Journal of Business and Information Systems
Vol. 4, No. 1, June (2022)
www.thejbis.org
DOI:10.36067/jbis.v4i1.120
ISSN:2685-2543

Accepted, May 2022
Revised, May 2022
Published, June 2022

A comparison study of machine learning techniques for phishing detection

Jathin Kolla
SCOPE, VIT-AP University, Amaravati, India
jathinkolla@gmail.com

Shinde Praneeth
SCOPE, VIT-AP University, Amaravati, India
shindepraneeth@gmail.com

Mirza Sameed Baig
SCOPE, VIT-AP University, Amaravati, India
mirzasameedbaig@gmail.com

Ganesh Reddy Karri
SCOPE, VIT-AP University, Amaravati, India
ganesh.reddy@vitap.ac.in

Abstract: *In the last few years phishing attacks have been increasing eventually. As the internet is developing, security for it is becoming a challenging task. Cyber-attacks and threats are increasing rapidly. These days many fake websites are created to deceive victims by collecting their login credentials, bank details, etc. Many anti phishing products are launched into the market and use blacklists, heuristics, visual and machine learning-based approaches, these products cannot prevent all the phishing attacks. However, unlike predicting phishing URLs, there are only few studies that compare machine learning techniques in predicting phishing. The present study compares the predictive accuracy of several machine learning methods including Decision tree, Random Forest, Multilayer Perceptions, Support Vector Machines and XGBoost for predicting phishing URLs. The results showed that random forest is the best model among the others and has the highest accuracy.*

Keywords: Phishing; Machine learning; Cyber attacks; Cyber security.

1. Background

When we talk about digitization the first thing that pops into our mind is the internet. The Internet is a network that computes all the information one could ever want. It is a vast network that connects all the computers across the world. Through the internet one could share information, pictures and make phone calls from anywhere across the world. Apart from all the advantages we get from the internet, there are some serious crimes taking place through the internet. One such crime is cybercrime with Phishing taking a major part in it.

Phishing is a type of social engineering attack often used to steal user data, including login credentials and credit card numbers. It occurs when an attacker masquerades as a trusted entity and dupes a victim into opening an email. The recipient is then tricked to open a malicious link, which can lead to the installation of malware, the freezing of the system as a part of a ransomware attack of the revealing of sensitive information.

Phishing attacks can also be done through phone calls or text messages, but emails are the most used tricks to lure. The word Phishing is derived from ‘fishing’ which refers to the victims, it has been over a decade since this type of attack has gained a lot of attention from researchers. Phishing is one of the most promising methods for attackers to lure people to believe it is a legitimate message (Ollmann, 2007). Phishing attacks have been growing over the years globally with an increase of 65% to a value of $ 1,220,523 in 2016 as compared to the previous year (APWG 2017) also APWG reported an increase of 5753% of average Phishing attacks per month over the period of 12 years, from 2004 to 2016 (Lininger & Vines, 2005). Over half a billion personal records were stolen in 2015, According to Kaspersky lab, Phishing in the financial sector has reached an all-time high in 2016. From 2013-to 2018, the FBI has reportedly lost $2.3 Billion over fake email scams.

The attackers are getting more innovative by introducing new Phishing methods, they are:

Spear phishing attack: A spear-phishing attack is a targeted attempt to steal sensitive information such as account credentials or financial information. The attackers then disguise themselves as trustworthy friends or entities to acquire sensitive information through mail or online messages. Whaling attack: A whaling attack is a method used by cybercriminals to masquerade as a senior player at an organization and directly target junior officials at an organization. Spoofing: Spoofing is the act of disguising a communication from an unknown source as being from a knowledgeable, trusted source. Spoofing can apply to emails, phone calls, and websites, or can be more technical such as a computer spoofing an IP address (OS) DNS server. Smishing: Smishing is a type of phishing that takes place via short message service (SMS) messages — otherwise known as the text messages that are received on phone through the cellular carrier.

Phishing campaigns can be difficult to spot. Cybercriminals have become experts at using sophisticated techniques to trick victims into sharing personal or financial information. But the best way to protect yourself is to learn how to spot a phishing scam before you take the bait. There is an array of methods developed by researchers to control phishing attacks but can't be guaranteed to detect 100% of attacks (APWG, 2013).

2. Theoretical Framework and Hypothesis Development

2.1. Phishing

In many ways, phishing hasn't changed a lot since its AOL heyday. In 2001, however, phishers turned their attention to online payment systems. Although the first attack, which was on E-Gold in June 2001, was not considered to be successful, it planted an important seed. In late 2003, phishers registered dozens of domains that looked like legitimate sites like eBay and PayPal if you weren't paying attention. They used email worm programs to send out spoofed emails to PayPal customers. Those customers were led to spoofed sites and asked to update their credit card details and other identifying information. By the beginning of 2004, phishers were riding a huge wave of success that included attacks on banking sites and their customers. Popup windows were used to acquire sensitive information from victims. Between May 2004 and May 2005, about 1.2 million users in the U.S. suffered losses caused by phishing, totaling approximately $929 million. Organizations lose about $2 billion per year to phishing. Phishing is officially recognized as a fully organized part of the black market. Specialized software emerges on a global scale that can handle phishing payments, which in turn outsources a huge risk. The software is implemented into phishing campaigns by organized crime gangs. In late 2008, Bitcoin and other cryptocurrencies were launched. This allows transactions using malicious software to be secure and anonymous, changing the game for cybercriminals. Table 1 shows growth rate of phishing starts from 1996 to 2014. According to the APWG report, the total number of unique phishing websites detected was 125,215 in the first quarter of 2014, which has increased approximately by 11 % in the last quarter of 2013 (APWG, 2014).

Table 1. Evolution of phishing during 1996–2015

Year	Extension
1996	Term ‘‘phishing’’ was first used
1997	Media declared the evolution of a new attack called ‘‘phishing’’
1998	Attackers started using message and newsgroups
1999	Use of mass mailing to escalate the phishing attacks
2000	First use of keyloggers, phishers used it for getting login credentials
2001	Use of URLs to direct victim to a fake site
2002	Use of screen loggers
2003	Use of IM and IRC
2004	Evolution of ‘‘pharming’’
2005	Term ‘‘spear phishing’’ was first used
2006	First phishing over VoIP
2007	More than $3 billion lost to phishing scams
2009	Symantec Hosted Services blocked phishing attacks impersonating 1079 different organizations.
2010	Facebook attracted more phishing attacks than Google and IRS
2012	6 million unique malware samples were identified
2013	Red October operation attacked more than 69 countries
2014	750,000 malicious emails were sent using IoT devices, i.e., refrigerators and smart TVs
2015	Spear phishing reached its peak in manufacturing and wholesale industries

2.2. Learning Methods for Phishing Detection:

From the dataset it is clear that this is a supervised machine learning task. There are two major types of supervised machine learning problems, called classification and regression. This data set comes under classification problems, as the input URL is classified as phishing (1) or legitimate (0). The supervised machine learning models (classification) considered to train the dataset in this notebook are: Decision Tree; Random Forest; Multilayer Perceptron’s; XGBoost; Support Vector Machines.

2.2.1. Decision Tree Classifier

Decision trees are widely used models for classification and regression tasks. Essentially, they learn a hierarchy of if/else questions, leading to a decision. Learning a decision tree means learning the sequence of if/else questions that gets us to the true answer most quickly. In the machine learning setting, these questions are called tests (not to be confused with the test set, which is the data we use to test to see how generalizable our model is). To build a tree, the algorithm searches over all possible tests and finds the one that is most informative about the target variable.

2.2.2. Random Forest Classifier

Random forests for regression and classification are currently among the most widely used machine learning methods. A random forest is essentially a collection of decision trees, where each tree is slightly different from the others. The idea behind random forests is that each tree might do a relatively good job of predicting, but will likely be overfit on part of the data. If we build many trees, all of which work well and overfit in different ways, we can reduce the amount of overfitting by averaging their results. To build a random forest model, you need to decide on the number of trees to build (the n_estimators’ parameter of Random Forest Regressor or Random Forest Classifier). They are very powerful, often work well without heavy tuning of the parameters, and don’t require scaling of the data.

2.2.3. Multilayer Perceptron’s (MLPs)

Multilayer perceptron’s (MLPs) are also known as (vanilla) feed-forward neural networks, or sometimes just neural networks. Multilayer perceptron’s can be applied for both classification and regression problems. A multilayer perceptron is a neural network connecting multiple layers in a directed graph, which means that the signal path through the nodes only goes one way. Each node, apart from the input nodes, has a nonlinear activation function. An MLP uses backpropagation as a supervised learning technique. Since there are multiple layers of neurons, MLP is a deep learning technique. MLPs can be viewed as generalizations of linear models that perform multiple stages of processing to come to a decision.

2.2.4. XGBoost Classifier

XGBoost is one of the most popular machine learning algorithms these days. XGBoost stands for extreme Gradient Boosting. Regardless of the type of prediction task at hand; regression or classification. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

2.2.5. Support Vector Machines

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

3. Methods

In this section we explain how we create a set of test data from criminal URL theft to sensitive information. In addition, we describe the analytical metrics we use in comparison. Finally, we describe the initial setup of the test.

3.1 Dataset

The dataset file has a list of URLs. The source of the dataset is taken from University of New Brunswick. One of the challenges faced by our research was the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using data mining techniques have been disseminated these days, no reliable training dataset has been published publicly, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features.

In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we proposed some new features, experimentally assigned new rules to some well-known features, and updated some other features.

3.2. Feature Selection

3.2.1. IP Address in the URL

Checks for the presence of an IP address in the URL. URLs may have IP address instead of domain name. If an IP address is used as an alternative of the domain name in the URL, we can be sure that someone is trying to steal personal information with this URL.

Rule	Feature value
If the domain part of the URL has an IP address	1(phishing)
Else	0(legitimate)

3.2.2. URL's having "@" Symbol

Checks for the presence of '@' symbol in the URL. Using "@" symbol in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol.

Rule	Feature value
If the URL has '@' symbol	1(phishing)
Else	0(legitimate)

3.2.3. Length of URL

Computes the length of the URL. Phishers can use long URL to hide the doubtful part in the address bar. In this project, if the length of the URL is greater than or equal 54 characters then the URL classified as phishing otherwise legitimate.

Rule	Feature value
If (Length of URL >= 54)	1(phishing)
Else	0(legitimate)

3.2.4. Redirecting Using “//”

The existence of “//” within the URL path means that the user will be redirected to another website. We examine the location where the “//” appears. We find that if the URL starts with “HTTP”, that means the “//” should appear in the sixth position. However, if the URL employs “HTTPS" then the “//” should appear in seventh position.

Rule	Feature value
If the URL has '//' symbol,	1 (phishing)
Else	0 (legitimate)

3.2.5. HTTPS (Hypertext Transfer Protocol with Secure Sockets Layer)

The existence of HTTPS is very important in giving the impression of website legitimacy, but this is clearly not enough. The authors in Mohammad, Thabtah, & Mccluskey (2013) suggest checking the certificate assigned with HTTPS including the extent of the trust certificate issuer, and the certificate age. Certificate Authorities that are consistently listed among the top trustworthy names include: “GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster and VeriSign”. Furthermore, by testing out our datasets, we find that the minimum age of a reputable certificate is two years.

Rule	Feature value
IF (Use https and issuer is trusted and age of certificate≥ 1 Year)	1 (phishing)
Else	0 (legitimate)

3.2.6. Using URL Shortening Services “Tiny URL”

URL shortening is a method on the “World Wide Web” in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an “HTTP Redirect” on a domain name that is short, which links to the webpage that has a long URL. For example, the URL “http://portal.hud.ac.uk/ ” can be shortened to “bit.ly/19DXSk4”.

Rule	Feature value
If tiny URL	1 (phishing)
Else	0 (legitimate)

3.2.7. Prefix or Suffix Separated by (-) to the Domain

The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate website. For example, http://www.Confirme-paypal.com/.

Rule	Feature value
IF (Domain Name Part Includes (-)	1 (Phishing)
Else	0 (legitimate)

3.2.8. DNS Record

For phishing websites, either the claimed identity is not recognized by the WHOIS database or no records found for the hostname.

Rule	Feature value
If the DNS record is empty or not found, then	1 (Phishing)
Else	0 (legitimate)

3.2.9. Web Traffic

This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites live for a short period of time, they may not be recognized by the Alexa database. By reviewing our dataset, we find that in the worst scenarios, legitimate websites ranked among the top 100,000. Furthermore, if the domain has no traffic or is not recognized by the Alexa database, it is classified as “Phishing”.

Rule	Feature value
If the rank of the domain < 100000	1 (Phishing)
Else	0 (legitimate)

3.2.10. Domain Age

Based on the fact that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.

Rule	Feature value
IF (Domains Expires on≤ 1 years)	1 (Phishing)
else	0 (legitimate)

3.2.11. End Period of Domain

This feature can be extracted from the WHOIS database. For this feature, the remaining domain time is calculated by finding the difference between expiration time & current time. The end period considered for the legitimate domain is 6 months or less for this project.

Rule	Feature value
If (end period of the domain is > 6 months)	1 (Phishing)
else	0 (legitimate)

3.2.12. IFrame Redirection

IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the “iframe” tag and make it invisible i.e. without frame borders. In this regard, phishers make use of the “frameborder” attribute which causes the browser to render a visual delineation.

Rule	Feature value
IF (Using iframe)	1 (Phishing)
else	0 (legitimate)

3.2.13. Status Bar Customization

Phishers may use JavaScript to show a fake URL in the status bar to users. To extract this feature, we must dig-out the webpage source code, particularly the “on Mouseover” event, and check if it makes any changes on the status bar. If the response is empty or on mouseover is found then, the value assigned to this feature is 1 (phishing) or else 0 (legitimate).

Rule	Feature value
If (response is empty or on mouse over is found then)	1 (Phishing)
else	0 (legitimate)

3.2.14. Disabling Right Click

Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code. This feature is treated exactly as “Using on Mouseover to hide the Link”. Nonetheless, for this feature, we will search for the event “event.button==2” in the webpage source code and check if the right click is disabled.

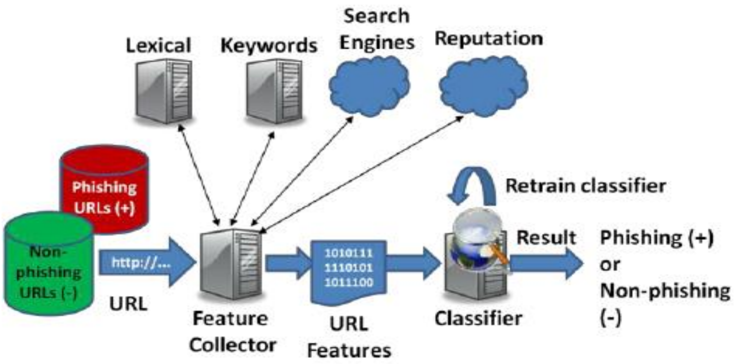
Rule	Feature value
IF (“Using on Mouse over to hide the Link”.)	1 (Phishing)
else	0 (legitimate)

3.2.15. Website Forwarding

The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. In our dataset, we find that legitimate websites have been redirected one time max. On the other hand, phishing websites containing this feature have been redirected at least 4 times.

Rule	Feature value
If (website forwarding is at least 4 times**)**	1 (Phishing)
else	0 (legitimate)

Figure 1. Phising Model



3.3. Evaluation metrics:

a. True Positive (TP): This denotes the ratio of the number of phishing emails identified correctly as:

a. $$\mathrm{TP} = \frac{n_{\mathrm{p}} \rightarrow P}{N_{\mathrm{P}}}.$$

b. True Negative (TN): This denotes the ratio of the number of ham emails identified correctly as:

a. $$\mathrm{TN} = \frac{n_{\mathrm{h}} \rightarrow H}{N_{\mathrm{H}}}.$$

c. False Positive (FP): This denotes the ratio of the number of ham emails classified as phishing, as:

a. $$\mathrm{FP} = \frac{n_{\mathrm{h}} \rightarrow P}{N_{\mathrm{H}}}.$$

d. False Negative (FN): Ratio denoting the number of phishing emails classified as ham, as:

a. $$\mathrm{FN} = \frac{n_{\mathrm{p}} \rightarrow H}{N_{\mathrm{P}}}.$$

e. Precision (P): Measures the rate of phishing emails which are identified as the emails detected as phishing:

a. $$p = \frac{n_{\mathrm{h}} \rightarrow P}{n_{\mathrm{p}} \rightarrow P + n_{\mathrm{h}} \rightarrow P}.$$

f. Recall (r): Measures the rate of phishing emails which are identified correctly as existing phishing emails:

a. $$r = \frac{n_{\mathrm{p}} \rightarrow P}{n_{\mathrm{p}} \rightarrow P + n_{\mathrm{p}} \rightarrow H}.$$

g. *f1* Score: This is the harmonic mean of Precision and Recall:

a. $$f_1 = \frac{2p.r}{p + r}.$$

h. Accuracy (ACC): Measures overall correctly identified emails:

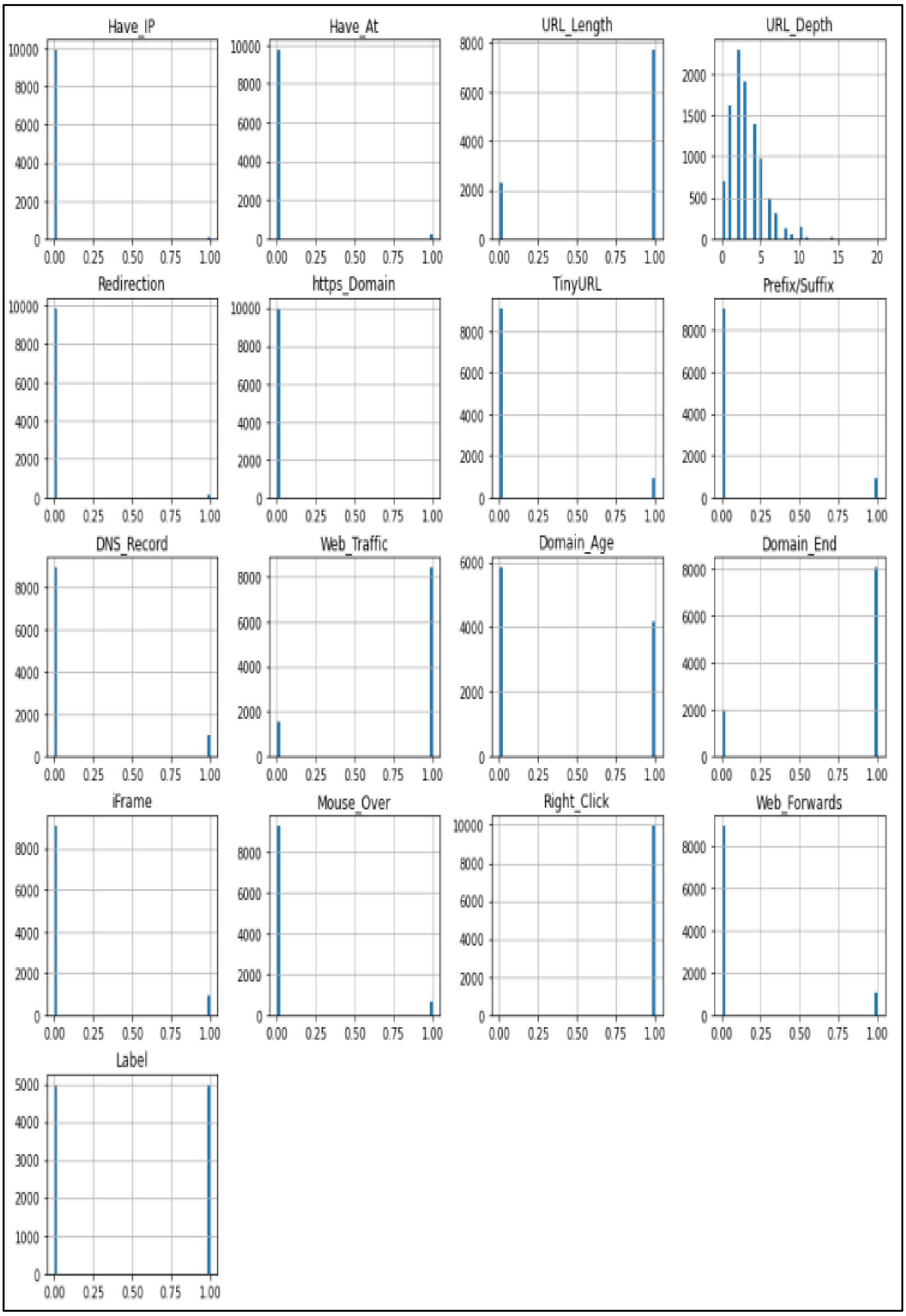
a. $$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}.$$

i. Specificity (S): Measures correctly identified ham emails:

$$S = \frac{\mathrm{TP}}{\mathrm{TN} + \mathrm{FP}}.$$

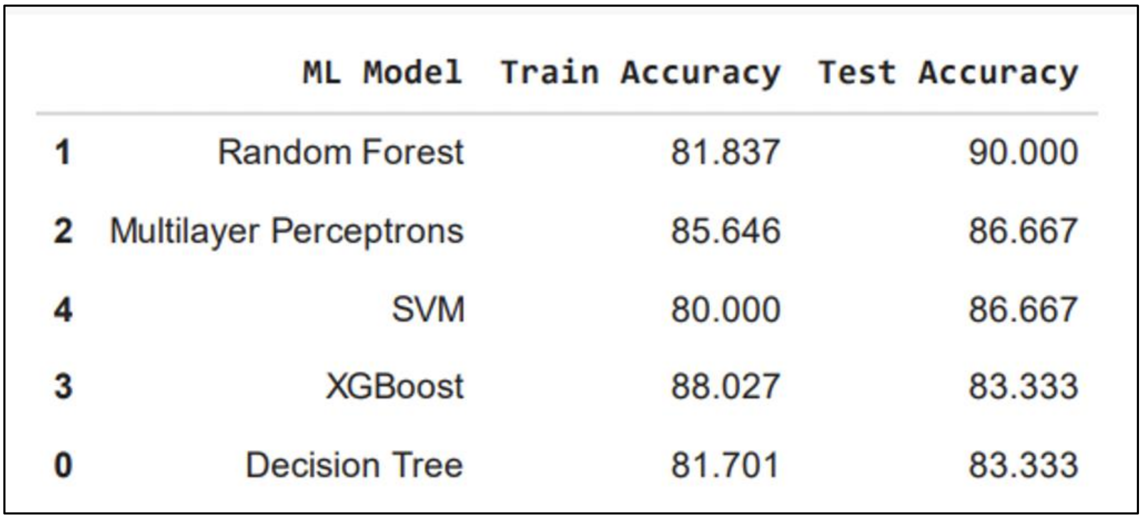
4. Results

Figure 2. Feature Distribution



This is a supervised machine learning task. There are two major types of supervised machine learning problems, called classification and regression. This data set comes under classification problems, as the input URL is classified as phishing (1) or legitimate (0).

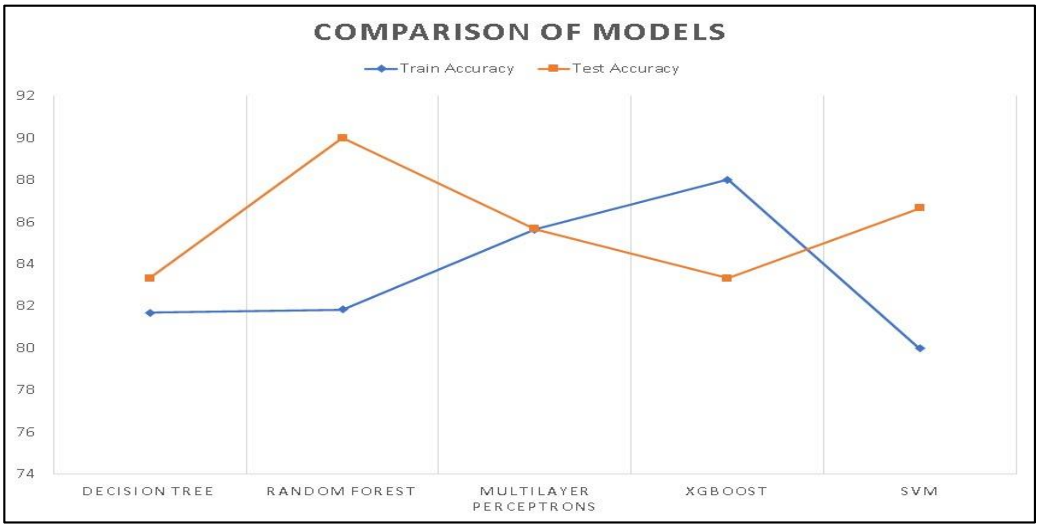
Figure 3. The accuracy comparison of machine learning models



	ML Model	Train Accuracy	Test Accuracy
1	Random Forest	81.837	90.000
2	Multilayer Perceptrons	85.646	86.667
4	SVM	80.000	86.667
3	XGBoost	88.027	83.333
0	Decision Tree	81.701	83.333

From the obtained results of the above models, Random Forest has the highest model performance of 90.0%.

Figure 4. The comparison of models



6. Conclusion

In this paper, we have developed a system for detecting identity theft by using five different machine learning algorithms, such as Decision Tree, Random Forest, Multilayer Perceptron's, SVM, XGBoost and various numbers / types of features such as Have_IP, Have_At, URL_length, URL_Depth, Redirection, https_Domain, TinyURL, Prefix / Suffix, DN _Record, Web_Traffic, Domain_Age, Domain_End, Frame, Mouse_Over, Right_Click, Web_Forwards, Label. To increase the accuracy of the acquisition system, the construction of an active feature list is an important task. Therefore, we compiled our feature list into two separate categories such as NLP-based features, which are highly determined by people and word vectors, focusing on the use of words in the URL without performing any other functions. Due to the absence of a worldwide acceptable test set for phishing systems, we needed to construct our own dataset with 9,375 URLs. This set contains 5,653 legitimate URLs and 3,722 phishing URLs.

7. References

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. *ACM International Conference Proceeding Series*, *269*, 60–69. https://doi.org/10.1145/1299015.1299021.

Aloul, F. A. (2012). The need for effective information security awareness. *Journal of Advances in Information Technology*, *3*(3), 176–183. https://doi.org/10.4304/jait.3.3.176-183

Anti-Phishing Working Group (APWG), ‘‘Phishing activity trends report—first quarter 2013. http://antiphishing.org/reports/apwgtrendsreportq12013.pdf

Anti-Phishing Working Group (APWG) (2014) Phishing activity trends report—first quarter 2014. http://antiphishing.org/reports/apwgtrendsreportq12014.pdf

Anti-Phishing Working Group (APWG) (2014) Phishing activity trends report—fourth quarter 2013. http://antiphishing.org/reports/apwgtrendsreportq42013.pdf

Anti-Phishing Working Group APWG (2017), Phishing activity trends report—fourth quarter http://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf

Babagoli, M., Aghababa, M. P., & Solouk, V. (2018). Heuristic nonlinear regression strategy for detecting phishing websites. *Soft Computing*, 23(12), 1-13.

Buber, E., Diri, B. & Sahingoz, O. K., (2017a). Detecting phishing attacks from URL by using NLP techniques, *International Conference on Computer Science and Engineering* (UBMK), 337–342.28

Buber, E., Diri, B., & Sahingoz, O. K. (2017b). NLP based phishing attack detection from URLs, in: A. Abraham, P. K. Muhuri, A. K. Muda, N. Gandhi (Eds.), *Intelligent Systems Design and Applications*, Springer International Publishing, Cham, 608–618.

Husna, H., Phithakkitnukoon, S., & Dantu, R. (2008) Behavior analysis of spam botnets, *Communication systems software and middleware and workshops. COMSWARE 2008. 3rd International Conference*, Bangalore, India. pp 246–253.

Levine, J. (2008). DNS blacklists and whitelists, IRTF anti-spam research group, Internet Draft draft-irtf-asrg-dnsbl-08.txt

Lininger, R., & Vines, R. D. (2005) Phishing: Cutting the Identity Theft Line Published by Wiley Publishing, Inc. 10475 Crosspoint Boulevard Indianapolis, IN46256

McAfee, Inc. (n. d.) McAfee Site Advisor. [Online] Available at: http://www.siteadvisor.com/. [Accessed: January 11, 2016].

Mohammad, R. M., Thabtah, F., & Mccluskey, L. (2015A). Tutorial and critical analysis of phishing websites methods, *Computer Science Review Journal*. 17, 1-24.

Mohammad, R. M., Thabtah, F., & Mccluskey, L. (2015B). Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites

Mohammad, R. M., Thabtah, F., & Mccluskey, L. (2014A) Predicting phishing websites based on self-structuring neural network, *Journal of Neural Computing and Applications*, 25(2), 443-458.

Mohammad, R. M., Thabtah, F., & Mccluskey, L (2014B) Intelligent rule based phishing websites classification. *Journal of Information Security* (2), 1-17.

Mohammad, R. M., Thabtah, F., & Mccluskey, L. (2013). Predicting Phishing Websites using Neural Network trained with Back-Propagation. In *World Congress in Computer Science, Computer Engineering, and Applied Computing* (pp. 682–686).

Netcraft Inc. (n. d.) Netcraft Anti-Phishing Toolbar. [Online] Available at: http://toolbar.netcraft.com/. [Accessed May 9th 2016].

Netscape Communications (n. d.) [Online] Available at: netscapenavigator.soft32.com. [Accessed May 8th 2016].

Ollmann, G. (2007). The Phishing Guide Understanding & Preventing Phishing Attacks, IBM Internet Security Systems.

Platt J. (1998). Fast training of SVM using sequential optimization, (Advances in kernel methods ± support vector learning, B. Scholkopf, C. Burges, A. Smola eds), MIT Press, Cambridge, 1998, 185-208.

Toolan, F., & Carthy, J. (2009). Phishing detection using classifier ensembles, *eCrime researchers summit, IEEE conference Tacoma*. WA, USA, pp 1–9

www.phishtank.com/

www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/

www.investors.proofpoint.com/releasedetail.cfm?releaseid=819799