

Volume : 53, Issue 7, July : 2024

# HYBRID APPROACH FOR PHISHING WEBSITE DETECTION USING M L.

#### M.PRAVEENA, HOD, Department of COMPUTER SCIENCE, SRI DURGA MALLESWARA SIDDHARTHA MAHILA KALASALA, Vijayawada, A.P., India.

### V.SIVA KRISHNAVENI, LECTURER, Department of Computer Science, SRI DURGA MALLESWARA SIDDHARTHA MAHILA KALASALA, Vijayawada, A.P., 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

### **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. Alot of useful data, information and datacan 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 phishingattack 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 artificial intelligence and data mining techniques



Volume : 53, Issue 7, July : 2024

[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 requireas 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.

### **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 detectphishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

### 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 GoogleSafe 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 forforecasting 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



Volume : 53, Issue 7, July : 2024

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 heapingrepresentations 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-termmemory (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 learningmodels. In which each model showsdifferent accuracies from different models.

### **EXISTING SYSTEM:**

Phishing is an internet scam in which an attacker sends out fake messagesthat 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:**

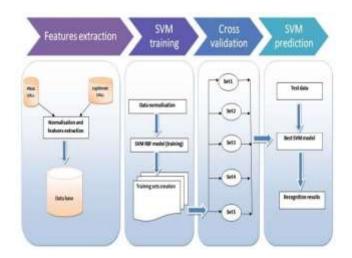
Phishing attacks have evolved in terms of sophistication and have increased in sheer number in recent years. This has led to corresponding developments in the methods used to evade the detection of phishing attacks, which pose daunting challenges to the privacy and security of the



Volume : 53, Issue 7, July : 2024

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 multipledomain-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.



#### **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



Volume : 53, Issue 7, July : 2024

various applications, we cannot discriminate which of the algorithms are superior ornot.

**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, classificationapplications 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 setpoints 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 perfectent to launch targeted phishingattacks. The phishing attacks taking place today are sophisticated and increasingly more difficult to spot. A study conducted by Intel found that 97% of security expertsfail at identifying phishing emails from genuine emails.

	5-60 (D) (to of the							1000	a the second second			0.0	14-1-10	
17h 26-10 11h	Laws-	- V.R.	76. C	1 10 1 1	D (0 )	- E	101 10			1001 1	Dimension (and	- 51		.41
ner Pres term	(0.7.10) A1	E- 12 0	4.0	1.2.21	新教 计			Q	1.12.25		direction 1	-	-	1
and a second second										and a				
41.	51 Tr													
			1.1.1.4						A WALLAND		W	n	P	A:
Corner Later	4 10					1		-	فجرا المحجز الغط	and distances and	, in a little in		de sente	÷. 1
8 24-12-1101 he 8 21-12-2007 Ad	0.0001014 0.000	10000 0.031330	07 G.G.M.1990	2.0C10000	0.0015130	0.000403.8	CONTRACTOR -	2-SPEAKOW	- 2	- 2.24	0.08/129			- 2
B Tothe Creek P		.0	10.000		1 10	0.0004466	CONTRACT.	L.Bronner F			and the second			- 2
17-39-3616 M	0.0.00	105.01				D. CONTRACTOR				0.01	F125464	. 10		
a long // service	8-8012786 8-000										********			
1 21-0-010 Ad	8.4011087 0.00	cont concern	4	a summine i	eronicies	0.00121082	continue.	0.0011587			CONTRACT OF STREET	and and		- 2
# John / gran h			a			6.0ahr/st					Addito The			
B. Print (Colorestill	-0.8011017 0.90		in in-president	r. 1		0.00000033	OLGENHOUS, 1	a description			TRITTICHE.	10.		- 2
12 18-34-807 M		- 18	H						- 18		0.0%	81.	18	
a marchane	0.0011197	- 10 D.DO.A.477	TE BORNATE			100404					ACRESSION		. 0	
ty it. to aming be	0.000	LINE DESIDE	2			electrony of	-	- Annotation			and shows a		- 2	- 2
G. N-AN	0.0.00	Contar.	4			d.contacts					11000			
M 11111 04 1111 M		AAL) DOMING									616LTT.410	44	18	
0 H-48-150 M											A AREA OF		- 18.	
B Holgs JVYrani B	2	O D.D.D.M.M.	AR GOLDANSKI			5.00mph/	contrast.	statutions.	1.00 m	0.01	KURCEPH .	- 5 -		21
B THE AVERAGE		- 2	2	· · · · ·		0.00177979			CONTRACTOR IN	coordinate and	AT BARDON	- 2-		- 2
A CONTRACTOR OF A		1.00				0.008449117					USJANS .		10.	
in an als sugge out			4 4							10.044	multane .			
17								10. B		[] [] [] [] []				+ 1
I 27 Topic horses			Contract of Contract	<b>~</b>	1000					and the second second		· · · ·	a. Supervise	
na mascom	derived types			n (unama)		+					(73)			-
e e e e e e	40004100 139 -= c+ + •		Page 1 (model)	ne (Sariacimus) Sari-cular	inia – Itala	+	- 1404	104 (j) = 0		d tannatianal		р <mark>сі</mark> "b	45 - 54 -	-
	demind tops 19 1-12 - 14 1		Page Lineard	a C		+	3#1 D	Sec.		10021	HD recent as	2 C	α- 10 · Σ · 3	
Base	40004100 139 -= c+ + •		Page Lineard	a C			AT ITC	Sec.		10021		2	45 - 54 - 52 - 5	
	demind tops 19 1-12 - 14 1		Page Lineard	a C		anterio - Attenue	ST ITTE	Sec.		10021	HD recent as	р (1 ф 1000-		- C) C) Fract
	annedice B = cr + • • 7 Ll - A ⊞ • • • • • • •		harrend w - Q - E			annen i Anerge	9 190 977 707	Sec.		1001 	ND too too too	0 00 0000 0000 0000 0000 0000 0000 000	ann an Andre	
	Annual Los Di		Partiend 		n coorace	n or hears		Sec.	re su ai	1001 	N 0.0070845	ы 0		. and
	Annual Los Di		Partiend 		n poorsos	u.o.iscosu	9 250 27 10 •	Sec.			ND	B (1)	4 0 0	. and
Competence of the second secon	Annualism Di = c = c = c 7 Li = A III 6 D 1000HEGGS BARIAN 9.007346				inni dill 4 dill 4 n. consol	8.0195039 0.0195039	9 250 27 10 •	Sec.	re su ai	1000 C	N 0.0070845	40 60 0	4 0 0	. and
A1 A A A A A A A A A A A A A A A A A A	4411101.000  2		harrent e e e e e e e e e e e e e e e e e e	needda		4 0.0195039 0.0191703 0.02191703 0.02191703		40 + 74. 8	re s⊫ 20	1000 A	N 1,2 cm hos 0,0370845 0,7647058 0,4490809 0,44921053	40 60 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	* 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	. and
A1 Insectors	Annel Los B = = + + + + + + + + + + + + + + + + +	· · · · · ·	10011001 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1000100 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1010 111 0.0000000 0.00000000 0.00000000	4.016000 0.0181731 0.02181731 0.0218158 0.0027322	100013601	4 4 0.0014045	re s⊫ 20	1000 C	N 12 <sup>2</sup> - 10 - 10 - 10 12 <sup>2</sup> - 10 - 10 - 10 0 - 03 - 20 0 - 40 - 20 0 - 40 - 20 0 - 40 - 20 0 - 21 - 20 0	40 60 00 00	* 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	. and
Transc com     Transc com     Transc man	400001300 () 2 = = = + + + + + + + + + + + + + + + +	K A	harrest 	needa a a 0 a 0	1000 1000 1000 1000 1000 1000 1000 100	1 0.0155039 0.0155039 0.02181731 0.0218181 0.0203180 0.02218029	10013000	* *	r	1000 000000000000000000000000000000000	N 12 <sup>2</sup> - 111 - 111 12 <sup>2</sup> - 111 - 111 - 111 12 <sup>2</sup> - 111 - 111 - 111 12 <sup>2</sup> - 111 - 111 - 111 - 111 12 <sup>2</sup> - 111 - 11	40 60 00 00 00 00 00 00 00	P 000	0.00
	Annualiza Di Contra di Contra Contra di Contra	A A A	hurrent 	needda a d a a a a a a a a a a a a a	1111	1 0.0155030 0.02191233 0.02191293 0.02191292 0.02180292 0.0218029	10 (10) 10	* * 0.0054645 0.000074	t 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1000 000000000000000000000000000000000	10	40 60 00 00 00 00 00 00 00	P 0	0.00
Compare Control Contro Control Control Control Control Control Control Control Control C			Puntiend 	0.0001148 0.0001148	1 111	1 1 1 1 1 1 1 1 1 1 1 1 1 1	Al (WC) (P <sup>-1</sup> The A A A A A A A A A A A A A	4 0 0.0054045 0.00054045 0.000014045 0.0000170 0.0000170	r 1 1 1 1 1 1 1 1 1 1 1 1 1 1	*	N 10 <sup>4</sup> cur nun 0.0370845 0.7647030 0.4047030 0.4449000 0.8421053 0.9248040 0.924577 0.920845 0.9224577 0.820845 0.8418443	80 00 00 00 00 00 00 00 00 00 00 00 00 0	P 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.00
Al      Al	Annual Los B	A A     A	Puntumoi 	1000010 0 0 0000000 0 00000000 0 00000000 0 00000000	1 111 1 111 1 111 1 11 1 100003003 0 00037002 0 0003700 0 0003700 0 0003700 0 0003700 0 0003700 0 000370 0 00000 0 0000 0 00000 0 0000 0 00000 0 000000 0 00000000	1 1 1 1 1 1 1 1 1 1 1 1 1 1	A) (100) 	60 - 35 60 - 35 6 0.0054645 0.000475 0.000375 0.00005000000000000000000000000000000	r 1	*	N 10 <sup>2</sup> cormon 10 <sup>2</sup> cormon 0.03790846 0.70470308 0.404900 0.4442055 0.4449000 0.4442055 0.4449000 0.9421053 0.9120046 0.9120046 0.9123200 0.0372091	80 00 00 00 00 00 00 00 00 00 00 00 00 0	P 0	4 0.000
A1			Functional 	0.00011748 0.00011748 0.00011748 0.00011748 0.00011748 0.00011811 0.00011811 0.00011811	1 (2) (4) 1 (2) (4)	1 1 1 1 1 1 1 1 1 1 1 1 1 1	4) (10) 4 4 4 5 5 6 6 6 6 6 6 6 6 6 6 6 6 6	60 m 0.00154645 0.000110 0.0000110 0.0000110 0.0000110 0.0000110 0.0000110 0.0000110 0.0000110 0.0000100		M 0000000	N 10 <sup>2</sup> cm mm 10 <sup>2</sup> cm mm 0.0070845 0.0070865 0.0004430 0.004430 0.004430 0.004430 0.004430 0.004430 0.004450 0.0025570 0.0025570 0.00270001 0.00272001 0.0027201	60 60 60 60 60 60 60 60 60 60 60 60 60 6	# 000000000000000000000000000000000000	6.600
Image: Control         Image: Control           Image: Contretee         <	Annual Los B	A A     A	Puntimed 	10000000000000000000000000000000000000	Han (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)	1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 190 191 191 191 191 191 191 191	60 m 0.00154645 0.000110 0.0000110 0.0000110 0.0000110 0.0000110 0.0000110 0.0000110 0.0000110 0.0000100		M 000	N 10 <sup>2</sup> correction 10 <sup>2</sup> correction 0.01700845 0.7047050 0.4047050 0.4047050 0.441055 0.9120000 0.9215077 0.800645 0.4121200 0.9222001 0.9222001 0.0012170 0.0012101 0.0012101	80 00 00 00 00 00 00 00 00 00 00 00 00 0	# 000000000000000000000000000000000000	0.00
A1     A2     A2     A	Annual Los B		Puntimed 	10000000000000000000000000000000000000	Han (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)	1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 190 191 191 191 191 191 191 191	60 m 60 m 60 m 60 m 60 m 60 m 60 m 60 m 70 m	0.000103		N 10 <sup>2</sup> cm mm 10 <sup>2</sup> cm mm 0.0070845 0.0070865 0.0004430 0.004430 0.004430 0.004430 0.004430 0.004430 0.004450 0.0025570 0.0025570 0.00270001 0.00272001 0.0027201	60 60 60 60 60 60 60 60 60 60 60 60 60 6	# 000000000000000000000000000000000000	0.00
Image: Control         Image: Control           Image: Control         Image: Control         Image: Control         Image: Control           Image: Control         Image: Control         Image: Control         Image: Control         Image: Control           Image: Control         Image: Control         Image: Control         Image: Control         Image: Control         Image: Control	Anneal Los 2	A A     A	Austived 4 4 6 0 0 0 0 0 0 0 0 0 0 0 0 0	Results	4000 400 4000 4	1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 (10) 1 (1) 1	90) 150 90) 150 100054045 0.0005404 0.000310 0.000300 0.000300 0.000300 0.0000000000	14	HEI Partitioner M 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ID         ID<	60 60 60 60 60 60 60 60 60 60 60 60 60 6	# 000000000000000000000000000000000000	6.600
A1     A1     A		A A A     A	Part Level 	10000000000000000000000000000000000000	Hanni	1 0.0186049 0 0.018734 0 0.018734 0 0.018734 0 0.0284359 0 0.024452 0 0.024469 0 0.024469 0 0.024469 0 0.024469 0 0.02469 0 0.02469 0 0.02469 0 0.02469 0 0.02469 0 0.02469 0 0.00469 0 0.00469	1 (11) 11 (11) 11 (11) 1 (1	\$ \$ 0.005405 0.005710 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.000210 0.0002000 0.0002000 0.0002000 0.0002000 0.0002000 0.0002000 0.0000000000	u+ 5∦ 28	100000000 0000000000000000000000000000	ND 1000000000000000000000000000000000000	60 60 60 60 60 60 60 60 60 60 60 60 60 6	# 000000000000000000000000000000000000	6.600
A1     A1     A1     A		A A A     A	Part Level 	10000000000000000000000000000000000000	Hanni	1 0.0186049 0 0.018734 0 0.018734 0 0.018734 0 0.0284359 0 0.024452 0 0.024469 0 0.024469 0 0.024469 0 0.024469 0 0.02469 0 0.02469 0 0.02469 0 0.02469 0 0.02469 0 0.02469 0 0.00469 0 0.00469	1 (11) 11 (11) 11 (11) 1 (1	¥2) - 32 * 0.005405 0.005415 0.000015 0.000015 0.000015 0.000015 0.000015 0.000015 0.000015 0.0000015 0.0000015 0.0000015 0.0000015 0.0000005 0.0000005 0.0000005 0.0000005 0.0000005 0.0000005 0.0000005 0.0000005 0.0000005 0.0000005 0.000005 0.000005 0.000005 0.000005 0.000005 0.000005 0.000005 0.000005 0.0005 0.0005 00	14 54 28	10000000000000000000000000000000000000	10	44 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	P 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	6.600
Image: Second		Ar A	Part Level 	Render R R R R R R R R R R R R R	1000 1 El 1 1 El 1	1 000000000000000000000000000000000000	1 (11) 11) 11) 10) 10) 10) 10) 10)	\$ \$ 0.005405 0.005710 0.0000710 0.0000710 0.0000710 0.0000710 0.0000000 0.0000000 0.00000000 0.00000000	un 21 28	1000/0000 0000 0000 0000 0000 0000 0000 0000 0000	HD 1000000000000000000000000000000000000	50 50 50 50 50 50 50 50 50 50 50 50 50 5	P	0.095
A1     A	40000100	A A	Part Level 	Render R R COMMISSION COMM	1000 1 El 1 1 El 1	4 0.014/00.000 0.014/00.000 0.014/00.000 0.014/00.000 0.000/000 0.0000 0.000/000 0.000/000 0.000/0000 0.000/000 0.000/000	11 (11) 11	9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	4 4 0.000103 0.000103 0.000103 0.000105	1000/0000 0000 0000 0000 0000 0000 0000 0000 0000	10	50 50 50 50 50 50 50 50 50 50 50 50 50 5	P 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.00
A1     A		A A	Part Level 	Render R R COMMISSION COMM	1000 1 El 1 1 El 1	4 0.014/00.000 0.014/00.000 0.014/00.000 0.014/00.000 0.000/000 0.0000 0.000/000 0.000/000 0.000/0000 0.000/000 0.000/000	11 (11) 11	9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	4 4 0.000103 0.000103 0.000103 0.000105	1011 200000000 000000000 00000000000000	III)	60 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	P 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.00
Image: Control         Image: Control           Image: Control         Image: Control         Image: Control <td>Annual Los D2</td> <td>A A</td> <td>Part Level </td> <td>Render R R COMMISSION COMM</td> <td>1000 1 El 1 1 El 1</td> <td>4 0.014/00.000 0.014/00.000 0.014/00.000 0.014/00.000 0.000/000 0.0000 0.000/000 0.000/000 0.000/0000 0.000/000 0.000/000</td> <td>11 (11) 11 (11) 11</td> <td>9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9</td> <td>4 4 0.000103 0.000103 0.000103 0.000105</td> <td>1011 200000000 000000000 00000000000000</td> <td>HD 1000000000000000000000000000000000000</td> <td>60 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td> <td>P 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td> <td>0.00</td>	Annual Los D2	A A	Part Level 	Render R R COMMISSION COMM	1000 1 El 1 1 El 1	4 0.014/00.000 0.014/00.000 0.014/00.000 0.014/00.000 0.000/000 0.0000 0.000/000 0.000/000 0.000/0000 0.000/000 0.000/000	11 (11) 11	9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	4 4 0.000103 0.000103 0.000103 0.000105	1011 200000000 000000000 00000000000000	HD 1000000000000000000000000000000000000	60 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	P 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.00



Industrial Engineering Journal ISSN: 0970-2555 Volume : 53, Issue 7, July : 2024 **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.

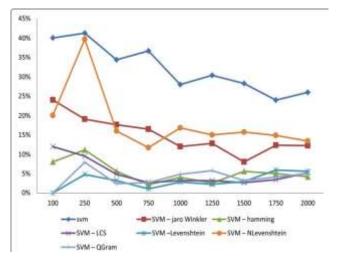
2. In Preprocessing, feature generation is done where eight features are generated from unstructured data. These features are length of url, http tokens, suspiciouscharacter, prefix/suffix, number of slashes, phishing term, length of subdomain, url IPaddress.

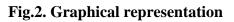
3. Next, a structured dataset is constructed, with binary values (0,1) for each feature, which is then sent to the various classifiers.

4. Next we train the two different classifiers and compare their performance on the basis of accuracy two classifiersnamely 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.





Affirmed by the results of our tests, we have demonstrated the potential impactof the use of



Volume : 53, Issue 7, July : 2024

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 mostobvious 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.

### **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. 7,2019R. Kiruthiga, D. Akila, Phishing Websites Detection using Machine Learning.

[4] Ademola Philip Abidoye, Boniface Kabaso, Hybrid Machine Learning: A Tool to detect Phishing Attacks in Communication Networks. (vol. 11 No. 6,2020)

[5] Andrei Butnaru, Alexios Mylonas and Nikolaos Pitropakis, Article Towards Lightweight URL-Based PhishingDetection.13 June 2021

[6] Ashit Kumar Dutta (2021), Detecting phishing websites using machine learning technique. Oct 11 2021

[7] Nguyet Quang Do, Ali Selamat, Ondrej Krejcar, Takeru Yokoi and Hamido Fujita (2021) Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical study.



Volume : 53, Issue 7, July : 2024

[8] Ammara Zamir, Hikmat Ullah Khan and Tassawar Iqbal, Phishing website detection using diverse machine learning algorithms.