IJARST

International Journal For Advanced Research In Science & Technology

A peer reviewed international journal www.ijarst.in

ISSN: 2457-0362

PHISHING WEBSITE DETECTION USING LIGHT GBM AND SVM ALGORITHM

N. Teja Sri1, K. Harshitha2, M. Sai Akshaya3, M. Manasa4

1Assistant Professor Department of Information Technology Malla Reddy Engineering College for Women (UGC-Autonomous) Maisammaguda,Hyderabad, TS, India.

2,3,4 UG students Department of Information Technology Malla Reddy Engineering College for Women (UGC-Autonomous) Maisammaguda,Hyderabad, TS, India.

ABSTRACT:

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as using light gbm and svm algorithm.

Keywords: URL, SVM, Light GBM, Cyber security, phishing website.

1. 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.

OBJECTIVE OF THE PROJECT

Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

2 LITERATURE SURVEY

Rashmi Karnik et al., proposed a model of classification method, kernel-based approach. In this we categories phishing . This method produces estimated accuracy of 95% in detecting the phishing and malware sites.

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.

Vahid 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.

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 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.

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 analyse behaviour 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.

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.

Atharva Deshpande et al., proposed a combination of machine learning algorithms and natural language processing methods to detect the phishing domain appearances, the feature that distinguish them from real domains. Ms.

Sophiya Shikalgar et al., proposed a machine learning classifiers and methods to detect phishing website using Hybrid machine learning approach is a combination of different classifiers working together which gives a good prediction result. Each of classifiers have its own way of working and classification. Uses a data of URLs which contains 2905 URLs which is in unstructured form.

Nureni Ayofe Azeez et al., tried to handle this challenge, attempts have been made to address two major problems. The first is how can the suspicious URL’s be recognized on social networks and how can internet users can be protected from unreliable and fake URLs on the social network. It adapts six machine learning methods – AdaBoost, Gradient Boost, random forest, Linear SVM, decision tree and Naïve Bayes classifier for training using features obtained from the social network and for additional processing. A total of 532,403 posts were analysed. At last 87,083 posts were considered suitable for training the models. AdaBoost performs well among all with an accuracy of 95% and a precision of 97%. Ademola Philip Abidoye and Boniface Kabaso proposed a machine learning technique to accurately classify the dataset to identify the phishing URLs features that can be used by the attackers.

R. Kiruthiga and D. Akila explained a novel way of detecting phishing websites using machine learning methods and proposes a classification model in order to classify the phishing attacks. Also presents a way to detect phishing email attacks using natural language processing and machine learning produces a good accuracy.

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. The proposed approach effectively classified the Phishing and Legitimate URLs with an accuracy of 87.0% and 82.4% for Random Forest and decision tree classifiers respectively

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.

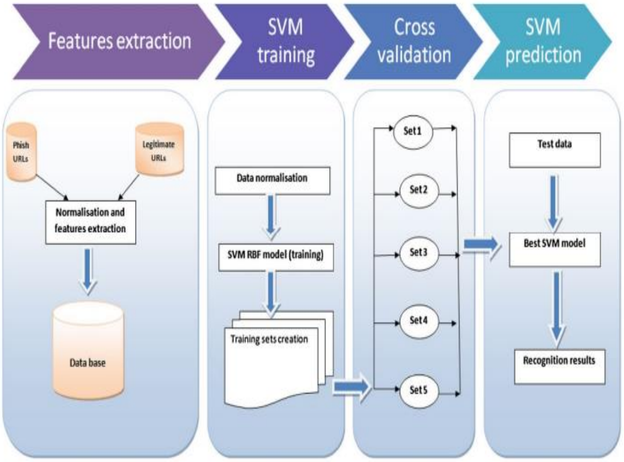


Fig.1. Phishing website process.

3 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.

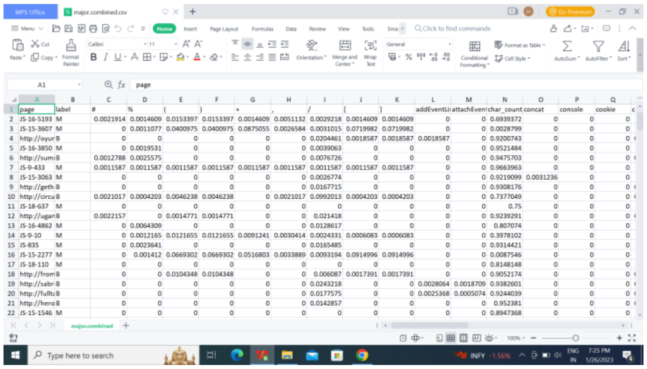
Support Vector Machine (SVM): This is also one of the supervised and simple to use classification algorithms. It can be

IJARST
International Journal For Advanced Research
In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

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.

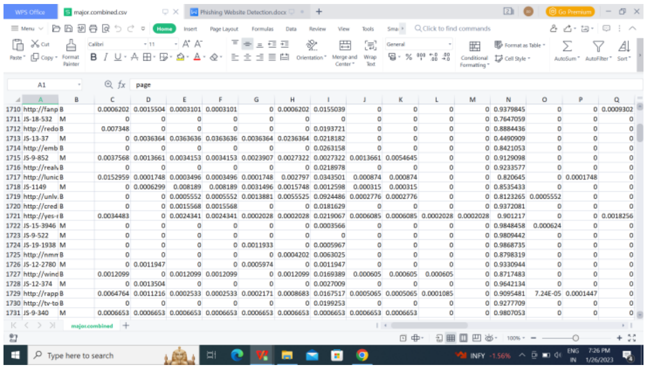
Data set:

Phishing continues to prove one of the most successful and effective ways for cybercriminals to defraud us and steal our personal and financial information. Our growing reliance on the internet to conduct much of our day-to-day business has provided fraudsters with the perfect environment to launch targeted phishing attacks. The phishing attacks taking place today are sophisticated and increasingly more difficult to spot. A study conducted by Intel found that 97% of security experts fail at identifying phishing emails from genuine emails.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	page	label	#	%	(	)	+	,	/	[	]	addEventLi	attachEven	char_count	concat	console	cookie
2	JS-16-5193	M	0.0021914	0.0014609	0.0153397	0.0153397	0.0014609	0.0051132	0.0029218	0.0014609	0.0014609	0	0	0.6939372	0	0	0
3	JS-15-3607	M	0	0.0011077	0.0400975	0.0400975	0.0875055	0.0026584	0.0031015	0.0719982	0.0719982	0	0	0.0028799	0	0	0
4	http://oyur	B	0	0	0	0	0	0	0.0204461	0.0018587	0.0018587	0.0018587	0	0.9200743	0	0	0
5	JS-16-3850	M	0	0.0019531	0	0	0	0	0.0039063	0	0	0	0	0.9521484	0	0	0
6	http://sum	B	0.0012788	0.0025575	0	0	0	0	0.0076726	0	0	0	0	0.9475703	0	0	0
7	JS-9-433	M	0.0011587	0.0011587	0.0011587	0.0011587	0.0011587	0.0011587	0.0011587	0.0011587	0.0011587	0	0	0.9663963	0	0	0
8	JS-15-3063	M	0	0	0	0	0	0	0.0026774	0	0	0	0	0.9219099	0.0031236	0	0
9	http://geth	B	0	0	0	0	0	0	0.0167715	0	0	0	0	0.9308176	0	0	0
10	http://circu	B	0.0021017	0.0004203	0.0046238	0.0046238	0	0.0021017	0.0992013	0.0004203	0.0004203	0	0	0.7377049	0	0	0
11	JS-18-637	M	0	0	0	0	0	0	0	0	0	0	0	0.75	0	0	0
12	http://ugan	B	0.0022157	0	0.0014771	0.0014771	0	0	0.021418	0	0	0	0	0.9239291	0	0	0
13	JS-16-4862	M	0	0.0064309	0	0	0	0	0.0128617	0	0	0	0	0.807074	0	0	0
14	JS-9-10	M	0	0.0012165	0.0121655	0.0121655	0.0091241	0.0030414	0.0024331	0.0006083	0.0006083	0	0	0.3978102	0	0	0
15	JS-835	M	0	0.0023641	0	0	0	0	0.0165485	0	0	0	0	0.9314421	0	0	0
16	JS-15-2277	M	0	0.001412	0.0669302	0.0669302	0.0516803	0.0033889	0.0093194	0.0914996	0.0914996	0	0	0.0087546	0	0	0
17	JS-18-110	M	0	0	0	0	0	0	0	0	0	0	0	0.8148148	0	0	0
18	http://from	B	0	0	0.0104348	0.0104348	0	0	0.006087	0.0017391	0.0017391	0	0	0.9052174	0	0	0
19	http://sabr	B	0	0	0	0	0	0	0.0243218	0	0	0.0028064	0.0018709	0.9382601	0	0	0
20	http://fullt	B	0	0	0	0	0	0	0.0177575	0	0	0.0025368	0.0005074	0.9244039	0	0	0
21	http://hero	B	0	0	0	0	0	0	0.0142857	0	0	0	0	0.952381	0	0	0
22	JS-15-1546	M	0	0	0	0	0	0	0	0	0	0	0	0.8947368	0	0	0

The provided dataset includes 11430 URLs with 87 extracted features. The dataset is designed to be used as benchmarks for machine learning-based phishing detection systems. Features are from three different classes: 56 extracted from the structure and syntax of URLs, 24 extracted from the content of their correspondent pages, and 7 are extracted by querying external services. The dataset is balanced, it contains exactly 50% phishing and 50% legitimate URLs.



A1 fx page

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1710	http://fanp	B	0.0006202	0.0015504	0.0003101	0.0003101	0	0.0006202	0.0155039	0	0	0	0	0.9379845	0	0	0.0009302
1711	JS-18-532	M	0	0	0	0	0	0	0	0	0	0	0	0.7647059	0	0	0
1712	http://redo	B	0.007348	0	0	0	0	0	0.0193721	0	0	0	0	0.8884436	0	0	0
1713	JS-13-37	M	0	0.0036364	0.0363636	0.0363636	0.0036364	0.0236364	0.0218182	0	0	0	0	0.4490909	0	0	0
1714	http://emb	B	0	0	0	0	0	0	0.0263158	0	0	0	0	0.8421053	0	0	0
1715	JS-9-852	M	0.0037568	0.0013661	0.0034153	0.0034153	0.0023907	0.0027322	0.0027322	0.0013661	0.0054645	0	0	0.9129098	0	0	0
1716	http://realv	B	0	0	0	0	0	0	0.0218978	0	0	0	0	0.9233577	0	0	0
1717	http://lunic	B	0.0152959	0.0001748	0.0003496	0.0003496	0.0001748	0.002797	0.0343501	0.000874	0.000874	0	0	0.820645	0	0.0001748	0
1718	JS-1149	M	0	0.0006299	0.008189	0.008189	0.0031496	0.0015748	0.0012598	0.000315	0.000315	0	0	0.8535433	0	0	0
1719	http://unlv.	B	0	0	0.0005552	0.0005552	0.0013881	0.0055525	0.0924486	0.0002776	0.0002776	0	0	0.8123265	0.0005552	0	0
1720	http://cred	B	0	0	0.0015568	0.0015568	0	0	0.0181629	0	0	0	0	0.9372081	0	0	0
1721	http://yes-r	B	0.0034483	0	0.0024341	0.0024341	0.0002028	0.0002028	0.0219067	0.0006085	0.0006085	0.0002028	0.0002028	0.901217	0	0	0.0018256
1722	JS-15-3946	M	0	0	0	0	0	0	0.0003566	0	0	0	0	0.9848458	0.000624	0	0
1723	JS-9-522	M	0	0	0	0	0	0	0	0	0	0	0	0.9809442	0	0	0
1724	JS-19-1938	M	0	0	0	0	0.0011933	0	0.0005967	0	0	0	0	0.9868735	0	0	0
1725	http://nmm	B	0	0	0	0	0	0.0004202	0.0063025	0	0	0	0	0.8798319	0	0	0
1726	JS-12-2780	M	0	0.0011947	0	0	0.0005974	0	0.0011947	0	0	0	0	0.9330944	0	0	0
1727	http://wind	B	0.0012099	0	0.0012099	0.0012099	0	0.0012099	0.0169389	0.000605	0.000605	0.000605	0	0.8717483	0	0	0
1728	JS-12-374	M	0	0.0013504	0	0	0	0	0.0027009	0	0	0	0	0.9642134	0	0	0
1729	http://rapp	B	0.0064764	0.0011216	0.0002533	0.0002533	0.0002171	0.0008683	0.0167517	0.0005065	0.0005065	0.0001085	0	0.9095481	7.24E-05	0.0001447	0
1730	http://tv-to	B	0	0	0	0	0	0	0.0199253	0	0	0	0	0.9277709	0	0	0
1731	JS-9-340	M	0.0006653	0.0006653	0.0006653	0.0006653	0.0006653	0.0006653	0.0006653	0.0006653	0.0006653	0	0	0.9807053	0	0	0

4 IMPLEMENTATION

In this section, we will going discuss about the actual steps which were executed while doing the experiment. We shall discuss the stepwise procedure used to analyse the data and to predict the phishing. We have used unstructured data which consists only url. There are 11064 urls obtained from the internet. Which consists of both phishing and genuine url where most of urls obtained are phishing.

1. First, we have unstructured data of urls from Phish tank website.

2. In Preprocessing, feature generation is done where eight features are generated from unstructured data. These features are length of url, http tokens, suspicious character, prefix/suffix, number of slashes, phishing term, length of subdomain, url IP address.

3. Next, a structured dataset is constructed, with binary values (0,1) for each feature, which is then sent to the various classifiers.

4. Next we train the two different classifiers and compare their performance on the basis of accuracy two classifiers namely SVM, lightgbm.

5. Then classifier detects the given url based on the training data that is if the site is phishing it shows error and if legitimate it opens that page in browser.

6. We compare the accuracy of different classifiers and found lightgbm is the best classifier which gives the maximum accuracy.

7. Below are the screen shots for the execution process.

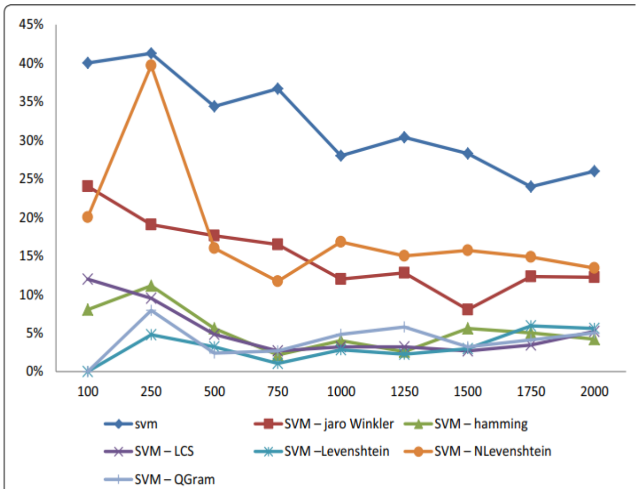


Fig.2. Graphical representation

Affirmed by the results of our tests, we have demonstrated the potential impact of the use of the similarity distance on the detection of phishing websites. Indeed, in three tests performed on four, the introduction of the distance of similarity has significantly improved the recognition rate of our detection system. In the same context, the only case where the similarity did not have a positive impact on the phishing websites recognition rate is the test with Probabilistic Neural networks that records the worst recognition rate among all of our tests. This impact is most obvious in tests performed using the SVM method since the use of the Hamming distance as one of the input characteristics of our system has improved the recognition rate of 21.8%.

5 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.

Feature Analysis

The features of the domain name used here can be obtained only by using known strings of domain names without obtaining information related to user privacy, such as traffic in the network.

Features of the domain name can be divided into two categories according to the acquisition method: features of the characters used in the domain name and features of information on the domain name. The features of information on the domain name can be obtained through the corresponding website or other query websites to this end, whereas the features of the characters used in the domain name can be obtained through a local feature-extraction algorithm without visiting the website.

6 REFERANCES

[1] Ms. Sophiya Shikalgar, 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, Alexios Mylonas 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, Ondrej Krejcar, 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] Vahid 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.