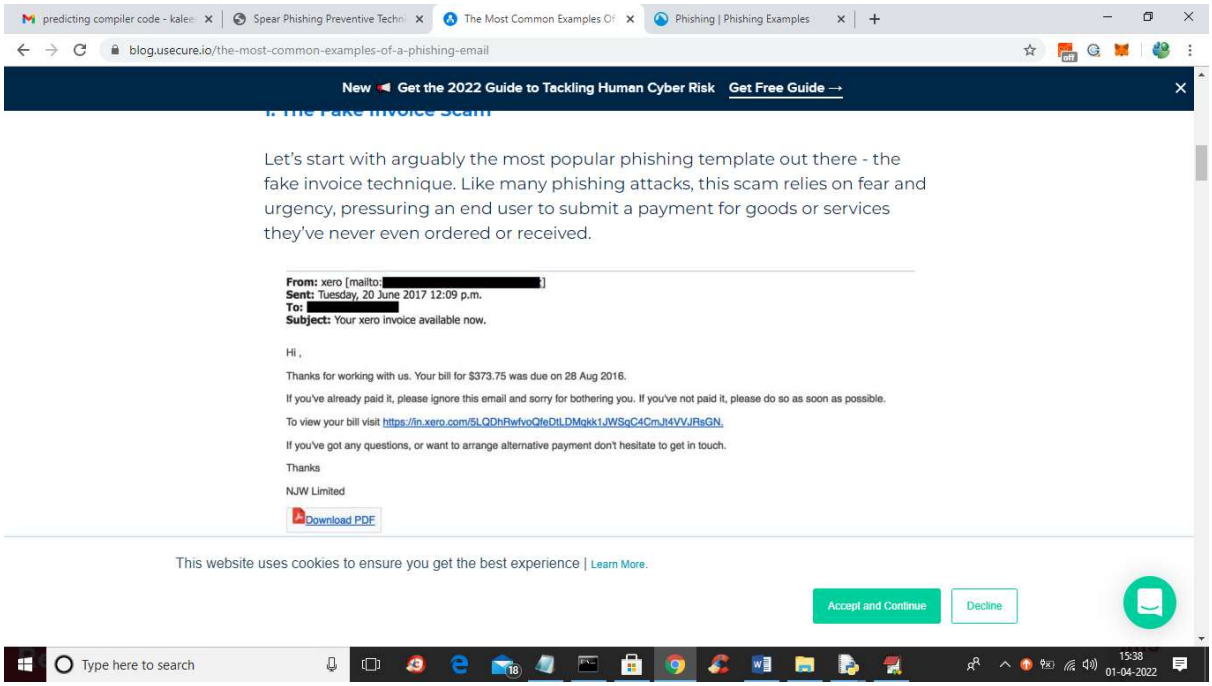
Algorithm Name	Accuracy	Precision	Recall	FScore
SVM	95.69486404833837	96.14835889512756	92.08515942262072	93.90369680582513
Light GBM	96.6012084592145	97.28403180486637	93.48840014474689	95.20561571161808

In above screen with Light GBM also we got 96% accuracy and now click on ‘Test Your URL’ link to get below screen

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.104 ISSN: 2581-4621



In above screen in blue colour text we can see given URL predicted as GENUINE (normal) and now test other URL.



In above screen blue colour URL is the phishing URL and I will input that to my application in below screen and below is the phishing URL from internet
'https://in.xero.com/3LQDhRwfvoQfeDtlDMqkk1JWSqC4CMJt4VVJRsGN'

DOI:10.46647/ijetms.2025.v09i02.104 ISSN: 2581-4621

Conclusion

This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

Referances

[1] Ms. SophiyaShikalgar, Mrs. Swati Narwane (2019), Detecting of URL based Phishing Attack using Machine Learning. (vol. 8 Issue 11, November – 2019)
[2] Rashmi Karnik, Dr. Gayathri M Bhandari, Support Vector Machine Based Malware and Phishing Website Detection.
[3] Arun Kulkarni, Leonard L. Brown, III2 , Phishing Websites Detection using Machine Learning (vol. 10, No. 7,2019)
[4] R. Kiruthiga, D. Akila, Phishing Websites Detection using Machine Learning.
[5] Ademola Philip Abidoye, Boniface Kabaso, Hybrid Machine Learning: A Tool to detect Phishing Attacks in Communication Networks. (vol. 11 No. 6,2020)
[6] Andrei Butnaru, AlexiosMylonas and Nikolaos Pitropakis, Article Towards Lightweight URL-Based Phishing Detection.13 June 2021
[7] Ashit Kumar Dutta (2021), Detecting phishing websites using machine learning technique. Oct 11 2021
[8] Nguyet Quang Do, Ali Selamat, OndrejKrejcar, Takeru Yokoi and Hamido Fujita (2021) Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical study.
[9] Ammara Zamir, Hikmat Ullah Khan and Tassawar Iqbal, Phishing website detection using diverse machine learning algorithms.
[10] Valid Shahrivari, Mohammad Mahdi Darabi and Mohammad Izadi (2020), Phishing Detection Using Machine Learning Techniques.
[11] A. A. Orunsolu, A. S. Sodiya and A.T. Akinwale (2019), A predictive model for phishing detection.
[12] Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.
[13] Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017, May). Malicious web content detection using machine leaning. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1432-1436). IEEE.