Performance Analysis of Machine Learning Algorithms Used for Web Based Phishing Detection.

[*1] **Shailendra Baliram Torane,**

[2] **Dr. Narendra Shekokar.**

[1] D. J. Sanghvi College of Engineering, Mumbai.

[2] Department of Computer Engineering, D. J. Sanghvi College of Engineering, Mumbai.

Email: [1] shailendratorane93@gmail.com, [2] narendra.shekokar@djsce.ac.in

D.O.I - 10.51201/JUSST/21/05-187

http://doi.org/10.51201/JUSST/21/05187

Abstract: Phishing is a cybercrime technique in which the attacker creates a copy of genuine websites with the same color pattern, layout, font, and logo and with a domain name that matches with the real one. Then, broadcast this fake website through various online modes like emails and social media. The attacker creates lucrative offers or discounts to lure in people to click on the phishing link. Once the user clicks on this phishing link, they are directed to the duplicate website that the attacker had created. The user believes that it is the real website and enters his/her login details and other confidential data. This data is stored on the attacker's server thus giving him full access to the victim's data. The phishing attack is mainly targeted to collect confidential data of the victim. This data includes Username, Passwords, Bank details, security Credit card numbers etc. Machine Learning algorithms are being used widely in detecting phishing websites. This paper shows performance analysis of three Machine learning algorithms used for URL phishing detection. These algorithms are Extreme Learning Machine, Support Vector Machine and Naïve Bayes algorithm. The paper analyses these algorithms on the parameters of Accuracy, Precision, Recall, F1 score and Confusion matrix. The dataset includes 11,000 entries and 30 features from UC Irvine dataset repository. The literature survey shows how only importance is given to only one parameter i.e., Accuracy parameter when analyzing performance of the URL phishing detection algorithms. This paper concludes on how Accuracy parameter does not show full picture on the overall performance of the URL phishing detection algorithms and also how Precision and Recall parameters are very important in understanding the working of these algorithms.

Keywords: Phishing, Machine learning, cybersecurity, performance analysis, Phishing detection.

1. INTRODUCTION

With the advent of innovations in Mobile networks and Internet of Things (IoT) devices, Internet usage has increased enormously. A new generation of mobile network technology like 4G and 5G has provided high-speed internet to all its users. Therefore, a large no of new users is now using the internet. This has created a favorable situation for attackers to target new users.

A phishing attack is performed in many ways. Email phishing uses emails to broadcast a phishing URL to a larger number of audiences. Spear Phishing is a technique in which a specific group of people is targeted, for example, employees of an organization. This kind of Phishing attack needs more knowledge about the target group. Whaling is similar to spear phishing but in Whaling the attacker masquerades as a senior employee of an organization and then directly target the attack on other important individuals in the organization with an aim to collect extremely sensitive information and then use it for criminal purposes. Smishing and Vishing are other Phishing attacks in which the medium of communication is an SMS text message and Voice call respectively. The SMS contains the same content as the phishing email whereas in Vishing the attacker poses as a fraud investigator and asks for sensitive information on phone calls.

Machine Learning algorithms are being used widely in detecting phishing websites. Machine learning is a field of computer science in which the algorithms are trained using a dataset and then made to perform the intended task. With training, the algorithm builds knowledge about the problem. The Algorithm then uses this knowledge and current inputs to the algorithm to make a particular decision. In this paper, three Machine Learning algorithms namely Extreme Learning Machine, Naive Bayes and Support Vector Machine are implemented on a Dataset of 11,000 Entries from UC Irvine website. The performance metrics calculated for this purpose are Accuracy, Precision, Recall, F1 Score, and Confusion Matrix.

2. DATASET:

The dataset used in this paper is 30 Features dataset from UC Irvine Machine Learning Repository database. It has 11,000 binary entries for all 30 features. These features include URL features and Web page features.

A large dataset would train the Machine Learning algorithm on different cases and possibilities which would ultimately increase the chances of Phishing detection algorithm.

The performance metrics of machine learning algorithms are inter-related. This relation will help in combining the metrics instead of looking at them individually. The accuracy metrics is widely used to measure the performance of the Algorithms. The value of accuracy metric would change if less significant features are selected for training the algorithm. The value of accuracy would change if dataset is changed even if we are using the same algorithm. The figure below shows the lists of 30 features from the dataset.

Input (Features)	*Output (Class)*
1.1. Address Bar based Features 1.1.1. Using the IP Address 1.1.2. Long URL to Hide the Suspicious Part 1.1.3. Using URL Shortening Services "TinyURL" 1.1.4. URL's having "@" Symbol 1.1.5. Redirecting using "//" 1.1.6. Adding Prefix or Suffix Separated by (-) to the Domain 1.1.7. Sub Domain and Multi Sub Domains 1.1.8. HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer) 1.1.9. Domain Registration Length 1.1.10. Favicon 1.1.11. Using Non-Standard Port 1.1.12. The Existence of "HTTPS" Token in the Domain Part of the URL	-1 Phishing 1 Legitimate
1.2. Abnormal Based Features 1.2.1.Request URL 1.2.2.URL of Anchor 1.2.3.Links in <Meta>, <Script> and <Link> tags 1.2.4.Server Form Handler (SFH) 1.2.5.Submitting Information to Email 1.2.6.Abnormal URL	
1.3. HTML and JavaScript based Features 1.3.1. Website Forwarding 1.3.2. Status Bar Customization 1.3.3. Disabling Right Click 1.3.4. Using Pop-up Window 1.3.5. IFrame Redirection	
1.4. Domain based Features 1.4.1. Age of Domain 1.4.2. DNS Record 1.4.3. Website Traffic 1.4.4. PageRank 1.4.5. Google Index 1.4.6. Number of Links Pointing to Page 1.4.7. Statistical-Reports Based Feature	

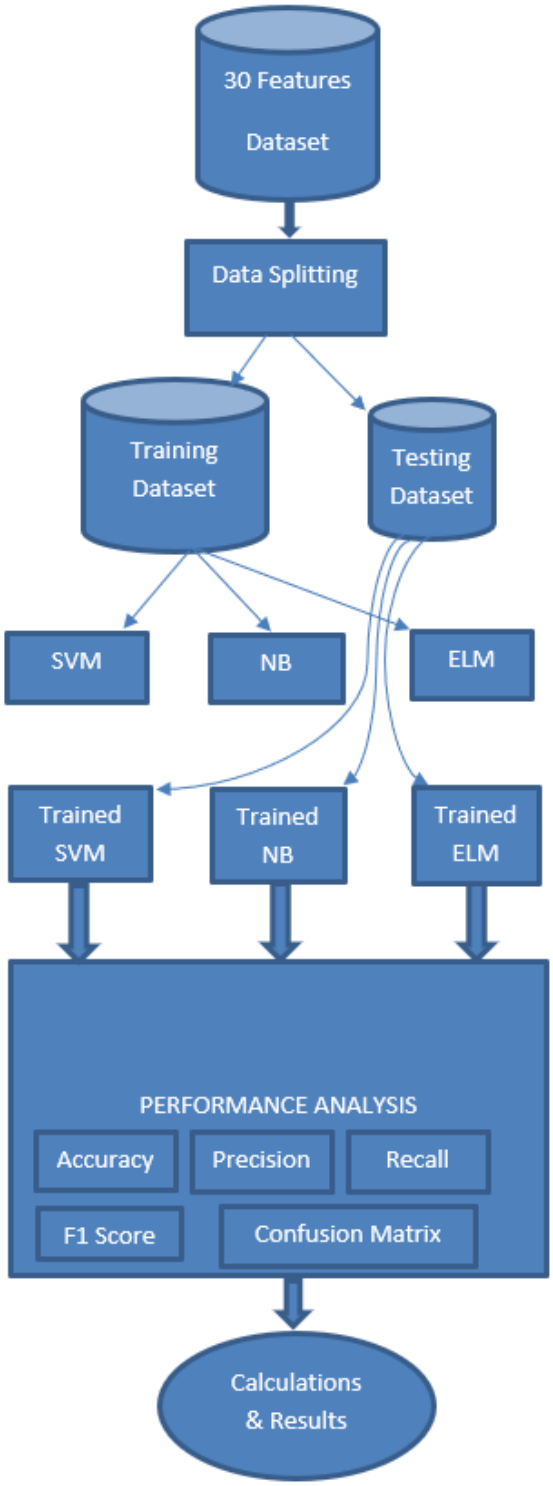
Figure.5. Table of 30 Features in the dataset[1].

3. IDENTIFICATION OF THE PROBLEM

There are lot of datasets and algorithms which are being used for URL phishing detection. Thus, every time we cannot say that we can get better results if we are using certain dataset or only half of the dataset features or if we use this algorithm, we will be getting highest accuracy. The ability of URL phishing detection changes from algorithm to algorithm and from dataset to dataset. Based on the current analysis it is clear that feature selection plays an important role if we want our algorithms to identify all types of cases in the dataset. This paper uses a 30 features dataset from UC Irvine dataset repository which has features of Phishing URLs. There is more emphasis on the Accuracy parameter when determining the overall performance of the Phishing detection algorithms. But Accuracy parameter does not show many other cases are there in the dataset which are detected by the algorithm. Detection of these cases can help us understand the working of the algorithm in a better way. Thus, Precision and Recall come into picture. This study attempts to calculate performance metrics like Accuracy, Precision, Recall, F1 Score and Confusion Matrix for ELM, NB and SVM algorithms and analyze their performance in URL phishing detection.

4. ARCHITECTURE DIAGRAM

The following architecture diagram shows three algorithms namely Support Vector Machine, Naïve Bayes and Extreme Learning Machine. It has a 30 features dataset split into two: Training set and Testing set.



Performance metrics for machine learning algorithms are calculated. These metrics include Accuracy, Precision, Recall, F1 Score and Confusion Matrix.

5. PERFORMANCE ANALYSIS

5.1 The Performance Metrics Calculated in this paper are as follow:

5.1.1 Classification Accuracy:

It is the ratio of number of correct predictions to the total number of predictions made.

$$\text{Acuracy} = \frac{\text{Number of Correct Preditions}}{\text{Total number of predictions made}}$$

Figure. 1. Formula to calculate Accuracy.

5.1.2 Precision:

It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}}$$

Figure. 2. Formula to calculate Precision.

5.1.3 Recall:

It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}}$$

Figure. 3. Formula to calculate Recall.

5.1.4 F1 Score:

F1 Score is used to measure a test's accuracy. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is.

High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model.

F1 Score tries to find the balance between precision and recall.

$$F1 = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

Figure.4. Formula to calculate F1 Score.

5.1.5 Confusion Matrix:

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.

There are 4 important terms:

- **True Positives**: The cases in which the model predicted YES and the actual output was also YES.
- **True Negatives**: The cases in which the moedl predicted NO and the actual output was NO.
- **False Positives**: The cases in which the model predicted YES and the actual output was NO.
- **False Negatives**: The cases in which the model predicted NO and the actual output was YES.

Measuring an algorithm's efficiency is important because your choice of an algorithm for a given application often has a great impact. This paper shows comparison of above performance metrics for following algorithms:

- Extreme Machine Learning.
- Support Vector Machine.
- Naïve Bayes.

The following shows calculation of performance metrics for Machine Learning Algorithms ELM, NB and SVM.

For each algorithm, Accuracy, Precision, Recall, F1 Score and Confusion Matrix is being calculated.

5.2. Accuracy Analysis:

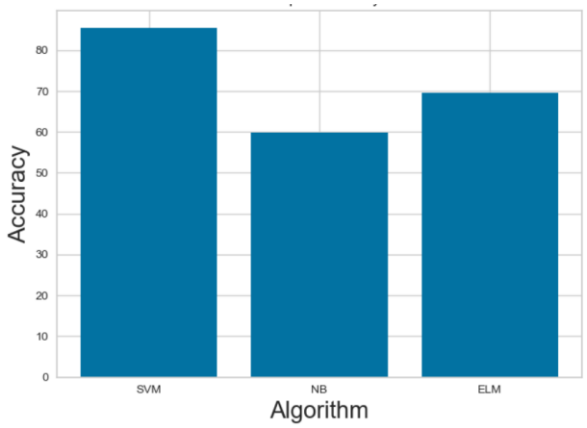


Figure.6. Accuracy Bar Graph for SVM, NB and ELM.
The above graph shows the Accuracy achieved by SVM, NB and ELM. Accuracy is the number of accurately predicted output out of total number of output predictions.

Results	SVM	NB	ELM
Accuracy	85.43%	59.90%	70.14%

In the above table we can see that SVM performs best on the Accuracy parameter followed by ELM and then NB. The Accuracy of the ELM algorithm changes with the number of neurons involved in the execution.

5.3 Confusion Matrix Analysis:

4.3.1 Confusion Matrix of Support Vector Machine Algorithm:

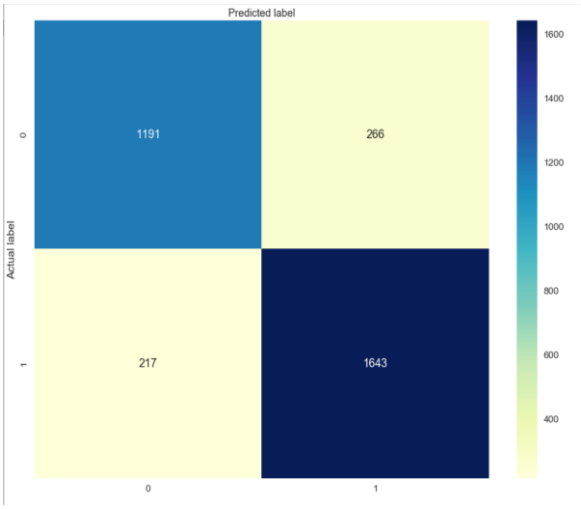


Figure.7. Confusion Matrix of SVM.

The above confusion matrix is calculated for 3317 entries.

True Negative = 1191	**False Positive** = 266
False Negative = 217	**True Positive** = 1643

With highest accuracy achieved, SVM has the high value for True Negative and True Positive in the confusion matrix. Total correct predicts are 2834 out of 3317 predictions. But we can see that the value of False Negative is 217 which means that out of 3317 entries, SVM algorithm predicted 217 Phishing URLs as Non-Phishing URLs.

5.3.2 Confusion Matrix of Naïve Bayes Algorithm:

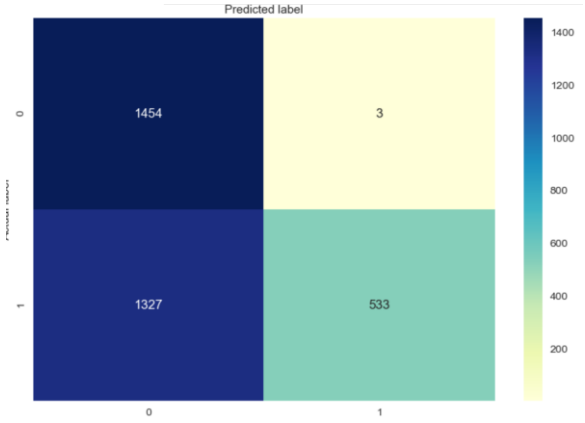


Figure.8. Confusion Matrix of NB.

True Negative = 1454	**False Positive** = 3
False Negative = 1327	**True Positive** = 533

Naive Bayes algorithm achieved highest value for False Negative thus lowering the overall accuracy of the algorithm. The False Positive value is lowest among all the three algorithms. That means the Naïve Bayes algorithm was predicting fewer wrong positives as compared to Naïve Bayes and ELM.

5.3.3 Confusion Matrix of Extreme Learning Machine Algorithm:

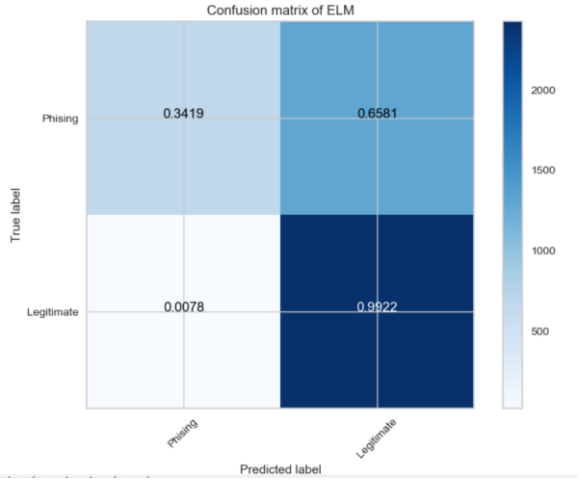


Figure.9. Confusion Matrix of ELM.

As we know, ELM is an Artificial Neural network algorithm and it has an analytical learning process. The performance of the ELM algorithm also depends on the number of neurons involved in the operation.

Out of total values that are actually phishing, 34.19% were predicted correctly as Phishing and 65.81% were predicted wrongly as Non phishing.

Out of total values that are actually Non-Phishing, 0.78% were predicted wrongly as Phishing and 99.22% were predicted correctly as Non-Phishing.

True Positives=0.3419	**False Positive**= 0.6581
False Negative= 0.0078	**True Negative**= 0.9922

5.4 Classification Report Analysis

The classification report shows us following four parameters:

1. Precision
2. Recall
3. F1-Score

To examine the overall working of a model we need to look in to both precision and recall values.

5.4.1 Classification Report of SVM:

Classifiaction Report of SVC

	precision	recall	f1-score	support
Yes	0.85	0.82	0.83	1457
No	0.86	0.88	0.87	1860
accuracy			0.85	3317
macro avg	0.85	0.85	0.85	3317
weighted avg	0.85	0.85	0.85	3317

Figure.10. Classification Report of SVM.

The precision value for Yes is 0.85, that means out of total predictions which were made as Phishing (Yes), 85% were precisely predicted as Phishing.

The Precision value for No is 0.86, that means out of total predictions which were made as Non-Phishing (No), 86% were precisely predicted as Non-Phishing.

The recall value for Yes is 0.82, that means out of total no of entries that were actually Phishing (Yes), 82% were correctly recalled as phishing.

The recall value for No is 0.88, that means out of total no of entries that were actually Non-Phishing, 88% were correctly recalled as Non-Phishing.

F1 score calculates the Harmonic mean between Precision and Recall values. The high F1 score indicates that the model has high Precision and Recall values.

The F1 score for Yes is 0.83 which is the Harmonic mean of Yes for Precision and Yes value for Recall i.e., 0.85 and 0.82 respectively.

The F1 sore for No is 0.87 which is the Harmonic Mean of No for Precision and No for Recall i.e., 0.86 and 0.88 respectively.

5.4.2 Classification Report of NB:

Classifiaction Report of NB

	precision	recall	f1-score	support
Yes	0.52	1.00	0.69	1457
No	0.99	0.29	0.44	1860
accuracy			0.60	3317
macro avg	0.76	0.64	0.57	3317
weighted avg	0.79	0.60	0.55	3317

Figure.11. Classification Report of NB.

The accuracy value for NB is lowest among three algorithms thus it has lower values Precision and Recall.

The value of Yes for Precision is 0.52, that means only 52% of total predictions made as Phishing (Yes) were correctly predicted as Phishing (Yes).

The value of No for Precision is rounded up to 0.99, that means 99% of total predictions made as Non-Phishing (No) were precisely predicted as Non-Phishing (No).

The value of Yes for Recall is rounded up to 1.0, that means 99% of predictions which were made as Phishing (Yes) were precisely predicted out of total actual predictions as Phishing (Yes).

The value of No for Recall is 0.29, that means out of total actual Non-Phishing (No) predictions, 29% of predictions which were made as Non-Phishing were recalled correctly.

The F1 score for Yes is 0.69 which is the harmonic mean of Yes for Precision and Yes for Recall i.e., 0.52 and 1.00 respectively.

The F1 score for No is 0.44 which is the Harmonic mean of No for Precision and No for Recall i.e., 0.99 and 0.29 respectively.

5.4.3 Classification Report of ELM:

Classification Report of ELM

	precision	recall	f1-score	support
Yes	0.96	0.34	0.50	1926
No	0.66	0.99	0.79	2496
accuracy			0.70	4422
macro avg	0.81	0.66	0.64	4422
weighted avg	0.79	0.70	0.66	4422

Figure.12. Classification Report of ELM.

The Precision value for yes is 0.96, that means out of all the predictions made as Phishing (Yes), 96% were correctly predicted.

The Precision value for No is 0.66, that means out of all the predictions made as Non-Phishing (No) only 66% were correct predictions as Non-Phishing.

The Recall value for Yes is 0.34, that means out of total no of entries that were actually Phishing (Yes) only 34% were correctly recalled as phishing.

The Recall value for No is 0.99, that means out of total no of entries that were actually Non-Phishing (No), 99% were correctly recalled as phishing.

The F1 sore for Yes is 0.50 which is the harmonic mean of Yes for Precision and Yes for Recall i.e., 0.96 and 0.34 respectively.

The F1 score for No is 0.79 which is the harmonic mean of No for Precision and No for Recall i.e., 0.66 and 0.99 respectively.

6. CONCLUSION

In this paper we did performance analysis of three algorithms namely Support Vector Machine, Naive Bayes and Extreme Learning Machine. The algorithms are analyzed on four parameters they are Accuracy, Precision, Recall and F1 Score.

SVM achieved highest accuracy but ranked 2nd in terms of predicting Phishing URLs after ELM algorithm. ELM algorithm has the highest Precision for predicting URLs as Phishing URLs but ranks 2nd in the race of achieving Accuracy. Naïve Bayes algorithm ranks third in the race of achieving Accuracy after SVM and ELM, but has the highest recall value for predicting URLs as Phishing. Even though ELM ranks 2nd in terms of Accuracy, it ranks 3rd in terms for Recalling URLs as Phishing.

Every algorithm ranks different when analyzed on different parameters. Not necessary that algorithm with highest accuracy will also rank high in terms of other parameters like Precision and Recall.

The Accuracy parameter is not enough when we are analyzing performance of models identifying URL phishing and thus need to focus on Recall and Precision parameter as well. The Recall parameter need to be focused more as it detects False Negatives i.e., predicting Phishing URL as Non-Phishing. Thus, achieving high recall for URL phishing detection model is very important.

7. FUTURE WORK

In this paper, performance analysis is done on three algorithms using 30 features dataset. In future, out of these 30 features, best minimum features will be identified using Forward selection technique and then performance analysis will be done on it. Also, a hybrid approach for URL phishing detection will be implemented with 30 features and also with best minimum features identified earlier and its performance analysis will be done based on parameters like Accuracy, Precision, Recall and F1 Score.

8. REFERENCES

[1] Yasin Sönmez, Türker Tuncer, Hüseyin Gökal, Engin Avci, "Phishing Web Sites Features Classification Based on Extreme Learning Machine", 2018 6th International Symposium on Digital Forensic and Security (ISDFS).

[2] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language processing and Machine Learning", Proc. – 12th OEEE Int, Conf. Semant.Comput. ICSC 2018, vol. 2018-Janua, pp. 300-301, 2018.

[3] Aniruddha Joshi, Tanuja Pattanshetti "Phishing Attack Detection using Feature Selection Techniques". Proeedings on International Conference on Communication and Information processing (ICCIP) 12 Jul 2019 Last revised: 30 Sept 2019.

[4] Bhagyashree E. Sananse, Tanuja K. Sarode "Phishing URL Detection: A Machine Learning and Web Mining-based Approach". International Journal of Computer Applications (0975 - 8887) Volume 123 - No.13, August 2015.

[5] L. MacHad0 and J. Gadge, "Phishing Sites Detection Based on C4.5 Decision Tree Algorithm," in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1-5.

[6] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. O, no. Icicct, pp. 949-952.

[7] Anjum N Shaikh, Antesar M Shabut, M. A. Hossain, "A Literature Review On Phishing Crime, Prevention Review and Investigation of Gaps", 2016 10th Internation conferences on Software, knowledge, Information Management and Applications (SKIMA).

[8] Joby James, Sandhya L, Ciza Thomas, "Detection of Phishing URLs Using Machine Learning Techniques", 2013 International conference on Control Communication and Computing.

[9] Sneha Mande, Prof. D.S.Thosar, "Detection Of Phishing Web Sites Based On Extreme Machine Learning",Vol-4 Issue-6 2018, IJARIIE-ISSN(O)-2395-4396.