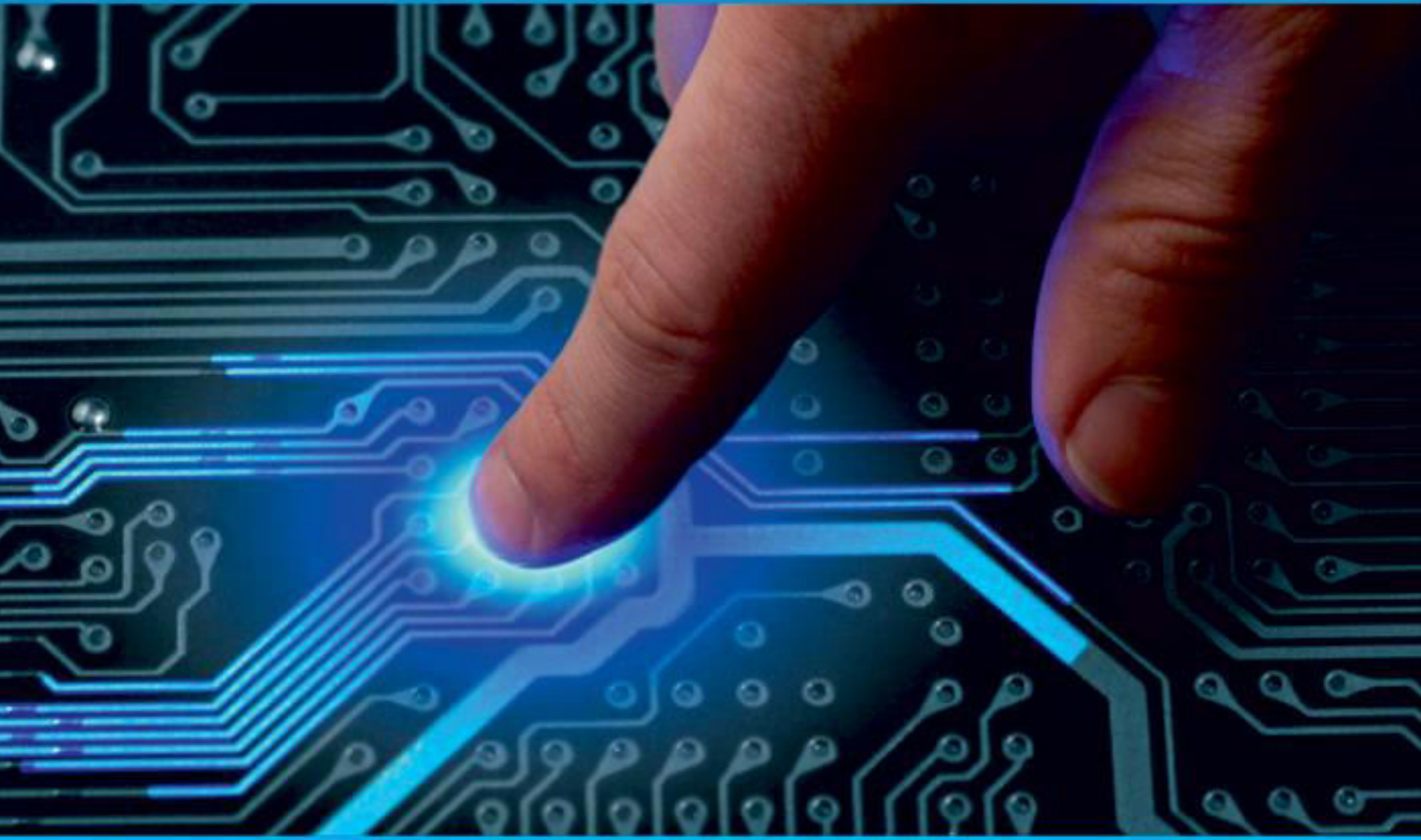
IJIRCCE

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



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005162|

Detecting Phishing Attacks Using Machine Learning

M.Subba Rao[1], G. Rajesh[2], M.Sai Sandeep[3], M.Siva Nagaraju[4], M.Gowtham[5]

Assistant Professor, Department of CSE, KKR & KSR Institute of Technology and Sciences, A.P, India. [1]

B.Tech Student, Department of CSE, KKR & KSR Institute of Technology and Sciences, A.P, India[2]

B.Tech Student, Department of CSE, KKR & KSR Institute of Technology and Sciences, A.P, India[3]

B.Tech Student, Department of CSE, KKR & KSR Institute of Technology and Sciences, A.P, India[4]

B.Tech Student, Department of CSE, KKR & KSR Institute of Technology and Sciences, A.P, India[5]

ABSTRACT: Malicious Web sites largely promote the growth of Internet criminal activities and constrain the development of Web services. As a result, there has been strong motivation to develop systemic solution to stop the user from visiting such Web sites. We propose a learning based approach to classifying Web sites into 3 classes: Benign, Spam and Malicious. Our mechanism only analyses the Uniform Resource Locator (URL) itself without accessing the content of Web sites. Thus, it eliminates the run-time latency and the possibility of exposing users to the browser based vulnerabilities. By employing learning algorithms, our scheme achieves better performance on generality and coverage compared with blacklisting service. URLs of the websites are separated into 3 classes:

- Benign: Safe websites with normal services
- Spam: Website performs the act of attempting to flood the user with advertising or sites such as fake surveys and online dating etc.
- Malware: Website created by attackers to disrupt computer operation, gather sensitive information, or gain access to private computer systems.

KEYWORDS: Phishing, attacks, machine leaning, web site.

I. INTODUCTION

While the Internet has brought unprecedented convenience to many people for managing their finances and investments; it also provides opportunities for conducting fraud on a massive scale with little cost to the fraudsters. Fraudsters can manipulate users instead of hardware/software systems, where barriers to technological compromise have increased significantly. Phishing is one of the most widely practiced Internet frauds. It focuses on the theft of sensitive personal information such as passwords and credit card details. Phishing attacks take two forms:

- Attempts to deceive victims to cause them to reveal their secrets by pretending to be trustworthy entities with a real need for such information
- Attempts to obtain secrets by planting malware onto victim's machines.

The specific malware used in phishing attacks is subject of research by the virus and malware community and is not addressed in this thesis. Phishing attacks that proceed by deceiving users are the research focus of this thesis and the term phishing attack will be used to refer to this type of attack. Malicious Web sites largely promote the growth of Internet criminal activities and constrain the development of Introduction to Machine Learning: Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Some machine learning methods:

Machine learning algorithms are often categorized as supervised or unsupervised. In addition to that there are ensemble learning and reinforcement learning.

- Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a recognized education dataset, the gaining knowledge of set of rules produces an inferred characteristic to make predictions approximately the output values. The machine is capable of offer objectives for any new input after sufficient education. The mastering algorithm also can compare its output with the proper, supposed output and find errors in order to regulate the model consequently. In evaluation, unsupervised gadget mastering algorithms are used while the facts used to train are neither categorized nor classified. Unsupervised getting to know research how structures can infer a function to describe a hidden shape from unlabeled records. The machine doesn't figure out the right output, but it explores the statistics and can draw inferences from datasets to explain hidden systems from unlabeled statistics.

● Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabelled data for training typically a small amount of labeled data and a large amount of unlabelled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data.

• Reinforcement Learning: It is set taking suitable movement to maximize reward in a particular situation. It is employed by using diverse software and machines to find the satisfactory viable behavior or direction it have to take in a selected scenario. Reinforcement learning differs from the supervised mastering in a manner that during supervised learning the schooling data has the solution key with it so the version is educated with an appropriate solution itself whereas in reinforcement getting to know, there may be no solution but the reinforcement agent comes to a decision what to do to perform the given undertaking. In the absence of education dataset, its miles bound.

II. LITERATURE SURVEY

[1] JavaScript is a browser scripting language that permits builders to create sophisticated patron-facet interfaces for internet programs. However, JavaScript code is likewise used to perform attacks against the user's browser and its extensions. These attacks typically bring about the download of extra malware that takes entire control of the sufferer's platform, and are, therefore, called "drive-by way of downloads." Unfortunately, the dynamic nature of the JavaScript language and its tight integration with the browser make it difficult to come across and block malicious JavaScript code. This paper offers a novel approach to the detection and evaluation of malicious JavaScript code.

[2] Proliferation of phishing attacks in recent years has presented an important cyber security research area. Over the years, there has been an increase in the technology, diversity, and sophistication of these attacks in response to increased user awareness and countermeasures. In this paper, we propose a novel scheme to automatically detect phishing URLs by mining and extracting Meta data on URLs from various Web services. Applying the proposed approach on real-world data sets, it is demonstrated that Logistic Regression classifier can achieve an overall accuracy of 97.2- 99.8%, false positive rate of 0.1-1% and false negative rate of 0.7-6.5% in detecting phishing and non-phishing URLs.

[3] Information and Communication Technology (ICT) has a splendid impact on social well-being, economic increase and country wide security in these dais's international. Generally, ICT includes computer systems, cell communique gadgets and networks. ICT is likewise embraced via a group of humans with malicious reason, also known as network intruders, cyber criminals, and so forth. Confronting these unfavorable cyber sports is one of the international priorities and crucial research place. Anomaly detection is an important information evaluation mission that's beneficial for figuring out the community intrusions. This paper provides an in-intensity evaluation of four predominant classes of anomaly detection strategies which encompass type, statistical, facts concept and clustering. The paper also discusses studies demanding situations with the datasets used for community intrusion detection.

[4] Phishing is an online crook act that occurs whilst a malicious website impersonates as valid webpage to be able to gather sensitive records from the person. Phishing assault keeps to pose a extreme hazard for net customers and annoying chance in the subject of digital commerce. This paper specializes in discerning the massive features that discriminate among valid and phishing URLs. These features are then subjected to associative rule mining—apriori and predictive apriori. The policies acquired are interpreted to emphasize the functions which might be greater conventional in phishing URLs. Analyzing the know-how available on phishing URL and considering self belief as a hallmark, the functions like shipping layer protection, unavailability of the top degree area inside the URL and keyword inside the course part of the URL had been located to be sensible indicators for phishing URL. In addition to this number of slashes in the URL, dot within the host portion of the URL and length of the URL also are the key elements for phishing URL.

III. EXISTINGSYSTEM

A poorly structured NN version may additionally motive the version to below fit the education dataset. On the opposite hand, exaggeration in restructuring the gadget to fit each unmarried item in the schooling dataset may additionally purpose the device to be over outfitted. One viable solution to keep away from the Over becoming problem is through restructuring the NN model in terms of tuning some parameters, including new neurons to the hidden layer or now and again including a new layer to the community. A NN with a small range of hidden neurons won't have a fine representational energy to model the complexity and diversity inherent in the data. On the alternative hand, networks with too many hidden neurons ought to over fit the facts. However, at a certain degree the model cannot be improved, consequently, the structuring manner needs to be terminated. Hence, a suitable blunders fee ought to be distinctive whilst developing any NN version, which itself is taken into consideration a hassle for the reason that it's far hard to determine the applicable blunders rate a priori. For example, the model fashion designer may additionally set the suitable mistakes price to a fee this is unreachable which causes the version to stick in nearby minima or sometimes the model dressmaker can also set the applicable error price to a fee which can further be improved.

Disadvantage:
1. It will take time to load all the dataset.
2. Process is not accuracy.
3. It will analyze slowly.

IV. PROPOSED METHODOLOGY

Lexical features are based on the observation that the URLs of many illegal sites look different, compared with legitimate sites. Analyzing lexical features enables us to capture the property for classification purposes. We first distinguish the two parts of a URL: the host name and the path, from which we extract bag-of-words (strings delimiters). We find that phishing website prefers to have longer URL, more levels (delimited by dot), more tokens in domain and path, longer token. Besides, phishing and malware websites could pretend to be a benign one by containing popular brand names as tokens other than those in second level domain. Considering phishing websites and malware websites may use IP address directly so as to cover the suspicious URL, which is very rare in benign case. Also, phishing URLs are found to contain several suggestive word tokens (confirm, account, banking, secure, ebayisapi, webscr, login, signin), we check the presence of these security sensitive words and include the binary value in our features. Intuitively, malicious sites are always less popular than benign ones. For this reason,
site popularity can be considered as an important feature. Traffic rank feature is acquired from Alexa.com. Host-based features are based on the observation that malicious sites are always registered in less reputable hosting centers or regions

Advantages:
1. All of URLs in the dataset are labeled.
2. We used two supervised learning algorithms random forest and support vector machine to train using scikit-learn library.

SYSTEM ARCHITECTURE

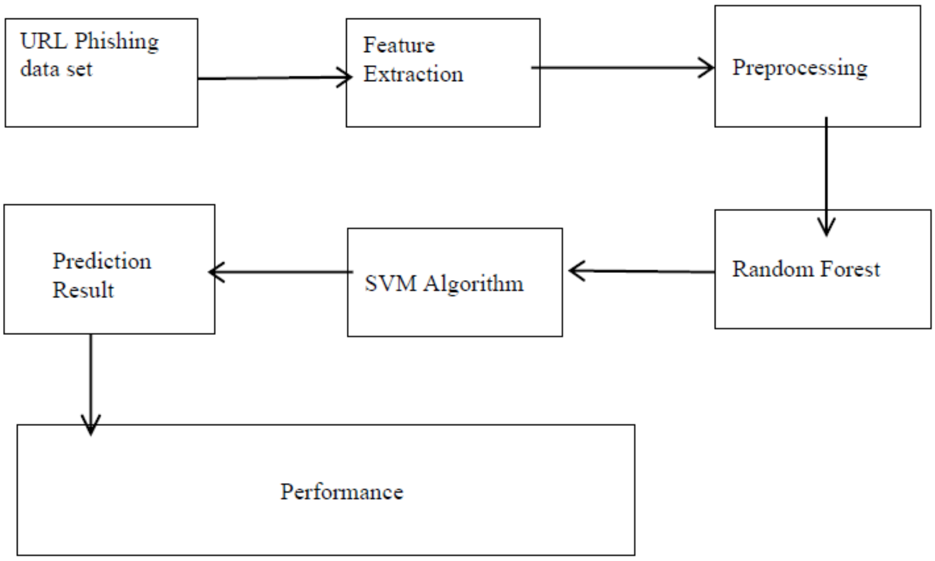


Fig: Propose System Architecture

V. CONCLUSION AND FUTURE SCOPE

CONCLUSION: By analyzing the results, we concluded that Support Vector Machine turned out to be best classifier for prediction of URL attacks using Machine learning and this model generates accurate results with high accuracy. We choose three popular classifiers considering their performance for the project. We select one dataset which is available at www.kaggle.com dataset repository. In order to compare the classification performance of three learning algorithms, classifiers are applied on same data and results are compared on the basis of misclassification and correct classification rate, it can be concluded that Support Vector Machines the best as compared to Decision Tree and Random Forest techniques.

FUTURE ENHANCEMENTS: By using different kinds of machine learning techniques to predict the phishing attacks have summarized. Determined the prediction accuracy of each algorithm and apply the proposed system for the area it needed. All these can be taken into consideration and even more reliable and more accurate algorithms can be used. We can make use of more learning techniques and Deep Learning techniques to predict heart attack chances with less time and more accuracy. Then the project will be more powerful to depend upon and even more efficient to depend upon.

REFERENCES

[1] Marco Cova, Christopher Kruegel, Giovanni Vigna, "Detection and analysis of drive by download attacks and malicious javascript code", Proceedings of the 19th International Conference on World Wide Web, pp.281-290, 2017.
[2] R. B. Basnet, A. H. Sung, "Mining web to detect phishing urls", Proceedings of the InternationalConference on Machine Learning and Applications, vol. 1, pp. 568-573, Dec 2016.
[3] Mohiuddin Ahmed, Abdun Naser Mahmood, Jiankun Hu, "A survey of network anomaly detection techniques", J. Netw. Comput. Appl., vol. 60, no. C, pp. 19-31, 2016.
[4] S. CarolinJeeva, Elijah Blessing Rajsingh, "Intelligent phishing url detection using association rule mining", Human-centric Computing and Information Sciences, vol. 6, no. 1, pp. 10, 2016.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

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