Phishing Diagnosis: A Multi-Feature Decision Tree-based Method

Pravin Kumar Pandey, Sandip Kumar Singh

***Abstract:** Phishing is an electronically connected criminal activity in which the attacker steals the user's personal information like username, countersign, internet banking account, credit/debit card number with the expiration date, password, pin, legitimacy, confidential patient record, CVV number, etc. to boon financially. Email-based phishing is the most common and traditional way of phishing scams, in which the phisher will send a suspicious email with an embedded URL and ask the user to click the URL. When the user clicks on the link, the link will be redirected to a spoofed site that looks the same to the original site to steal their credentials and displays some error message. Later the phishing uses those credentials for malicious purposes. To overcome these scams, many anti-phishing tools have developed. Among that the machine learning-based approaches can give a better result. This paper is an extensive study of the various machine learning-based anti-phishing approaches and their results that detect the phishing URL's from the URLs with URLs features. Six most important models of machine learning have been examined for the phishing detection problem. The Decision Tree-based method outperforms other methods.*

***Keywords:** Phishing, Anti-phishing, Machine learning, Phish tank, Legitimate, Suspicious, Decision Tree.*

I. INTRODUCTION

Phishing is a wide term used to describe a group of scam people with their personal information shared such as consumer name, password, credit/debit card number, etc., that manipulate information for disseminating rea-sons. Earliest contact is sent to a bulky group of people at once, so anyone can be a victim. They will contact their victims with the help of URLs, social media, emails, and phones. The only target through this attack of these people is to send a fake correspondence, which appears to have originated from the actual organization, hoping that a large group will follow the links provided to them from these contacts and disclose their personal information to the phishers (Figure 1). Phishing is an automated detection method used to cheat billions of dollars to out-siders and phishing technology uses human nature as well as the power of the internet to deceive millions of people in the world [1]. The social media platforms are used for deceitful, cultivated and perceptive information from internet users by covering through a legitimate entity. The basic goal of phishing technology is to illegally commit deceitful financial transactions on behalf of internet users [2]. According to the anti-phishing working group (APWG), which is an NGO community (a non-profitable group) report issued on June 2017, the global phishing survey 2016 has shown all the phishing attacks from 2012-2016 (Figure 2) [3]. The anti-phishing working group (APWG), which has also reported on the 1st quarter of 2019 (January, February, March) that there were 180,768 phishing incidents detected [4]. Various methodologies are being adopted at present to identify phishing web sites and emails. Sajid Yousuf Bhat *et al.* proposes an approach for "Spammer classification using ensemble methods over structural social network features" [5]. In [5] finds out whether the URL is spam/legitimate on the social network with community-based features. Mouad Zouina *et al*. proposes an approach for "A novel lightweight URL phishing detection using SVM and similarities index" [6]. In [6] phishing detection from the URL with the help of 6 features. SVM and similarity index is targeted to improve overall recognition of the phishing detection system. Alejandro Correa *et al.* explore "Classifying phishing URLs using recurrent neural networks" [7]. In [7] we explored the use of URLs as input for machine learning models applied for phishing site prediction with the help of 14 features. Suh *et al.* used "Comparing writing style feature-based classification methods for estimating user reputations in social media" it evaluates the performance of classifiers depend on the state-of-art methods 4 writing style features such as lexical, syntactic, structural, content-specific [8]. Gunikhan Sonowal *et al.* using "Masphid: a Model to Assist Screen Reader Users for Detecting Phishing Sites Using Aural and Visual Similarity measures" [9]. In [9] URL is phishing, suspicious or legitimate based on the 10 features. Kshitij Tayal *et al* explore "Particle swarm optimization trained class association rule mining: Application to phishing detection" [10]. In [10] class association rules used to detect the URLs are phished or legitimate.

1.1 Motivation for phishing

1.1.1 Financial gain: E-mail spoofing is also called spear phishing which is used for fraud that targets a specific organization, seeking unauthorized access to confidential data.

1.1.2 Theft the login credentials: Typically, the networked login credentials of prominent high-street banking organizations and successive access to funds ready to transfer and seize the home address mobile, number and other personal information.

1.1.3 Theft the bank credentials: More recently, the increase in networked share trading businesses have portended that a customer's trading attention gives an uncomplicated direction for global money transfers.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author

Pravin Kumar Pandey, Department of Computer Science & Engineering, UNSIET, VBS Purvanchal University, Jaunpur, UP, 211001
Email: pravin108786@gmail.com
Sandip Kumar Singh, Department of Mechanical Engineering, UNSIET, VBS Purvanchal University, Jaunpur, UP, 211001
Email: sandipkumarsingh25@gmail.com

1.1.4 Malware distribution: Although outdated, but malware families are still used to spreading malware and infecting online user computers with the help of email attachments. This type of infection depends on the consumer pressing on the email attachment. A present methodology that uses email as a dispersal implement is to put links to malicious websites.

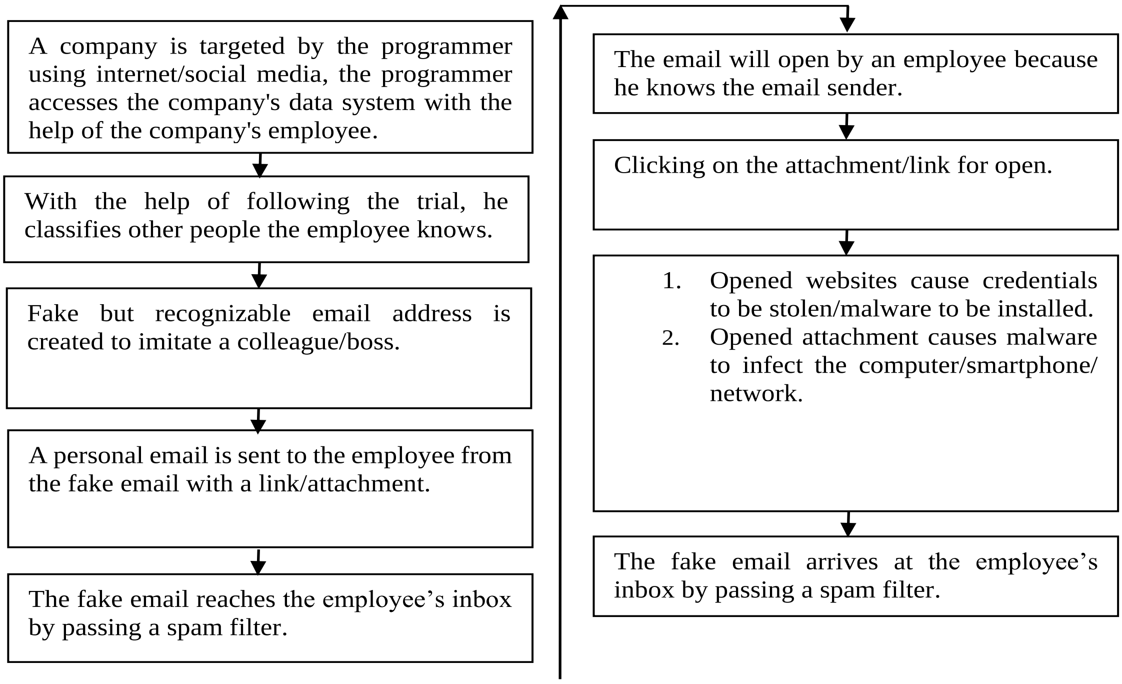

A company is targeted by the programmer using internet/social media, the programmer accesses the company's data system with the help of the company's employee.
With the help of following the trial, he classifies other people the employee knows.
Fake but recognizable email address is created to imitate a colleague/boss.
A personal email is sent to the employee from the fake email with a link/attachment.
The fake email reaches the employee's inbox by passing a spam filter.
The email will open by an employee because he knows the email sender.
Clicking on the attachment/link for open.
1. Opened websites cause credentials to be stolen/malware to be installed.
2. Opened attachment causes malware to infect the computer/smartphone/ network.
The fake email arrives at the employee's inbox by passing a spam filter.

Figure 1: Phishing working procedure

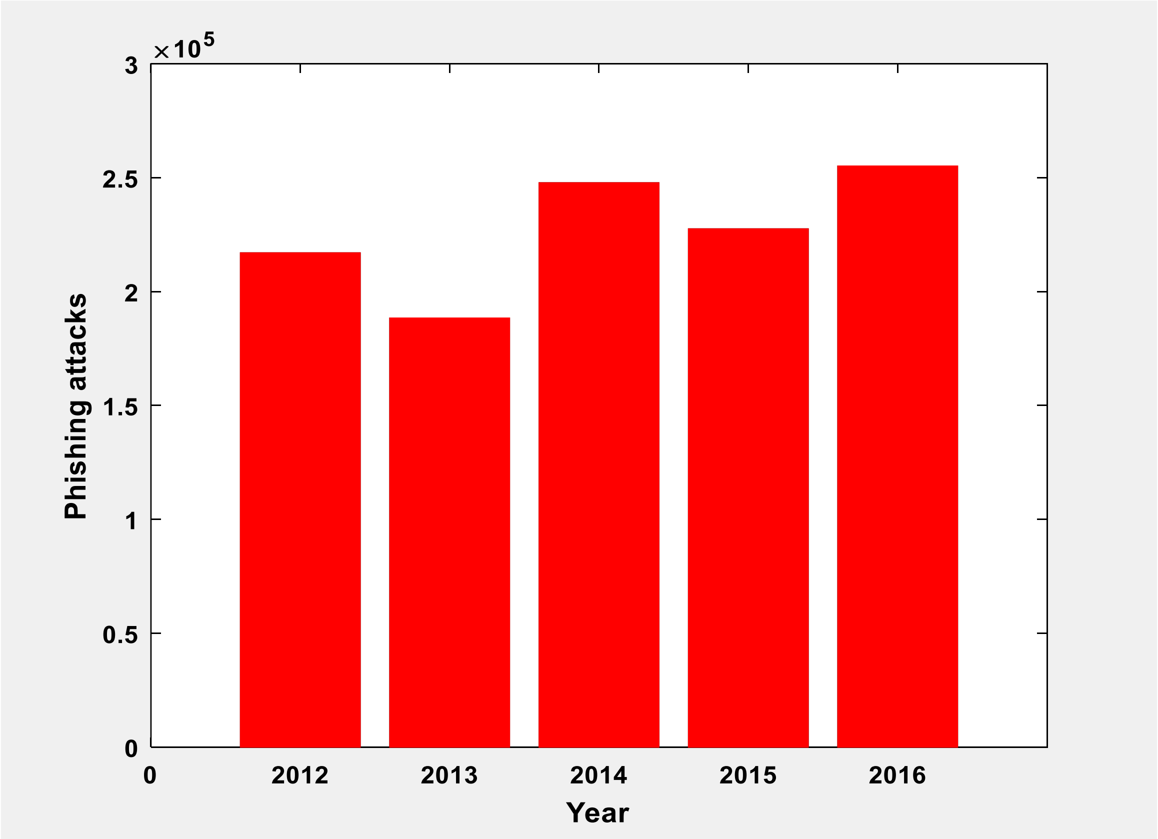


Figure 2: Phishing attacks 2012-2016

1.1.5 Distribution of Botnet and DDoS agent: Illegitimate person uses phishing hustles to install special bot and DDoS instrumentalities on unsuspecting computers and append them to their distribution networks.

II. THE PROPOSED METHOD

The following machine learning algorithms have been used for phishing detection problem in recent researches which have certain limitations:

2.1 Support Vector Machine: Support Vector Machine is a supervised machine learning algorithm that can be used for classification problems involving two classes. SVM with the maximizing margin (i.e. the distance between the closest data point and the hyperplane) yields an improved outcome [11].

2.2 Neural Network: An Artificial Neural Network or Neural Network is structure and/or function as a set of interconnected identical units (neurons). With the help of interconnections are used to send signals from one neuron to the other neuron. Other than this, the weight of the interconnection is carried to increase the distribution between the neuron [12].

2.3 Naïve Bayes: Bayesian methods are those that exceptionally apply Bayes' theorem for problems such as classification and regression. The Bayesian classifier is designed in such a way that we use it only when its features are independent in each class, but this is not the case. Even if this is not a valid case, the Bayesian classifier works very well also [13].

2.4 Random Forest: Random Forest is a machine learning classifier. This classifier integrates with a series of tree predictors, each tree votes one unit for the most popular class, then integrating these results yields the final type of result. We use random forest classifiers when it comes to classification accuracy, overfitting, and tolerant utilities [14].

2.5 k Nearest Neighbor: k Nearest Neighbor is an instance-based learning model which is based on a decision problem with instances and the k-Nearest Neighbor classification algorithm is a non-parametric classification algorithm.

We propose a new method (Figure 3) based on a decision tree that offers better classification accuracy as com-pared to the above methods.

2.6 Decision Tree: Decision Tree is a supervised learning algorithm (pre-defined outcome) that is frequently used in classification problems. Whatever is output by the Decision Tree will be visible in binary tree format. The Decision Tree keeps such rules with which it can easily predict the target variable. Where the database has a lot of similarities in large-scale training replacements, the J48 algorithm works well and measures well. J48 is an extension of ID3. The additional features of the J48 algorithm usage greedy technique to the induced tree for classification. A Decision Tee is built by analysing the training data and classify missing data [15].

2.6.1 Algorithm basics steps [16]

(a) If in each case instances exist to the same features class, the tree produces a leaf so the leaf is returned with labelling with this same class.

(b) Compute the potential information which is calculated for every attribute given for a test for the mode. After this compute the gain in information is that would get a result from analysis on the attribute.

(c) Find the favourite attribute is based on the present procedure for attribute branching selection.

2.6.2 Counting gain

This process uses the "Entropy" which is a measure of the URLs phishing. The Entropy of $\vec{y}$ is calculated

$$\boldsymbol{Entropy}\ (\vec{y}\,) = -\textstyle\sum_{j=1}^{n} \frac{|y_i|}{|\vec{y}|} \log \left(\frac{|y_i|}{|\vec{y}|}\right),$$

$$\boldsymbol{Entropy}\ (j\,|\vec{y}) = \frac{|y_i|}{|\vec{y}|} \log \left(\frac{|y_j|}{|\vec{y}|}\right)$$

$$\boldsymbol{Gain}\ (\vec{y}, j) = \boldsymbol{Entropy}(\vec{y} - \boldsymbol{Entropy}\ (j\,|\vec{y}))$$

Dividing by entire entropy due to split argument $\vec{y}$ by value j for maximizing the gain.

2.6.3. Pruning

Pruning is used to reduce the classification errors that are caused by being specializing in training sets. This procedure has to done to make the Decision Tree classifier is more general.

III. FEATURE SELECTION AND MODEL IMPLEMENTATION

In this paper, we are implementing a machine learning-based approach for detecting phishing URLs. When compared to the other approaches, machine learning-based approaches can give better results. In machine learning-based approaches, first we have to choose the features and then we check whether that features are present in the given data or not for classification or clustering. Next, the very important thing is the dataset and without this, we can't work with machine learning. Training and testing are the other steps, where we train our classifier with the data and then it classifies the upcoming data automatically based on the training. Better training can give a better outcome. By referring more than 20 research articles and the dataset (i.e. 418 phishing URLs from phishtank and 110 legitimate URLs from alexa) from phishtank and alexa, we found more than 70 features. From these features, we apply the Apriori algorithm to find the most frequent and very important features (i.e. 10 features). Later the 10 features are used for detecting the phishing, legitimate, and suspicious URLs by training the model these features. Decision Tree classifier is used to calculate the accuracy in classifying the phishing URLs from legitimate URLs. The implementation work can be further explained in detail in the following sections.

3.1 Features

3.1.1 Structured based feature

3.1.1.1 Structure feature

3.1.1.1.1 Total number of the body part

3.1.1.1.2 Total number of alternative part

3.1.1.2 Link feature

3.1.1.2.1 Total number of link

3.1.1.2.2 Number of IP based link

3.1.1.2.3 Number of deceptive link

3.1.1.2.4 Number of the link behind an image

3.1.1.2.5 Maximum number of dots in a link

3.1.1.2.6 Boolean indicator of whether there is a link that conations click/here/login/update

3.1.1.3 Element feature

3.1.1.3.1 Boolean indicator of whether it is in HTML format
3.1.1.3.2 A Boolean indicator of whether it contains JAVA script
3.1.1.3.3 A Boolean indicator of whether it contains <FORM> tag

3.1.1.4 Word-list feature

3.1.1.4.1 Boolean indicators of whether the nodes/stems listed below appear in the email body such as account, update, confirm, verify, secure, not if, log, click, internet convenience

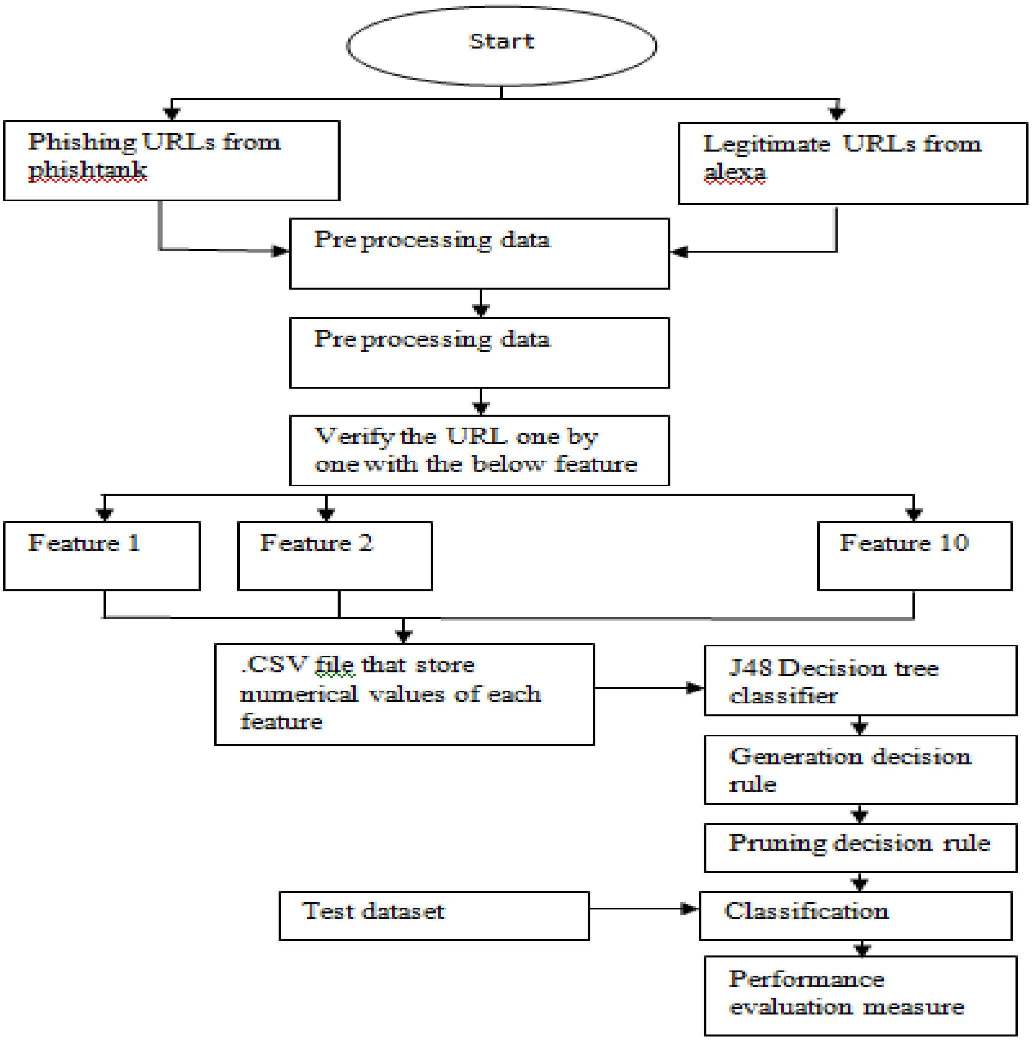

Start
Phishing URLs from phishtank
Legitimate URLs from alexa
Pre processing data
Pre processing data
Verify the URL one by one with the below feature
Feature 1
Feature 2
Feature 10
.CSV file that store numerical values of each feature
J48 Decision tree classifier
Generation decision rule
Pruning decision rule
Test dataset
Classification
Performance evaluation measure

Figure 3: The proposed model

3.1.2 Web-based feature

3.1.2.1 IP address
3.1.2.2 Long URL
3.1.2.3 URL have used as some special feature such as “-”, “\\”, “@”, “`”, “#”, “%”, “=”, “&”, “.”
3.1.2.4 Prefix and suffix in URL
3.1.2.5 Miss use of HTTP protocol
3.1.2.6 Request URL
3.1.2.7 URL of anchor
3.1.2.8 Server form handler
3.1.2.9 Abnormal URL
3.1.2.10 Redirect page
3.1.2.11 Using pop up
3.1.2.12 Hiding the suspicious link
3.1.2.13 DNS record
3.1.2.14 Web site traffic
3.1.2.15 Age of the domain
3.1.2.16 Disable right link
3.1.2.17 Port number
3.1.2.18 On mouse over to hide the link
3.1.2.19 Popup window
3.1.2.20 Irregularity form
3.1.2.21 Absent of title
3.1.2.22 SSL certificate
3.1.2.23 Web address length
3.1.2.24 Black-list keyword
3.1.2.25 Nil anchor
3.1.2.26 Foreign anchor
3.1.2.27 Foreign anchor in identity set
3.1.2.28 Foreign request
3.1.2.29 Foreign request URL in identity set
3.1.2.30 Cookies

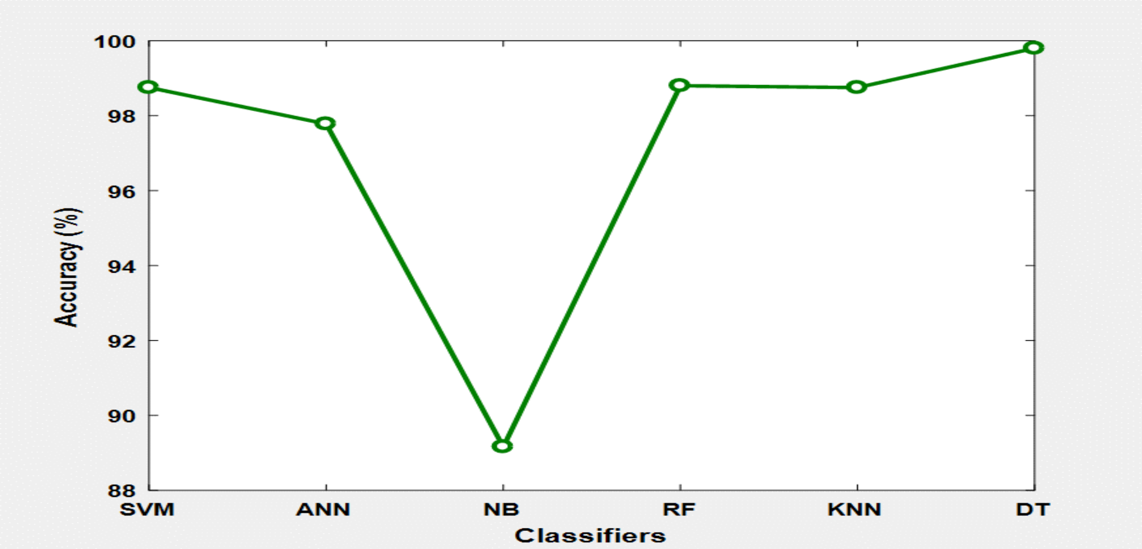
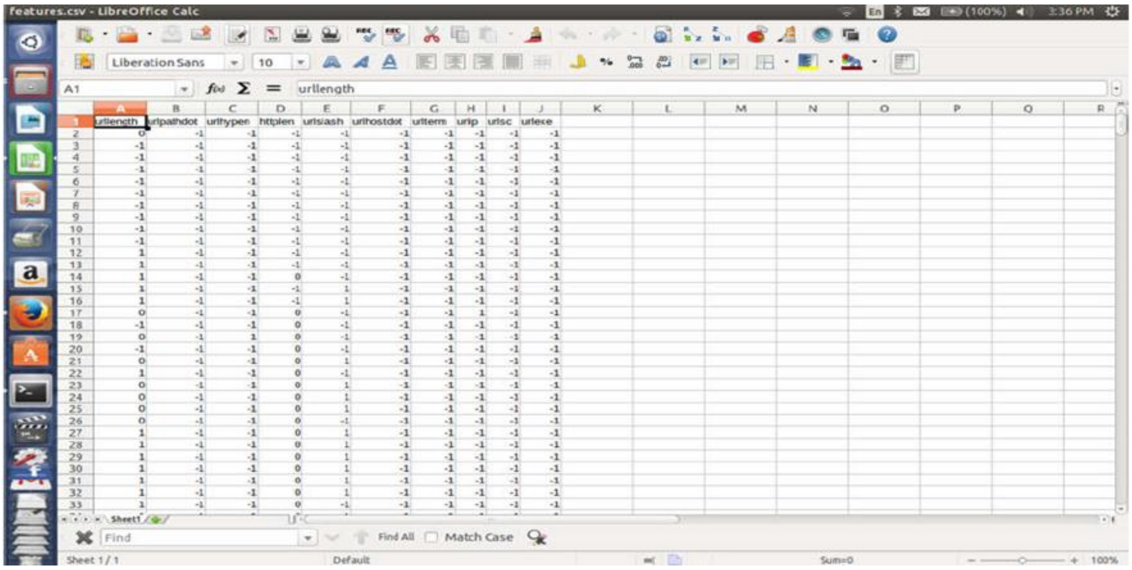
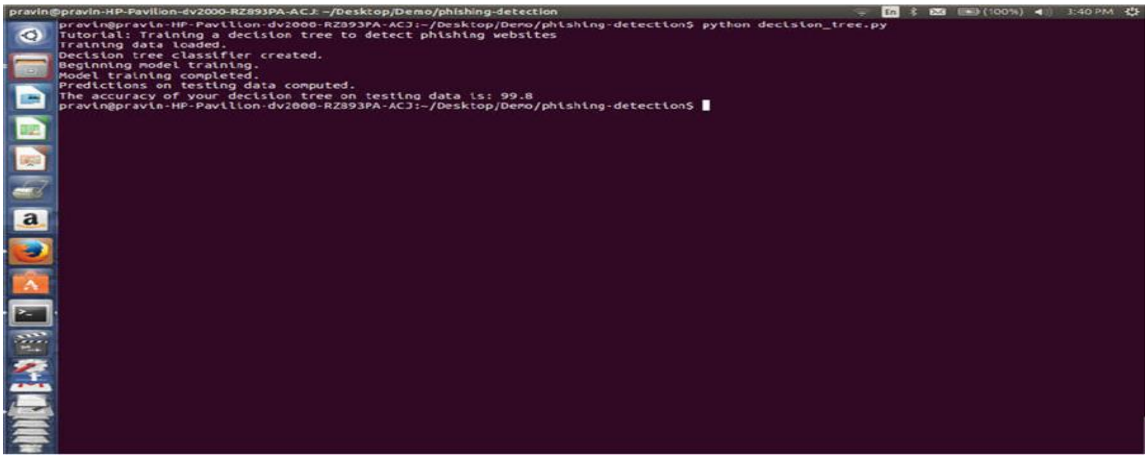


Figure 4: The accuracy rate based on different classifiers



	A	B	C	D	E	F	G	H	I	J
1	urllength	urlpathdot	urlhypen	httplen	urlslash	urlhostdot	urlterm	urlip	urlsc	urlexe
2	0	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
7	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
10	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
12	1	-1	-1	-1	-1	-1	-1	-1	-1	-1
13	1	-1	-1	-1	-1	-1	-1	-1	-1	-1
14	1	-1	-1	0	-1	-1	-1	-1	-1	-1
15	1	-1	-1	-1	1	-1	-1	-1	-1	-1
16	1	-1	-1	-1	1	-1	-1	-1	-1	-1
17	0	-1	-1	0	-1	-1	-1	1	-1	-1
18	-1	-1	-1	0	-1	-1	-1	-1	-1	-1
19	0	-1	1	0	-1	-1	-1	-1	-1	-1
20	-1	-1	-1	0	-1	-1	-1	-1	-1	-1
21	0	-1	-1	0	1	-1	-1	-1	-1	-1
22	1	-1	-1	0	-1	-1	-1	-1	-1	-1
23	0	-1	-1	0	1	-1	-1	-1	-1	-1
24	0	-1	-1	0	1	-1	-1	-1	-1	-1
25	0	-1	-1	0	1	-1	-1	-1	-1	-1
26	0	-1	-1	0	-1	-1	-1	-1	-1	-1
27	1	-1	-1	0	1	-1	-1	-1	-1	-1
28	1	-1	-1	0	1	-1	-1	-1	-1	-1
29	1	-1	-1	0	1	-1	-1	-1	-1	-1
30	1	-1	-1	0	1	-1	-1	-1	-1	-1
31	1	-1	-1	0	1	-1	-1	-1	-1	-1
32	1	-1	-1	0	1	-1	-1	-1	-1	-1
33	1	-1	-1	0	-1	-1	-1	-1	-1	-1

Figure 5: Screenshot of CSV file with 528 records based on 10 phishing features


pravin@pravin-HP-Pavilion-dv2000-RZ893PA-ACJ: ~/Desktop/Demo/phishing-detection

```
pravin@pravin-HP-Pavilion-dv2000-RZ893PA-ACJ:~/Desktop/Demo/phishing-detection$ python decision_tree.py
Tutorial: Training a decision tree to detect phishing websites
Training data loaded.
Decision tree classifier created.
Beginning model training.
Model training completed.
Predictions on testing data computed.
The accuracy of your decision tree on testing data is: 99.8
pravin@pravin-HP-Pavilion-dv2000-RZ893PA-ACJ:~/Desktop/Demo/phishing-detection$
```

Figure 6: Screenshot of calculating accuracy with decision tree classifier

3.1.2.31 Search engine
3.1.2.32 Keyword/description
3.1.2.33 Abnormal DNS record
3.1.2.34 Abnormal anchor
3.1.2.35 Abnormal server from the handler
3.1.2.36 Abnormal request URL
3.1.2.37 Abnormal cookies
3.1.2.38 Abnormal certificate in SSL
3.1.2.39 Chain identity (double URL)
3.1.2.40 Numeric and alphabet

Retrieval Number: B2321129219 /2019©BEIESP
DOI: 10.35940/ijeat.B2321.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

3.1.3 Email based feature
3.1.3.1 Email filtering
3.1.3.2 Grammar/spelling error
3.1.3.3 Age of linked to the domain name
3.1.3.4 Non-matching URL
3.1.3.5 “Here” links to the non-model domain
3.1.3.6 HTML email
3.1.3.7 Number of domains
3.1.3.8 Number of dots
3.1.3.9 Contain javascript
3.1.3.10 Spam filter
3.1.3.11 Number of links

3.1.4 Third-party feature
3.1.4.1 Traffic rank
3.1.4.2 Domain record
3.1.4.3 Domain age

For detecting the phishing URLs, the maximum number of features has been used earlier. We analyzed those features and found the minimum number of features for detecting the phishing URLs. 10 features are finalized from more than 70 features by using apriori algorithm. The proposed system is more flexible and can be added for additional features for better decision making when there is any new way of phishing attacks occur. All those features are URL length, number of dots in the path of the URL, number of hyphens in URL host, presence of SSL, number of slash present in host part of the URL, number of terms in the host part of URL, presence of IP based URL, presence of any special character in URL, presence of .exe in the URL, number of dots present in the host part of the URL. Most of the features are already implemented and we tried in our way to combine all these 10 features to detect the phishing URLs more accurately.

3.2 Classification Model

Once the features are selected, we train our model with those features to detect the phishing URLs more accurately. In the training process, we load the list of URLs to the model and for each URL, all the 10 features are extracted and the outcome (i.e. 1: phishing, -1: legitimate, 0: suspicious) and stored those results in a CSV file (Figure 5). For all the URLs, these 10 features are extracted and the outcome is stored in one CSV file. Later this CSV file is used to calculate the accuracy of the model by using any classifier. In this work, we used Support Vector Machine, Neural Network, Naïve Bayes, Random Forest, k Nearest Neighbor, and Decision Tree.

3.3 Dataset

To implement the machine learning-based phishing detection approach, we choose the dataset (which is very important) from two different sources. Phishing URLs from phishtank.com and legitimate/binge URLs from alexa.com. We took 418 phishing URLs from phishtank.com and 110 legitimate URLs from alexa.com.

IV. RESULT ANALYSIS

The machine learning classifier has generated better results for detecting phishing URLs and the output screen-shot of our work has been included with the explanation.

Table-1: Performance of different classifiers on phishing

Machine Learning Classifier	Accuracy
Support Vector Machine (SVM)	98.80
Artificial Neural Network (ANN)	97.80
Naïve Bayes (NB)	89.10
Random Forest (RF)	99.00
k-Nearest Neighbor (KNN)	98.80
Proposed Decision Tree based method (DT)	99.80

Amongst all machine learning classifiers, the Decision Tree method (DT) (table-1 and figure 4, 6) classifier yields maximum 99.80 % accuracy in classifying the phishing URL from the legitimate ones. The minimum accuracy is exhibited by Naïve Bayes (NB). The performance of DT is due to better pruning. We applied 528 phishing and legitimate URLs for training and testing. The result may vary if we increase the size of the data for training and testing.

V. CONCLUSION

In this paper, we used the Support Vector Machine, Neural Network, Naïve Bayes, Random Forest, k Nearest Neighbor, and proposed Decision Tree method. The figures show the comparative performance of different algorithms in the phishing detection problem. It is observed that the performance of the proposed Decision Tree method excels with the other techniques. It is also observed that the classification accuracy is significantly affected by the number of URLs / mails i.e. the size of available phishing data. It is also observed that the recognize 10 features govern the phishing detection problem most effectively. The Decision Tree utilizes these features in the best possible way to identify phishing.

In the future, supervised and unsupervised classifiers will be used to detect cell phone phishing, sound phishing, spear phishing and phishing in many important areas.

REFERENCE

1. Lininger, R., & Vines, R. D. (2005). Phishing: Cutting the identity theft line. John Wiley & Sons.
2. H. Tout and W. Hafner, “Phishpin: An identity-based anti-phishing approach,” Proc.- 12th IEEE Int. Conf. Comput. Sci. Eng. CSE 2009, vol. 3, pp. 347–352, 2009.
3. https://docs.apwg.org/reports/APWG_Global_Phishing_Report_2015-2016.pdf
4. https://docs.apwg.org/reports/apwg_trends_report_q1_2019.pdf
5. Bhat, Sajid Yousuf, Muhammad Abulaish, and Abdulrahman A. Mirza. "Spammer classification using ensemble methods over structural social network features." Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02. IEEE Computer Society, 2014.
6. Zouina, Mouad, and Benaceur Outtaj. "A novel lightweight URL phishing detection system using SVM and similarity index." Human-centric Computing and Information Sciences 7.1 (2017): 17
7. Bahnsen, Alejandro Correa, et al. "Classifying phishing URLs using recurrent neural networks." 2017 APWG Symposium on Electronic Crime Research (eCrime). IEEE, 2017.
8. Suh, Jong Hwan. "Comparing writing style feature-based classification methods for estimating user reputations in social media." SpringerPlus 5.1 (2016): 261.
9. Sonowal, Gunikhan, and K. S. Kuppusamy. "Masphid: a model to assist screen reader users for detecting phishing sites using aural and visual similarity measures." Proceedings of the International Conference on Informatics and Analytics. ACM, 2016.
10. Tayal, Kshitij, and Vadlamani Ravi. "Particle swarm optimization trained class association rule mining: Application to phishing detection." Proceedings of the International Conference on Informatics and Analytics. ACM, 2016.
11. Adewumi, Oluyinka Aderemi, and Ayobami Andronicus Akinyelu. "A hybrid firefly and support vector machine classifier for phishing email detection." Kybernetes 45.6 (2016): 977-994.

12. Abu-Nimeh, Saeed, et al. "A comparison of machine learning techniques for phishing detection." Proceedings of the anti-phishing working groups 2nd annual eCrime researchers' summit. ACM, 2007.
13. Lakshmi, V. Santhana, and M. S. Vijaya. "Efficient prediction of phishing websites using supervised learning algorithms." Procedia Engineering 30 (2012): 798-805.
14. RANDOM FORESTS, Leo Breiman. "Statistics Department." University of California, Berkeley, CA 94720 (2001).
15. Lakshmi, V. Santhana, and M. S. Vijaya. "Efficient prediction of phishing websites using supervised learning algorithms." Procedia Engineering 30 (2012): 798-805.
16. Korting, Thales Sehn. "C4. 5 algorithm and multivariate decision trees." Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP, Brazil (2006).

AUTHORS PROFILE

Sandip Kumar Singh, is working as an Associate Professor in the Department of Mechanical Engineering at V B S Purvanchal University Jaunpur (U.P.), India. He has done B.Tech. From KNIT Sultanpur and M.Tech. From National Institute of Technology (N I T) Kurukshetra. He has done Ph.D. from IIT (BHU) Varanasi. His area of interest is Machine Learning and Structural Health Monitoring.

Pravin Kumar Pandey, is working as an Assistant Professor in Dept. of Computer Science and Engineering, UNSIET, VBS Purvanchal University Jaunpur (UP). He has done B. Tech in CSE Department from AAIDU Allahabad, PGDSRM (PG Diploma Statistics and Research Methods), from Pondicherry University and M.Tech. in Computer Science and Engineering form Pondicherry University.