ISSN No:-2456-2165

Classification of Phishing Website Using Hybrid Machine Learning Techniques

T.Pavansai
Vignan University
Information Technology,
Guntur, AP

Ziaul Haque Choudhury
Vignan University
Information Technology,
Guntur, AP

G.Gowtham sai
Vignan University
Information Technology,
Guntur, AP

Abstract:- The problem with cyber security involves scam websites, stilling the information that exploit people's trust. It could be reduced to the act of enticing internet users even though that they can get their personal data, including user names and passwords. In this study, we present a method for identifying phishing websites. The technology works as an add-on to a web browser, alerting the user when it finds a phishing website. A machine learning technique, specifically supervised learning is proposed in our study. The Logistic regression, Principal Component Analysis (PCA) and Apriori algorithms are chosen because of its success in classification. By examining the characteristics of phishing websites and selecting strongest combination of them, we developed a classifier that performs better.

Keywords:- *Phishing Website, Cyber Security, Machine Learning.*

I. INTRODUCTION

In this era, the modern world, technologies are merged completely. One of those technologies, that is advancing quickly each day and has a significant effect on people's lives, is the web and internet. It has evolved into a valuable and handy platform for facilitating public transactions such as e-banking and e-commerce. Users now believe that giving their private information to the internet is convenient as a result of this. A significant security issue has arisen as a result of the security thieves who have started to target this material. One of these issues is what are known as phishing sites. They are using social engineering, which may be characterized as con artists trying to influence the consumer into providing personal information? In accordance with the Anti-Phishing Task force, statistics indicate that such frequency of phishing assaults is rising, posing a threat to user data. (APWG) [1] as well as Mcafee Lab [2], which noted phishing assaults, reported an increase of 47.48% compared to all phishing attacks discovered in 2016.

Internet-connected gadgets and their services are becoming increasingly widely used all over the world as a result of technical advancement. IoT devices, despite being regarded as novel technology garner more attention for other web system security challenges as well. Several efforts have been made to address these difficulties, and machine learning methods are frequently used in their execution [1–3]. An attack called as phishing tricks victims into accessing malicious files and divulging personal details. The majority of fake sites utilize the same Domain and web experience as trustworthy websites. There is a great need for an intelligent plan to protect customers from cyber-attacks [3].

The person got redirected to that website if they click on a phishing link. The attacker uses the victim's information to gain access to other official websites after taking it. Several alternative detection procedures are developed and used in the literature to identify this kind of phishing attempt. Use of signature-based/rule-based detection techniques is the simplest strategy [4]. The signature of the phishing assault is listed in this method. This link might, for the purpose of detecting attacks, become the description of the URL addresses.

Many studies are recently been conducted in an effort to address to the phishing issue. Some researchers utilized the URL and compared it with already-existing watch lists and include lists of harmful websites that they have been developing, while others used the URL in the other way, comparing it with a whitelist of trustworthy websites [4]. The latter strategy makes use of heuristics and a database of signatures. Additionally, some studies have used methods of machine learning. Computer programming, a sub field of artificial intelligence (AI), that executes jobs and has the capacity to learn or behave intelligently, includes the discipline of machine learning. It really has supervised learning and unsupervised learning as its two separate active learning. A model is prepared for supervised learning by providing it with a collection of measurable characteristics of data linked to a target label corresponding to this data. Once the classifier is developed, it may create a new label with unknown data. Unsupervised learning, in contrast hand, is based on creating fresh data without providing a goal label throughout the training phase.

Among the main problems with data security was phishing. Users can click on links which take them straight to a fake website or they may receive malicious email that connect to the phoney website. Nevertheless, the two approaches have one thing in common: rather than technical flaws, the attacker focuses on human weaknesses [3]. Phishing is the practise of fraudsters tricking victims into divulging their personal information, including usernames,

passwords, and credit card numbers. Users are experiencing economical and financial troubles as a result of these frauds.

Throughout the 1990s, phishing set up a fraudulent profile on America Online (AOL), a corporation that offered an online system and had a web portal, using a false name and a counterfeit credit card. The phishers may be using its services in this way at no expense to themselves. Since then, in the middle of the 1990s, AOL has upgraded its anti-phishing system. Sadly, the phishers used a different strategy, hijacking legitimate accounts by impersonating an AOL representative and asking customers to surrender their passwords for safety. Both emails and text messaging were used for this the algorithms technique compares any attack patterns with the signature of a systematic pattern using signatures database of those attacks. Intuitions have the drawback of failing to identify novel threats since signatures may be easily evaded through obfuscation. Considering the rise in novel assaults, particularly 0 threats, upgrading the authentication system is also a laborious process [7]. In order to identify phishing websites, content analysis uses well-known methods like phrase intensity often the (TF-IDF). In order to determine if a website is scamming or not, it examines the message content of each page on the site. Other techniques used by researchers to identify phishing websites include monitoring site traffic with Alexia. Machine-learning makes use of this predictive capability. After learning the traits of the fake website, it makes predictions about new phishing traits. There are various techniques, including artificial neural network (ANN), Naive Bayes, support vector machine (SVM), logistic regression, and naive Bayes (NB) (BN). From algorithm each engine, phishing detection performance differs.

In the study, we describe a method for recognize URLs using ML techniques. Using Addresses, a recurrent neural technique is utilized to identify phishing websites. Our work aims to increase cypher attack detection rates by offering good performance with low false-negative and false-positive rates as phishing schemes grow more prevalent. False-negative sites are those that are misidentified as authentic websites, whereas false-positive sites are those that are misidentified as legitimate websites Figure 1 shows a straightforward description of phishing. When a client goes to a website and opens on an email, phishing begins. Outside connection, for example a pop-up or advertising [5].

A. Problem statement

When new phishing strategies are launched, phishing detection solutions do suffer from low detecting quality and high rates of false alarms. Additionally, because registration new domains has gotten simpler, the most popular methodology, the exclusion list method, is ineffective at reacting to phishing assaults that are on the rise. No complete blacklisted can guarantee a flawlessly a go dataset. In addition, several solutions have made use of page content analysis to address the false negative issues and strengthen the weaknesses of the expired lists. Additionally, different web pages analysis techniques each employ a unique approach to accurately identifying malicious URLs. Because aggregation may mix overall similarities in correctness and various probability qualities in chosen methods, it can be viewed as a better option. As a result, the following research will be included in this study.

II. LITERATURE SURVEY

Anti-Phishing is employed to stop consumers from visiting shady websites, which might result in phishing attacks. Here, Anti Phish tracks the sensitive information that the user is required to fill out and warns the user anytime he or she tries to exchange that information with an unreliable website. The most effective explanation for this is to encourage consumers to only visit reliable websites. This strategy, nevertheless, is impractical. In any case, the user could be duped [4]. As a result, it is necessary for the associates to provide these justifications in order to combat the phishing issue. Options that are widely recognized rely on spooky websites to identify "clones" and keep track of malicious scams that are on the hit list. An effective procedure of machine dependability on a trait meant for the reflecting of the beleaguered user deceit via telecommunications is an option for identifying these attacks. This method may be used to identify spam emails or texts and emails that are transmitted via emails and used to capture victims. Roughly 800 scam emails and 7,000 non phishing emails have been tracked so far, and over 95% of them have been properly identified. Moreover, 0.09% of the actual emails have been used to classify the emails. We can just finish with the techniques for spotting the trick and the evolving the majority of e-banking providers employ [7]

Phishing websites, which nature of the attacks. Identification and classification are quite dynamic and intricate. As it deals with taking into consideration numerous quality criteria rather than precise numbers, some vital data mining techniques may offer an excellent way to maintain the e-commerce websites safe due to the participation of various uncertainties in the detection [8]. This research proposes an intelligent robust and efficient model for detecting e-banking phishing websites in order to overcome the "fuzziness" in the evaluation of e-banking, email phishing. The used model uses data mining methods and fuzzy logic to take into account the many elements that make an e-banking scam website effective. Here, two methods are described for correlating data from several DNS servers and numerous suspects in the FF domain. Real-world examples may be utilized to demonstrate how our correlation techniques, which are based on an analytical solution that can quantify multiple DNS queries needed to validate an FF domain, speed up the identification of the FF domain [9].

The publish N subscription association model, also called as LARSID, illustrates how correlation schemes may be implemented on a large scale utilizing a distributed architecture that is more salable than a centralized one. Since that the FF Mother Ship is protected by a proxy screen, it is quite challenging to accurately and quickly identify the FF domains. The number of DNS requests necessary to provide a specific number of distinct IP addresses is calculated as a theoretical approach to the issue of FF detection. Several models are offered for various locations due to the variance

in the allocation of attributes in the various phishing areas [10]. Gaining sufficient information from a new location to recover the detection algorithm and utilize the transfer learning technique to modify the current model is almost impossible. Use of our URL-based technique is a suitable strategy for phishing detection [6]. We must use the transferring approach to create a more effective mode in order to deal with all the conditions for identifying characteristic failure. Comparative study of the classifiers' model-based features is shown in the table1.

III. PROPOSED SYSTEM

We have proposed a novel anti-phishing technique that's projected to assure strong protection is the anti-phishing technique employing deep learning. The study deal both URLs (Universal Resource Locators) and URLs (Universal Resource Identifies) in this method, check them using machine learning, and anticipate whether or not they are phishing websites. Here, a web application for viewing inferred URLs is developed. Every time we visit a website, the associated URL (Universal Resource) is verified using a machine learning algorithm. Users employed logistic regression, Principal Component Analysis and Apriori algorithms techniques to construct our train-out model, with the end, our system chose logistic regression since it provided a much more precise estimate. Phishing is a sort of widespread fraud that occurs when a malicious website poses as a legitimate one with the intention of obtaining sensitive data, such as usernames, account login information, or MasterCard numbers. Phishing is a deception technique that combines social engineering and technology to get private information, such as usernames and credit card numbers, by impersonating a reliable person or organization in a digital communication.

A. Logistic regression

Whenever the response variable (output) is in binary code, including such 0 (False) or 1, logistic regression is employed as a classification algorithm (True). Because of this, logistic regression is an effective technique for your job's goal of determining if a URL is indeed a scam URL (1) or not (0), in the case presented here.

The logistic regression model is expanded upon by logistic regression. Let's use a clear example to better grasp this. If we use a regression analysis model to determine if a message is spam or not, we would only regularly receive values ranging from 0 and 1, such as 0.4, 0.7, etc. The Logistic Regression, in contrast hand, expands this linear regression model by establishing a cutoff at 0.5; as a result, the data point will be classed as spam if the output value is larger than 0.5 and not spamming if the resulting value is less than 0.5. By using Logistic Regression to categorization issues, we may get specific estimates.

Statistical methods first employed the logistic regression, sometimes known as the nonlinear function, to characterize characteristics of population expansion in ecological. The projected values are transformed to probabilities using a mathematical tool called the logistic function. The S-shaped curve of the logistic regression has a range of values around 0 and 1, but just never completely at such boundaries.

$$\text{Logistic}(n) = 1/(1+\exp(-n)) \quad (1)$$

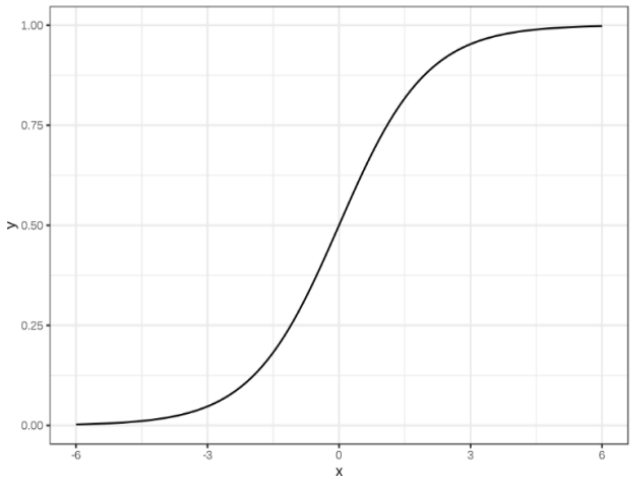

Fig 1: logistic regression

B. Principle Component Analysis

An unsupervised learning approach called principal component analysis is used in machine learning to reduce dimensionality. With the use of orthogonal transformation, it is a statistical procedure that transforms the observations of correlated characteristics into a collection of linearly uncorrelated data. The Principal Components are these newly altered features. One of the widely used tools for exploratory data analysis and predictive modelling is this one. It is a method for identifying significant patterns in the provided dataset by lowering the variances. PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.

The PCA algorithm is based on some mathematical concepts such as:

- Variance and Covariance
- Eigenvalues and Eigen factors

Mathematical equation of PCA: $Z=XS-1$ Z = X S – 1

➢ *Variance and Covariance*

In statistics and probability theory, the mathematical concepts of variance and covariance are often utilized. A covariance is a measure of the directional connection between two random variables, whereas variation is the dispersion of a data set around its mean value.

➢ *Eigenvalues and eigen factors*

Eigenvalues are a unique collection of scalar values connected to a set of linear equations that are most likely seen in matrix equations. The characteristic roots are another name for the eigenvectors. It is a non-zero vector, after applying linear transformations, can only be altered by its scalar component.

➢ *Apriori algorithm*

The Apriori algorithm is made to operate on databases that include transactions and construct association rules using frequent item-sets. It establishes how strongly or weakly two items are associated using these association rules. This approach calculates the item-set associations quickly by using a breadth-first search and a hash tree. Finding the common item sets from a huge dataset is an iterative procedure.

➢ *Apriori Algorithm Procedure:*

The stages for the apriori algorithm are as follows:
Step 1: Establish the minimal support and confidence for item-sets in the transactional database.
Step 2: Take all transaction supports that have a greater support value than the minimum or chosen support value in step two.
Step 3: Track down all the rules in these subgroups with confidence values greater than the threshold or minimum.
Step 4: Arrange the rules in order of lowering lift.

IV. EXPERIMENTS

There are several traits and patterns that might be regarded as elements on fake sites. In this paragraph, we discuss all scamming web page aspects that were employed in the past investigations as feasible. In addition, when we examine the patterns and characteristics of phishing, we discover a few fresh traits that qualify as features. There are 37 scamming characteristics in all, of which 3 are new characteristics. With the characteristics in table I, we divide these in three major categories as follows [9].

→The URL can be used to automatically extract.
→Web page may be used to feature extraction.

Because the goal of a fake website is to collect sensitive data like an e-mail address and a password, we utilize the number of entered emails and the number of input passwords as additional features. The quantity of login or pass code inputs is regarded as a scamming website characteristic. Some other novel characteristic is the number of icons. As we were researching phishing features, we discovered that many scam websites utilize.

A. Types of phishing websites benign:

Synonym of maliciousness, harmless or well-intentioned. In the above fig benign has highest count and the count is 400000.

➢ *Defacement*

Long after the hacker's message has been removed, the damage a defacement assault does to a website's identity and credibility serves as a visible sign that a website has been hacked. In the above fig defacement has a count of 100000.

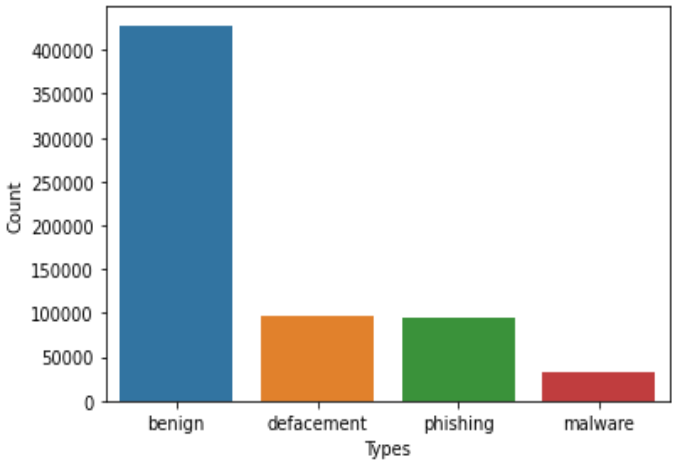

Fig. 2. Different types of attacks

➢ *Phishing*

Phishing means to rob data from users with fake websites which will send by mail to the user. In the above fig phishing has a count of 100000.

➢ *Malware*

Malware is any program that is utilized to gain unauthorized access to IT systems in order to obtain data, obstruct system performance, or otherwise impair IT networks. Hackers that use ransomware encrypt data or hold devices hostage until they are paid or given a ransom are considered to be a subset of malware. In the above fig malware has a count of 50000.

➢ *Abnormal Url*

An extremely lengthy URL is produced by a hacker who attempts to assault the web email system's parsing process since the system might not be designed for handling very long strings. This typically indicates the existence of an overflow of buffers or threat of service attack.

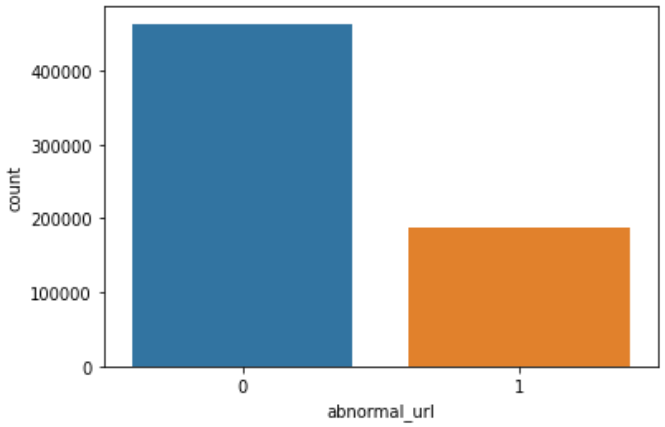

Fig. 3. Abnormal Url

➢ *Https*

The HTTP is a fusion of the SSL with the HTTP. TLS is a popular authentication and security tool for web servers and browsers.

ISSN No:-2456-2165

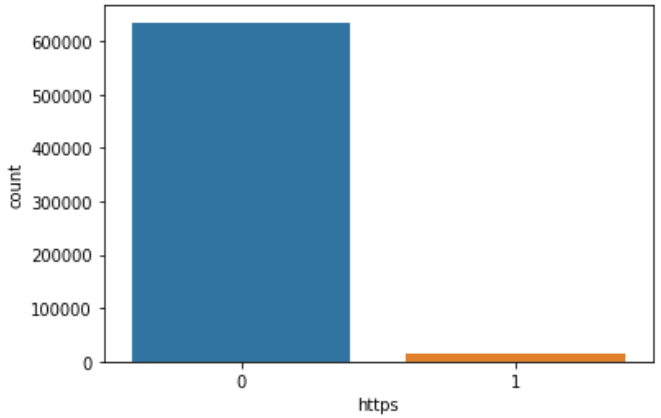

Fig. 4: HTTPS

➢ *Shortening service*

A third-party website known as a URL shortening service changes the lengthy URL into a short, dependent on case numeric code. Simply said, this indicates that a URL shorten service reduces the amount of characters of absurdly lengthy URLs (web addresses).

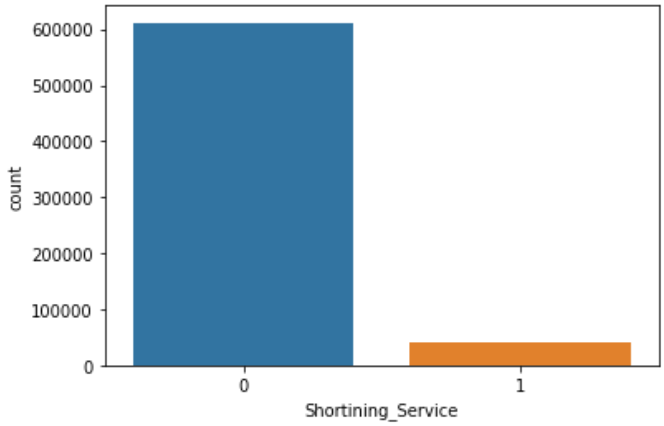

Fig 5. Shortening services

➢ *Confusion Matrices*

A confusion matric is worked to demonstrate the ppresentation of a classification system. The result of an algorithm for classification is presented visually in a confusion matrix.

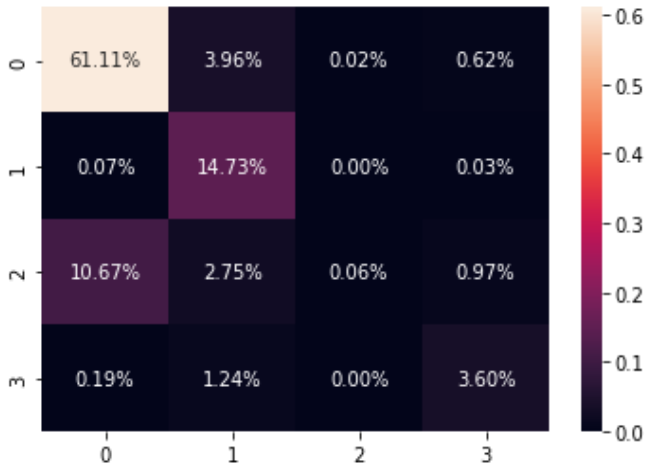

Fig. 6. Confusion Matrix

The total accuracy of the confusion matrix is 91.28%

B. Accuracy of the phishing models

➢ *Decision tree classifier*

A deep learning-supervised tool called a decision tree may be used to categories or predict data based on how previous questions have been resolved. The model goes through supervised learning in that it is trained on a set of data containing the target category before being put to the test on that set of data. In the above fig the accuracy of decision tree classifier is 0.91%

➢ *AdaBoost Classifier*

The boosting technique used by ML ensemble techniques is the AdaBoost algorithm, also known as Adaptive Boosting. Every time, the weights are redistributed, with samples that were wrongly categorized obtaining higher weights—hence, the phrase "adaptive boosting". In the above fig the accuracy of adaboost classifier is 0.82%.

➢ *K-Neighbor Classifier*

The K-Neighbors Classifier looks for the five nearest neighbours. The classifier has to be explicitly told to use Euclidean distance to calculate how close neighboring points are to one another. Using our recently learned model, we assess the benignity of a tumor based on its average compactness and area. In the above fig the accuracy of K-neighbors is 0.89%.

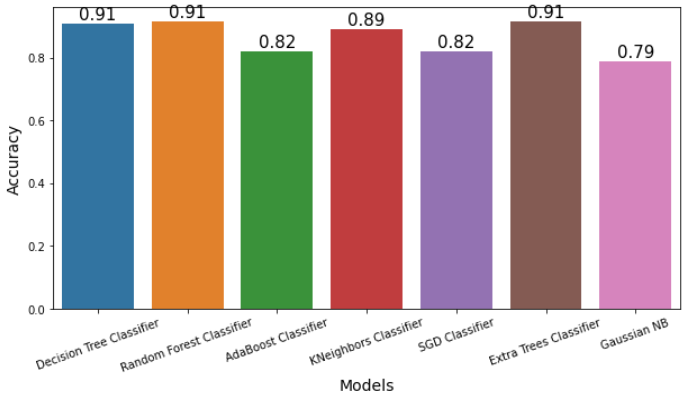

Fig. 7: Accuracy of models

➢ *SGD classifier*

SGD classifier. Essentially, the SGD classifier employs a simple SGD learning technique that supports a variety of categorization loss equations and penalties. Sci kit Learn provides the Classifier module to implement SGD classification. In the above fig the accuracy is 0.82%.

➢ *Technique*

We examine every conceivable pairing of the 36 features in order to identify the best and poorest traits as well as to eliminate any. Unnecessary features. This equation may be used to determine the length of any mixture:

$$\Sigma = \frac{n!}{k!(n-k)!} \quad (2)$$

The k is the number of attributes that range from 1 to 36 that were selected. n is the number of characteristics, which really is 36. The research will be condensed into obtaining the greatest and the lowest outcome for every k combo because there are an enormous number of conceivable combinations. As in conclusion, the optimum combination will be the greatest correctness also with fewest characteristics. It explains the method for selecting features in Fig. 1.

This system's major purpose is to determine if a web page is real or a phishing site and classify it accordingly. The algorithm depicted in Fig. 2 can be used to carry out this purpose. Every time the person visits a new web page, this algorithm is activated. Its job is to retrieve the web page characteristics using the URL and DOM objects. the URL that was utilized to retrieve the characteristics of the URL and site rank. Whereas the DOM, which is a link among scripts, was developed to remove the characteristic of the table.

- *Training and testing of model*

Process of teaching methods with a portion of the Information and evaluating how well They do in accurately categorizing the datasets are referred to simply as the "training and Evaluation model." Use the logistic regression which is loaded by using classifier from the Sci-Kit module in order to build the datasets.

- *Regression Analysis*

The different URLs may now be provided to the training sample as inputs. It makes a prediction about the quality of the Address and outputs good or terrible. First, load the Logistic Regression package, then use the Logistic Regression Method to build a Logistic Regression classification instance.

Table 1. Models and its accuracy:

	Model	Accuracy
1	Decision Tree Classifier	0.909528
2	Random Forest Classifier	0.914749
3	AdaBoost Classifier	0.820077
4	K-Neighbors Classifier	0.890409
5	SGD Classifier	0.820591
6	Extra Trees Classifier	0.914672
7	Gaussian NB	0.789548

V. CONCLUSION

Because we utilize the Internet more often in our everyday lives, cybercriminals target their victims through this medium. One of the most common attacks is "phishing," which involves creating a faked website to steal customers' private data, including their user-ID and password, from financial websites utilizing social media tools. The malicious website is made to seem exactly like a legal website, even down to replicating the original website word for word. Due to the sentence meaning of such pages, which exploits human weaknesses, their sensing therefore represents a very simple issue to address. Software packages are only effective as a support mechanism for this sort of threat identification and avoidance, and they employ a whitelist/blacklist strategy in particular to thwart such attacks. These are fixed methods, though, therefore they are unable to detect newly introduced assaults on the system. As a result, we suggest using a logistic regression computer vision system to recognize incoming URLs as an effective option. According to the research observations, this technique produces an acceptable effectiveness rate of roughly 98%.

REFERENCES

[1]. R. B. Basnet, A. H. Sung, "Mining web to detect phishing URLs", Proceedings of the International Conference on Machine Learning and Applications, vol. 1, pp. 568-573, Dec 2012.
[2]. Abdelhamid N., Thabtah F., Ayesh A. (2014) Phishing detection based associative classification data mining. Expert systems with Applications Journal. 41 (2014) 5948-5959.
[3]. Mohammad, R. M., Thabtah, F. & McCluskey, L. (2013) Predicting Phishing Websites using Neural Network trained with Back Propagation. Las Vigas, World Congress in Computer Science, Computer Engineering, and Applied Computing, pp. 682-686.
[4]. Aburrous M.., Hossain M., Dahal K.P. and Thabtah F. (2010) Experimental Case Studies for Investigating E-Banking Phishing Techniques and Attack Strategies. Journal of Cognitive Computation, Springer Verlag, 2 (3): 242-253.
[5]. Mohammad R., Thabtah F., McCluskey L., (2014B) Intelligent Rule based Phishing Websites Classification. Journal of Information Security (2), 1-17. ISSN 17518709. IET.
[6]. Jain, Ankit Kumar, and B. B. Gupta. "Comparative analysis of features-based machine learning approaches for phishing detection." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. IEEE, 2016, pp. 2125-2130.
[7]. R.Aravindhan, Dr.R.Shanmugalakshmi, Certain Investigation on Web Application Security: Phishing Detection and Phishing Target Discovery, January 2016.
[8]. L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, “A novel approach for phishing detection using URL-based heuristic,” 2014 Int. Conf. Compute. Manage. Telecommun. ComManTel 2014, pp. 298–303, 2014.
[9]. A. Berthold, et al., ”Improved phishing detection using model-based features,” in Proc. Conference on Email and Anti-Spam (CEAS). Mountain View Conf, CA, aug 2008
[10]. L. Ma, et al.,”Detecting phishing emails using hybrid features,”IEEE Conf, 2009, pp. 493-497