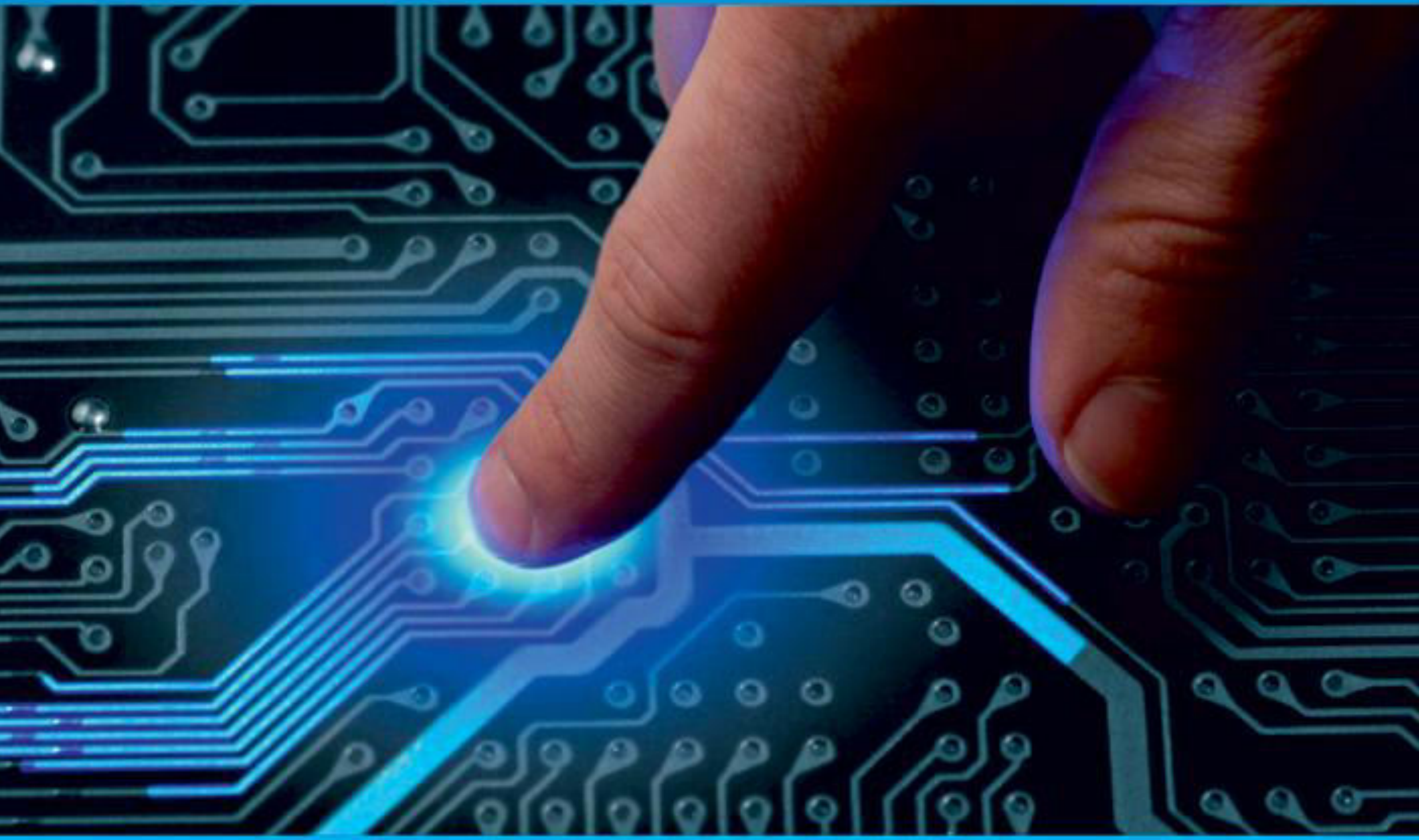
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907119 |

Detection of Suspicious URL with Naive Bayes and Support Vector Machine Learning Approach

Ashwini Nhavi[1], Shrutika Jawale[2], Pragati Patil[3], Pooja Shinkar[4], Ashish T. Bhole[5]

Under Graduate Student, Department of Computer Engineering, Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon, Maharashtra, India[1,2,3,4],

Associate Professor, Department of Computer Engineering, Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon, Maharashtra, India[5]

ABSTRACT: Suspicious URLs are widely used for performing various cyber attacks such as spamming, phishing and malware. Detection of malicious URL and threat type is critical and it finds difficult to beat these attacks.Knowing the type of threat helps in adoption of an effective countermeasure. Many existing methods are available to detect malicious URL and threat type but only for single type of attack at a time. The proposed hybrid machine learning approach is used to detect suspicious URLs. Supervised machine learning algorithms, Naive Bayes and Support Vector Machine alongwith a novel hybrid approach is used for better detection of suspicious URLs.

KEYWORDS: Cyber security, suspicious URL, Heuristic detection, Machine Learning (ML), Naive Bayes, Support Vector Machine (SVM)

I. INTRODUCTION

The world wide web become a killer application on the internet due to an immense risk of cyber attacks. In [1], attackers use web as vehicle to deliver malicious attacks such as phishing, spamming and malware infection.Phishing involves sending email to user so that user click the URL contained in the email and redirected to the malicious website submitting personal as well as financial data.Username, passwords, credit card/debit card details, national security ID and other important information.E-commerce, banks, and money transfer companies are the most targeted industries by phishing attacks.One of the major problems of 2015 is the Business Email Compromise scam. The attacker fools industries into transferring large amounts of money using spearphishing techniques. Statics determined, the attacks cause by malicious URL technique are ranked first among the ten most common attack techniques.To foil such types of attacks great efforts are directed towards detection of malicious URL's/links.

The effective solutions to prevent a phishing attack is to combine security features with the webbrowser so which it raise the alerts whenever a phishing site is accessed by an internet user.Web provides security against phishing attack through list solution such as black list and white list.Various techniques are used previously to detect suspicious URL's such as blacklist. As per [1], blacklist is effective only for known malicious URL.It involves human effort and feedback which is accurate but more time consuming.The weakness of blacklist it does not detect the unknown malicious URL. Many malicious websites are not blacklisted because they are too recent or are never or incorrectly evaluated [2].Several studies in the literature survey tackle the problem from a Machine Learning standpoint.It compile a list of URLs and classified as either malicious or benign and characterize each URL through a set of attributes.There are three main URL spreading techniques such as malicious URL, botnet URL and phishing URL which increases the number of attack upto danger level.

A number of Online Social Networks are now developing malicious content detection systems for such attacks.For detection of suspicious URL's Machine learning algorithm provides better and accurate result.Regarding the problem of malicious URL detection, two methods are in trends, based on signs or sets of rules and based on behaviour analysis techniques [3].First method detect quickly and accurately malicious URL but are not capable of detecting new malicious URL which are not in set of rules.Second method adopt machine learning and deep learning algorithms to detect malicious URL based on behaviour.A lot of software, ways and algorithms area are used for phishing detection [3].

II. RELATED WORK

Before designing the application and the system environment, the background gives the techniques used to develop the system.

A. *Database-oriented Detection*

Most of the phishing detection systems use database oriented detections such as blacklist and white list. These approaches are common in earlier detection systems since they can give faster output and more convenient detection. However, both of them have several limitations, i.e., in blacklist approach, as long as there is no prior similar information, then result could be highly false positive. Similarly, in white list approach, there can be misclassification even when a user is authentic if the user accesses to unfamiliar website because of no history of often-accessed links [4].

B. *Heuristic-oriented Detection*

Heuristic-oriented detection varies from content to visual based detection. Basically, heuristic-oriented detection results more precise performance in terms of accuracy, precision and recall. They are more robust than database oriented detection. They are visual, content and URL based techniques. However, there are also a few drawbacks in these techniques. Since visual detection checks if the two websites are visually similar, it consumes a longer execution time leading to be unrealistic. In content-based detection such as extracting keywords using tf-idf (term frequency inverse document frequency) gives wrong detection result when phishers rarely use texts in a web page. Then, it lead to high false positive values. Furthermore, although URL-based technique can detect more accurately, it highly relies on features used in the system. The more the number of features, the better accuracy, however, the longer training time and unnecessary features which even lead to reduce performance [4].

C. *Signature based Malicious URL Detection*

Most of studies often use lists of known malicious URLs.When a new URL is accessed, a database query is executed. If the URL is found blacklisted, it is considered asmalicious and a warning will be generated, otherwiseURL consider as safe. The main disadvantage is, it will be very difficult to detect new maliciousURLs that are not in the given list [5].

D. *Machine Learning based Malicious URL Detection*

There are three types of machine learning algorithms,applied on malicious URL detection methods, includingsupervised learning, unsupervised learning, and semi supervised learning. The detection methods are based on URL behaviours. A number of malicious URL systems based onmachine learning algorithms [5].

III. METHODOLOGY

The proposed system consists of hybrid algorithm which is nothing but combination of Naïve Bayes algorithm and support vector machine algorithm.First the URLs are loaded from the standardKaggle dataset [6] and then feature extraction of URL is done with following parameters.

1) URL length
2) Symbol to total character ratio
3) Number of suspicious symbols
4) Path length to URL ratio
5) Number of suspicious keywords
6) Protocols used
7) Number of dash(-)
8) Presence of symbol at last character
9) Redirection occurs
10) presence of '@'

11) Number of slash(/)
12) Presence of IP address
13) Number of question mark
14) number of subdomains
15) Presence of 'www'
16) Presence of 'htttp' word in URL
17) Presence of port number
18) Presence of Unicode characters

After the feature extraction is done, URLs are classified with the following algorithms.

A. Naïve Bayes Algorithm

The steps used to detect malicious URL with Naive Bayes algorithm are as follows.

Step 1: Separate by class.
Step 2: Summarize Dataset [trainInput, trainOutput, testInput, testOutput].
Step 3: Summarize Data by class.
Step 4: Gaussian Probability Density Function [GaussianNB()] [7].
Step 5: Class Probabilities.

B. Support Vector Machine Algorithm

The steps used to detect malicious URL with Support Vector Machine algorithm are as follows.

Step 1: Load Pandas library and the dataset using Pandas.
Step 2: Define the features and target.
Step 3: Split the dataset into train and test using sklearn before building the SVM algorithm model [trainInput, trainOutput, testInput, testOutput].
Step 4: Import the support vector classifier function or SVC function from sklearn SVM module. Build the Support Vector Machine model with the help of the SVC function [svm.LinearSVC()] [8].
Step 5: Predict values using SVM algorithm model.
Step 6: Evaluate the Support Vector Machine model.

C. Hybrid Algorithm

The steps used to detect malicious URL with Hybrid algorithm are as follows.

Step 1: Naive Bayes would be executed first.
Step 2: SVM would be executed on result of Naive Bayes algorithm.

After all the classifiers are applied, performance matrix is calculated. The parameters used for evaluating the performance [9] are as follows.

True Positive (TP): The number of phishing URLs that are classified as phishing URLs correctly by classifier.
True Negative (TN): The number of non-phishing URLs that are classified as non-phishing URLs correctly by the classifier.
False Positive (FP): The number of non-phishing URLs that are classified as phishing URLs incorrectly by the classifier.
False Negative (FN): The number of phishing URLs that are classified as non-phishing URLs correctly by the classifier.

Accuracy: The percentage of correct decisions among all testing samples [9]. The accuracy is calculated as shown in Equation (1).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} * 100 \text{..............Eq. (1)}$$

Recall: Recall is the percentage of suspicious URLs correctly labelled (TP) among all suspicious URLs of the testing data (TP+FN) [9, 10]. The Recall is calculated as shown in Equation (2).

$$Recall = \frac{TP}{TP+FN} * 100 \text{...............................Eq. (2)}$$

Precision: Precision is the percentage of suspicious URLs correctly labelled (TP) among all suspicious URLs labelled by the classifier (TP+FP) [9, 10]. The Precision is calculated as shown in Equation (3).

$$Precision = \frac{TP}{TP+FP} * 100 \text{....................Eq. (3)}$$

F-score: It is the harmonic mean of precision and recall [9, 11]. Higher the value means the classifier is good. The F-Score is calculated as shown in Equation (4).

$$F - Score = 2 * \frac{(Precision*Recall)}{(Precision+Recall)} \text{....................Eq. (4)}$$

IV. RESULTS AND DISCUSSION

The Table 1 shows the results of the performance of different classifier used to detect malicious URL. It is observed that hybrid algorithm gives more accuracy(86.88%) as compared to other algorithms. The hybrid algorithm gives higher precision (0.88), recall (0.98) and f-score (0.87) among Naive Bayes and Support Vector Machine algorithms alone. The hybrid algorithm takes average time as 0.18400 seconds while Naive Bayes proved to be efficient in terms of time execution which took lesser time of 0.17899 seconds.

The comparison between different classifiers for various parameters is as shown in Figure 1. Figure 2 compares the accuracyof different classifiers for detection of Malicious URLs.

Table 1: Classifier Performance for detection of Malicious URLs

Classifier	Naïve Bayes	SVM	Hybrid
Precision	0.78	0.82	0.88
Recall	0.84	0.92	0.98
f-score	0.78	0.80	0.87
Accuracy	78.14%	80.44%	86.88%
Execution time (sec)	0.17899	0.20099	0.18400

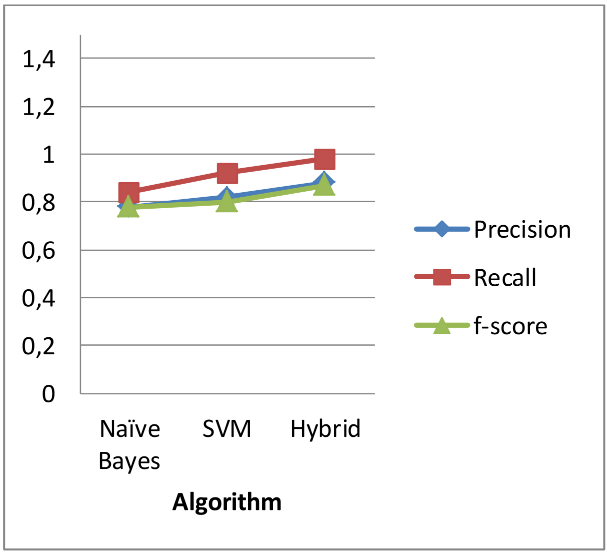


Fig.1: Comparing classifiers for various parameters

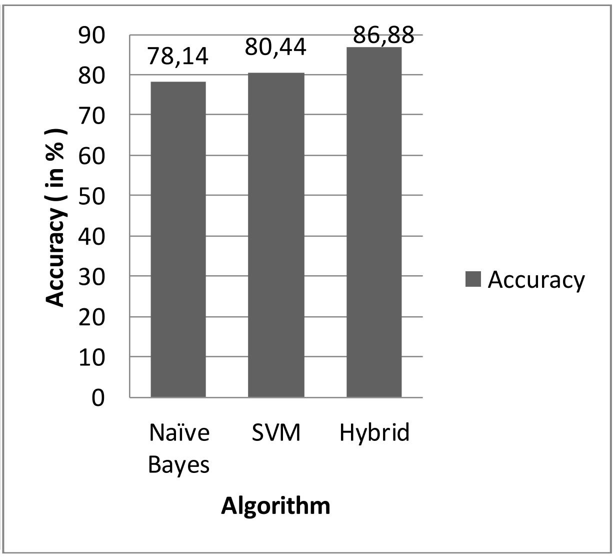


Fig.2: Comparing accuracy of different classifiers for detection of Malicious URLs

From the results, it is observed that

- Phishing URLs and domain show some characteristics that are different from other URLs and domains.
- Phishing URLs and domain names have altogether different lengths contrasted with different URLs and domain names inside the internet.
- A large number of the phishing URLs contained the name of the brand they focused on.

IV. CONCLUSION AND FUTURE WORK

Among the malicious URL detection techniques, Machine Learning techniques proved to be promising one. A systematic and comprehensive survey on malicious URL detection using machine learning technique is followed. It provided an example of how to employ parameter tuning and feature based methods to develop machine learning tool for malicious URL detection. To detect suspicious URL, the proposed hybrid machine learning approach consisting of Naive Bayes & Support Vector Machine algorithms yielded more accuracy.

In future, other machine learning algorithms can be combined to further improve accuracy for detection of malicious URLs.

REFERENCES

1. Choi, Hyunsang, Bin B. Zhu, and Heejo Lee. "Detecting Malicious Web Links and Identifying Their Attack Types." *WebApps* 11, no. 11 (2011): 218.
2. Vanhoenshoven, Frank, Gonzalo Nápoles, Rafael Falcon, KoenVanhoof, and Mario Köppen. "Detecting malicious URLs using machine learning techniques." In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1-8. IEEE, 2016.
3. Jain, Manish, Kanishk Rattan, Divya Sharma, KritiGoel, and Nidhi Gupta. "Phishing Website Detection System Using Machine Learning." *International Research Journal of Engineering and Technology (IRJET)* 7, no. 5 (2020).
4. Aung, Eint Sandi, and Hayato YAMANA. "Malicious URL Detection: A Survey."
5. Do Xuan, Cho, HoaDinh Nguyen, and Tisenko Victor Nikolaevich. "Malicious url detection based on machine learning." (2020).
6. https://www.kaggle.com
7. Sayamber, Anjali B., and Arati M. Dixit. "Malicious URL detection and identification." *International Journal of Computer Applications* 99, no. 17 (2014): 17-23.
8. Kumar, Rajesh, Xiaosong Zhang, Hussain Ahmad Tariq, and RiazUllah Khan. "Malicious URL detection using multi-layer filtering model." In *2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 97-100. IEEE, 2017.
9. Banik, Bireswar, and AbhijitSarma. "Phishing URL detection system based on URL features using SVM." *Int. J. Electron. Appl. Res.(IJEAR)* 5, no. 2 (2018): 40-55.
10. Mamun, Mohammad Saiful Islam, Mohammad Ahmad Rathore, ArashHabibiLashkari, Natalia Stakhanova, and Ali A. Ghorbani. "Detecting malicious urls using lexical analysis." In *International Conference on Network and System Security*, pp. 467-482. Springer, Cham, 2016.
11. Kazemian, Hassan B., and Shafi Ahmed. "Comparisons of machine learning techniques for detecting malicious webpages." *Expert Systems with Applications* 42, no. 3 (2015): 1166-1177.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 ✆ 6381 907 438 ✉ ijircce@gmail.com

Scan to save the contact details

www.ijircce.com