MACHINE LEARNING BASED APPROACHES FOR EFFICIENT DETECTION OF MALICIOUS URLs

[1]Dr. K. Radhika, [2]Ranjith Reddy Gaddam, [3]Saiprakash Bollam

[1,2,3]Department of Information Technology, Chaitanya Bharathi Institute of Technology (A), Hyderabad, India

ABSTRACT:

Malicious URL is a common and serious threat to cyber security. Malicious URLs host unsolicited content and lure unsuspecting users to become victims of scams and cause losses of billions of dollars every year. Detection and blocking of malicious URLs is of utmost important issue for organizations as well as individuals to protect their information assets. Though these techniques can prevent most of the malicious URLs, these solutions lead to complex User Interfaces, high computational time, cost and also may require changes on the website. In this paper, we developed 3 machine learning models namely Random Forest (RF), Artificial Neural Networks (ANN) and Support Vector Machine (SVM) for classifying and detecting Malicious URLs. Further, we analyzed the performance of these models.

Keywords: Malicious URLs, Machine Learning, Random Forest, Artificial Neural Networks, Support Vector Machine

[1] INTRODUCTION

The increased use of Internet paved way to cyber attackers to perform malicious activities such as unauthorized information access and to alter, destroy, disable or expose an information asset. These attacks include Phishing, spam, drive-by downloads. According to the statistics reported by global cloud service Kaspersky Security Network (KSN), 10.18% of Internet user computers experienced malware attacks worldwide and 173,335,902 unique URLs were recognized as malicious by Web Antivirus during the period Nov 2019 to Dec 2020.

Malicious URLs are the links embedded in e-mails or web by clicking on which, causes a Virus or Trojan horse or Ransomware to be downloaded on to the user’s system to compromise the user’s machine or an Organization’s Network. Cyber attackers use these malicious URLs to obtain sensitive information such as usernames, passwords, and credit card details. These attacks are typically carried out by email spoofing which lures users to enter personal information at a fake website which matches the look and feel of the legitimate site. Users are often enticed by e-mails purporting to be from trusted parties such as social web sites, auction sites, banks, online payment processors.

Detection and blocking of malicious URLs is of utmost important issue for organizations as well as individuals to protect their information assets [7][8]. Though there exist several mechanisms for malicious URL detection, they often result in more false negatives which means that the detection mechanism fails to detect the malicious URLs and treats them as genuine and further it does not block those URLs.

Machine learning algorithms allow computers to perform complex tasks without human intervention. Machine learning algorithms allow training the machines with sufficiently large datasets there by predicting the outputs for a new set of input data. Machine learning algorithms are classified as Supervised, Unsupervised and Semi- supervised machine learning algorithms. Malicious websites have certain characteristics and patterns and to identify those features can help us to detect attacks. Identification of such features is a classification task and can be solved using Machine Learning techniques. In this paper, we present 3 classification models to overcome malicious URL problem. To evaluate this model, we have used the dataset from UCI repository [9], which contains 30 attributes and 11055 instances.

In this paper, we developed 3 machine learning models namely Random Forest (RF), Artificial Neural Networks (ANN) and Support Vector Machine (SVM) for classifying and detecting Malicious URLs. Further, we analyzed the performance of these models.

The rest of the paper is organized as follows: Section 2 deals with the existing malicious URL prevention and detection schemes and the flaws of existing methods. Section 3 presents the machine learning approaches and describes the dataset used in this research work. Section 4 presents the results of implementation and the analysis. Conclusions and scope of future work are discussed in Section 5.

[2] STATE OF THE ART

Studies that deal with malicious URLs are classified as Malicious URL prevention and Malicious URL detection schemes as described below:

2.1 Malicious URL Prevention Schemes:

Malicious URL prevention schemes showed in Fig. 1 deal with the prevention of malicious URLs by including an additional layer of security. Though these techniques can prevent most of the malicious URLs, these solutions lead to complex User Interfaces, high computational time, cost and also may require changes on the website.

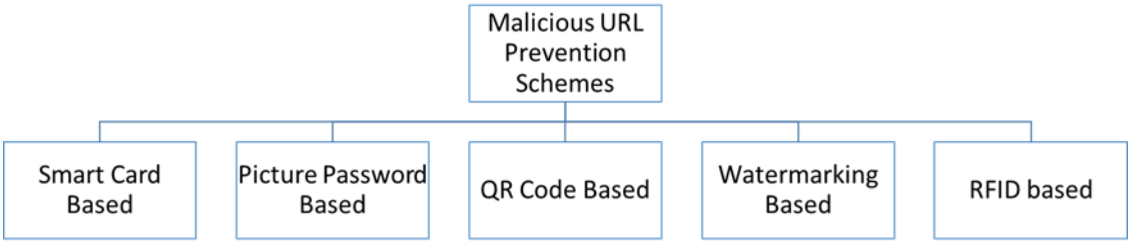

Fig. 1 Malicious URL Prevention Schemes

2.2 Malicious URL Detection Schemes

A broad classification of malicious URL detection schemes based on the underlying technique used for identifying malicious URL detection is shown in Fig. 2.

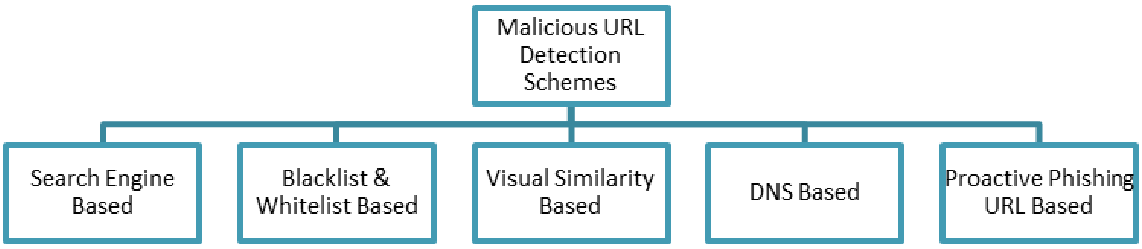
Malicious URL Detection Schemes
Search Engine Based
Blacklist & Whitelist Based
Visual Similarity Based
DNS Based
Proactive Phishing URL Based

Fig. 2 Malicious URL Detection Schemes

Search engine-based techniques extract features such as text, images, and URLs from websites, then search for them using single or multiple search engines. Malicious URLs are detected based on their lower Index as compared to the normal websites and that they remain active for a short duration of time.

Black-list and white-list based detection schemes use a whitelist of normal websites and a blacklist containing anomalous websites. The blacklist is obtained either by user feedback or via reporting by the third parties who perform phishing URL detection.

Visual similarity-based detection technique utilizes the visual similarity between phishing websites and the authentic websites. It then checks whether the URL is on the authentic domain URL list. If not, the website is identified as a phishing website.

In DNS based detection, DNS is used to validate the IP address of a malicious website. In this approach, DNS will identify whether the IP address over which the malicious website is running is in the list of authentic website IPs. If it is not, the website is marked as malicious.

Proactive phishing URL detection-based approach detects probable phishing URLs by generating different combinatorial URLs from existing authentic URLs and determining whether they exist and are involved in phishing-related activities on the web.

Different techniques for detecting and blocking malicious URLs are proposed in the Literature. In [1] authors proposed a multidimensional feature phishing detection approach by using deep learning. A Hybrid Ensemble Feature Selection Framework for Machine Learning- Based Phishing Detection System was proposed in [2]. In this research work, authors derived baseline features, when coupled with Random Forest classifier, are highly effective in distinguishing between phishing and legitimate websites. In [3], authors used Data pre-processing models and Word Decomposer models. Authors of [4] [6] used ML algorithms-KNN and SVM with a true positive rate of 98% and false positive and false negative rates of 2%. But it suffers from lazy learning. Authors of [5] proposed a Framework for Auto-Detection of Phishing Websites, but it requires more time is required to build hybrid model. Evaluating webpage similarity may not be accurate.

Malicious websites have certain characteristics and patterns and to identify those features can help us to detect attacks. To identify such features is a classification task and can be solved using Machine Learning techniques.

[3] PROPOSED MACHINE LEARNING APPROACHES

Random Forest: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision tree's habit of over fitting to their training set. Random forest, like its name implies, consists of many individual decision trees that operate as an ensemble.

Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model"s prediction. Many relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

Artificial Neural Networks: A neural network is structured as a set of interconnected identical units (neurons). The interconnections are used to send signals from one neuron to the other. In addition, the interconnections have weights to enhance the delivery among neurons. The neurons are not powerful by themselves, how-ever, when connected to others they can perform complex computations. Weights on the interconnections are updated when the network is trained, hence significant interconnection play more role during the testing phase. Since interconnections do not loop back or skip other neurons, the network is called feed-forward. The power of neural networks comes from the nonlinearity of the hidden neurons. In consequence, it is significant to introduce nonlinearity in the network to be able to learn complex mappings. The commonly used function in neural network research is the sigmoid function. Although competitive in learning ability, the fitting of neural network models requires some experience, since multiple local minima are standard and delicate regularization is required.

Support Vector Machine: The basic idea of SVM classifier is to draw hyper- plane(s) to separate the classes. Suppose, we have „n" number of features in the dataset. Then, each data has to plot a point in n-dimensional space where each feature value is the particular coordinate. The algorithm starts by identifying a subset of training dataset known as support vectors. The main aim is to separate the support vectors of two different classes in an efficient way. In two dimensional spaces, SVM focuses on to draw line in order to achieve the maximum distance from the support vectors of each class and minimize the wrong occurrences in each side. But drawing linear hyper-plane is not suitable in case of n-dimensional space where n is relatively higher. SVM functions are taken from kernel function f(x, y) where its value is chosen based on the situation. In our case, SVM is used as a binary classifier to separate the data points of malicious and non-malicious classes.

[4] RESULTS & ANALYSIS

To evaluate our machine learning techniques, we have used the Phishing Websites Dataset from UCI Machine learning repository. It consists of 11,055 URLs (instances) with 7738 phishing instances and 3317 legitimate instances. Each instance contains 30 features. Each feature is associated with a rule. If the rule satisfies, it is termed as phishing. If the rule doesn't satisfy then it is termed as legitimate. The features take three discrete values. „1" if the rule is satisfied, „0" if the rule is partially satisfied, „-1" if the rule is not satisfied. Each instance consists of 30 features comprising of various attributes typically associated with phishing or suspicious web pages such as presence of IP address in the URL domain or presence of JavaScript code to modify the web browser address bar information. Each feature is associated with a rule. If the rule is satisfied, we take it as an indicator of phishing and legitimate otherwise. The dataset has been normalized to contain only discrete values. Each feature of each instance will contain '1', if the rule associated with that feature is satisfied, '0' if partially satisfied and '-1‟ if unsatisfied.

We tested the three machine learning algorithms on the "Phishing Websites Dataset‟ from the UCI Machine Learning Repository and reviewed their results. This section presents the results of the performance of all the classifier algorithms described in Section 3. Table 1 shows the confusion matrix for Random forests. With 1293 true positives, 182 false positives, 162 false negatives and 1680 true negatives. Table 1 shows the confusion matrix for Artificial Neural Networks. With 1246 true positives, 155 false positives, 209 false negatives and 1707 true negatives. Table III shows the confusion matrix for Support Vector Machine. With 1254 true positives, 129 false positives, 209 false negatives and 1733 true negatives. Accuracy, Specificity and Sensitivity of the proposed models are calculated using the following formulae.

Accuracy=(TP+TN)/(TP+FP+TN+FN)

Sensitivity=(TP/(TP+FN)

Specificity=(TN)/(FP+TN)

The results obtained are presented in Fig. 3 to Fig. 5

Table 1: Confusion Matrix using Random Forest

	Predicted Malicious URLs	Predicted Legitimate URLs
Ground Truth Malicious URLs	1293	162
Ground Truth Legitimate URLS	182	1680

Table 2: Confusion Matrix using Artificial Neural Network

	Predicted Malicious URLs	Predicted Legitimate URLs
Ground Truth Malicious URLs	1246	209
Ground Truth Legitimate URLS	155	1707

Table 3: Confusion Matrix using Support Vector Machine

	Predicted Malicious URLs	Predicted Legitimate URLs
Ground Truth Malicious URLs	1254	209
Ground Truth Legitimate URLS	129	1733

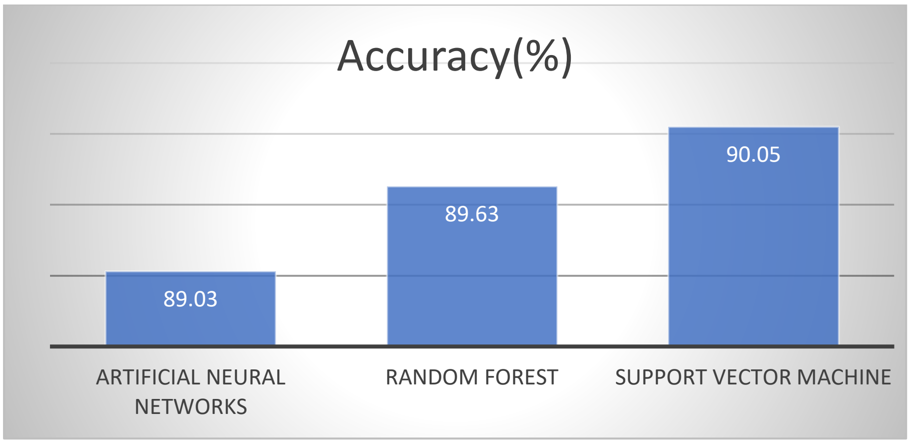

Fig. 3 Accuracy of the proposed Models

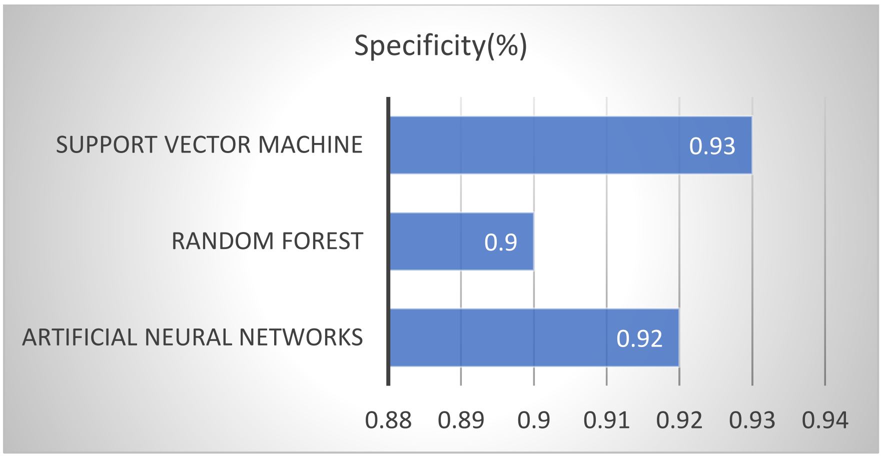

Fig. 4 Specificity of the proposed Models

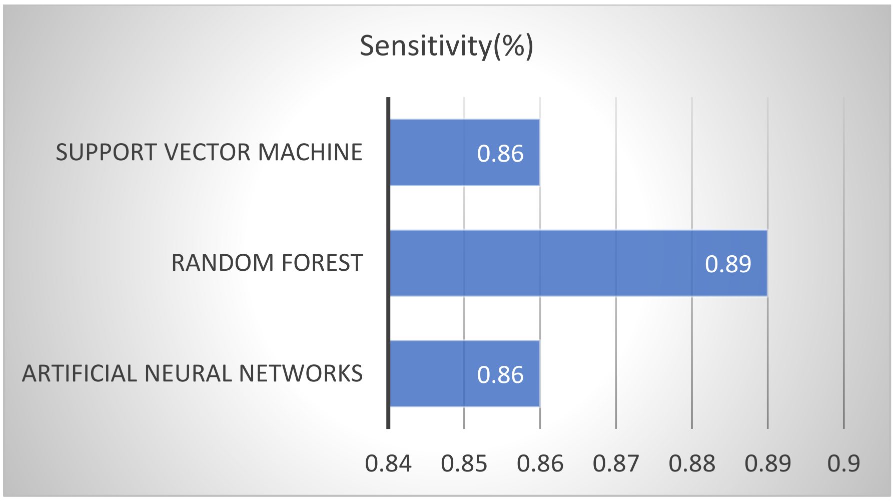

Fig. 4 Sensitivity of the proposed Models

[5] CONCLUSION & FUTURE SCOPE

This paper emphasizes on how Malicious URLs are huge threat to the security and safety of the web and the research importance of malicious URL detection. We reviewed some of the prevention and detection approaches for malicious URL detection in the Literature. Three machine learning models are developed and their performance is evaluated using Phishing Websites Datasetfrom the UCI Machine Learning Repository. This work can further be extended by building a Chrome extension for detecting malicious web pages. Also, we intend to build the malicious URL detection system as a scalable web service which will incorporate online learning so that new attacks patterns can easily be learned and improve the accuracy of our models.

REFERENCES

[1] Peng Yang, Guangzhzen Zhao, “Phishing Website Detection based on Multidimensional features driven by deep learning, 2019”.

[2] Kang Leng Chiew , Choon Lin Tan , KokSheik Wong , Kelvin S.C. Yong , WeiKing Tion, “A New Hybrid Ensemble Feature Selection Framework for Machine Learning-Based Phishing Detection System,2019”.

[3] Ozgur Koray Sahingoz , Ebubekir Buber , Onder Demir, “Machine Learning Based Phishing Detection from URL, 2020“.

[4] Jun Ho Huh, Hyoungshick Kim,” Phishing Detection with Popular Search Engines: Simple and Effective, 2017“.

[5] Waleed Ali,” Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection, 2017 “.

[6] M. Amaad Ul Haq Tahir, Sohail Asghar, Ayesha Zafar, Saira Gillani, “Fresh Phish: A Framework for Auto-Detection of Phishing Websites, 2017“.

[7] Jian Mao, Jingdon Bian, Tao Wei, Shishi Zhu, “Detecting Phishing Websites via Aggregation Analysis of Page Layouts, 2018“.

[8] Hyunsang Choi, Bin Zhu, Heejo Lee, “Detecting Malicious Web Links and Identifying Their Attack Types, 2015”.

[9] UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences. Available: http://archive.ics.uci.edu/ml

[10] Anoop Joyti Sahoo, and Rajesh Kumar Tiwari “A Novel Approach for Hiding Secret data in Program Files” International Journal of Information and Computer Security. Volume 8 Issue 1, March 2016,

[11] Abu Salim, Sachin Tripathi and Rajesh Kumar Tiwari “A secure and timestamp-based communication scheme for cloud environment” Published in International Journal of Electronic Security and Digital Forensics, Volume 6, Issue 4, 319-332.

[12] Rajesh Kumar Tiwari and G. Sahoo, “A Novel Watermark Scheme for Secure Relational Databases” Information Security Journal: A Global Perspective, Volume 22, Issue 3, July 2013

Author[s] brief Introduction

1. Prof. Radhika Kavuri is working as a Professor in the Department of Information Technology, CBIT, , Hyderabad. Prof. Radhika Kavuri did her B.Tech.(EEE) from VRSEC, Vijayawada. She completed her Post Graduation in Computer Science and Engineering (CSE) from JNTU, Hyderabad. She received her Doctorate from Osmania University for her Research work titled "Efficient Mobile-Centric Vertical Handoff Decision Models for Heterogeneous Wireless Networks". She has a total of about 23 years of experience in both Industry and Academia, Prof. Radhika's research interests include Mobile Computing, Cloud Computing, Machine Learning, Decision Support Systems (Game Theory, Multiple Criteria Decision Making-MCDM, Analytic Hierarchy Press-AHP, PCA for Dimensionality Reduction), Blockchain Technology.

Corresponding Address-
(Pin code and Mobile is mandatory)

Dr. K. Radhika,
Professor & Head,
Department of Information Technology,
Chaitanya Bharathi Institute of Technology(A),
Gandipet, Hyderabad.