iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50907

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Safeguarding your Data from Malicious URLs using Machine Learning

Koppula Mounika[1], Tunuguntla Naga Harshitha[2], Shaik Kashipha[3], Vadithe Dhana Lakshmi Bai[4], Batthula Venkata Lakshmi[5], Asst. Prof. Mrs. Bethapudi Haritha[6]
[1, 2, 3, 4, 5, 6]*Dept of Computer Science and Technology, BWEC, Bapatla, Andhra Pradesh, India*

Abstract: ***Malicious URLs host a wide range of unwanted content and can be extremely dangerous to potential victims. Thus, a quick and effective detection method is required. The topic of identifying harmful URLs based on data gleaned from URLs using machine learning methods is the main subject of this thesis. The simplest method of obtaining sensitive information from unwitting people is through a phishing attack. The goal of phishers is to obtain crucial data, such as username, password, and bank account information. Cybersecurity professionals are now looking for stable and reliable detection techniques to detect phishing websites. In order to distinguish between legal and phishing URLs, this article uses machine learning technology. It extracts and analyses many aspects of both types of URLs. Algorithms such as Support Vector Machine, Decision Tree, and Random Forest are used to identify phishing websites. By evaluating each algorithm's accuracy rate, false positive and false negative rates, the study aims to identify phishing URLs and identify the best machine learning method.***
Keywords: URL; SVM; malware; classification model; malicious URL detection; feature extraction; feature selection; machine learning

I. INTRODUCTION

As the name suggests, a malicious URL can only do harm. The usual motivation is to achieve malicious goals such as pushing a political agenda, stealing confidential information about individuals or companies, or simply making a quick buck.to fraudulent sites. Both fake and genuine websites can contain dangerous links, this should be taken in to account.

Online services are an integral part of today's business, school, banking and personal life. With their growing popularity, the number of malicious websites is increasing. A malicious website contains unwanted content designed to collect sensitive data or install malicious software on your computer. Some user interaction is usually required, but Drive downloads install the malware automatically without asking your permission.

A. Motivation

The motivation of our project is to increase the accuracy of predicting whether a particular hosted website is malicious or benign. In these projects, we implement machine learning model like supervised learning and we use binary classifier like SVM (support vector machine) and logistic regression.

B. Objective

Web applications are becoming more and more vulnerable to threats that take advantage of their flaws. Web application flaws were exploited by an attacker to compromise URLs and utilise them for despicable ends. For instance, attackers have attacked websites using URL. Attackers add a redirect code to a compromised URL so that the user is taken to a malicious Site automatically. Additionally, these malicious URLs cause the user to download a malicious programme like a botnet into their computer, which allows the attacker to gather private data like banking information and contact details.

The objective of our project is to detect Malicious URL based on the technology which can help users identify malicious URLs and prevent users from being attacked by malicious URLs. Traditionally, malicious URL detection research has applied blacklist-based methods to detect malicious URLs.

C. Existing System

The current model is a blacklisting approach where it only includes a list of malicious URLs. So, when the user enters the URL, it will check this blacklist if there is one, it will send a warning to the user that it is malicious code that harms the user's information.

Likewise, new URLs will be generated daily, which will prevent them from detecting new threats, and their downside is that they do not apply to a large volume of data sets. And it is not suitable for real time analysis.

D. Proposed System

The Proposed System is used to classify the given URL is Legitimate or malicious by using Machine Learning. In this approach we try to analyse the information of a URL and its corresponding web pages or web pages, extract good representations of the function of the URLs and train a predictive model of the data. training malicious and benign URLs. There are two types of functions that can be used - static functions and dynamic functions**.** In static analysis, we perform web page analysis based on available information without running the URL. By using Supervised learning techniques such as SVM, Random Forest, Logistic Regression and Adaboost we can detect URLs in real time and it can be performed out of the dataset also. The advantages of the proposed system is new feature will be added such as sending mails to the user whether it benign or malicious.

II. LITERATURE SURVEY

1) CANTINA: A content-based approach to detecting phishing web sites

AUTHORS: Jason Hong, Yue Zhang, Lorrie Cramer

Phishing is spoofing a website for the purpose of tracking and stealing sensitive information from online users. Attackers fool users with social engineering techniques such as SMS, voice, email, websites, and malware. In this article, we implemented a desktop application called Phish Shield, which focuses on phishing page URLs and website content.

2) Performance study of classification techniques for phishing URL detection.

AUTHORS: Pradeepthi K V and Kannan

This article presents an overview of the research work done on different researchers' classification techniques to detect phishing URLs. The tests were performed using 4,500 URLs and several classification algorithms. Observational results show that tree classification gives the highest accuracy.

III. SYSTEM MODEL

A. Data Exploration

We've collected this dataset that includes a large number of malicious URLs so we can develop a machine learning-based model to identify malicious URLs, we can stop them before they attack to infect computer, systems or spread over the Internet. The dataset contains attributes extracted from web pages that can be used to classify websites as malicious or benign. The dataset also includes raw page content, including JavaScript code that can be used as unstructured data.

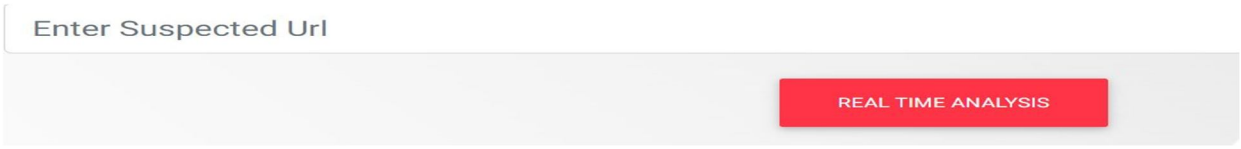
While there are many features one can use to classify whether a website is spam or not, this project aims to use only URLs and limited metadata information for classification. see if the web pages are spam or not. This way, potentially malicious URLs can be de-prioritized during crawling, and these resources can be used to crawl more useful pages that are less likely to be malicious.

B. Modules

1) Uploa Dataset

	A	B	C	D	E	F	G	H	I
1	index	having_IPha	URLURL_Le	Shortining_S	having_At_S	double_slas	Prefix_Suffi	having_Sub_	SSLfinal_Sta
2	1	-1	1	1	1	-1	-1	-1	-1
3	2	1	1	1	1	1	-1	0	1
4	3	1	0	1	1	1	-1	-1	-1
5	4	1	0	1	1	1	-1	-1	-1
6	5	1	0	-1	1	1	-1	1	1
7	6	-1	0	-1	1	-1	-1	1	1
8	7	1	0	-1	1	1	-1	-1	-1
9	8	1	0	1	1	1	-1	-1	-1
10	9	1	0	-1	1	1	-1	1	1
11	10	1	1	-1	1	1	-1	-1	1
12	11	1	1	1	1	1	-1	0	1
13	12	1	1	-1	1	1	-1	1	-1
14	13	-1	1	-1	1	-1	-1	0	0
15	14	1	1	-1	1	1	-1	0	-1
16	15	1	1	-1	1	1	1	-1	1

2) Static



C. Algorithms And Techniques

1) Random Forest

Random Forest algorithm is a powerful and versatile supervised machine learning algorithm and which helps to grows and combines multiple decision trees to create a "forest." It can be used for both classification and regression problems in R and Python. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

2) Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

D. Model Evaluation And Validation

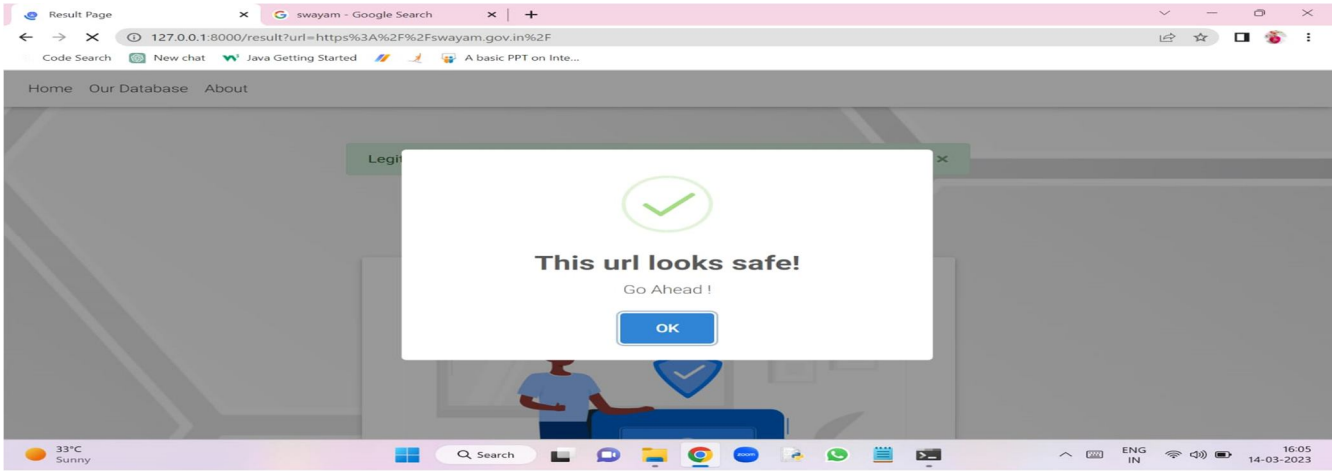
We executed and evaluated our experiments on the models created by SVMlight and TensorFlow library, one representative of batch learning, and one of the online learning approaches. When we used all features and a training set of size 352 096 URL samples, both achieved accuracy more than 97.3%. For this project, we need to ensure that our models can detect as many of the malicious websites as possible, even at the cost of predicting that some good websites are malicious. In fact, declaring a malicious website harmless is very expensive, while declaring a harmless website malicious is not that expensive.

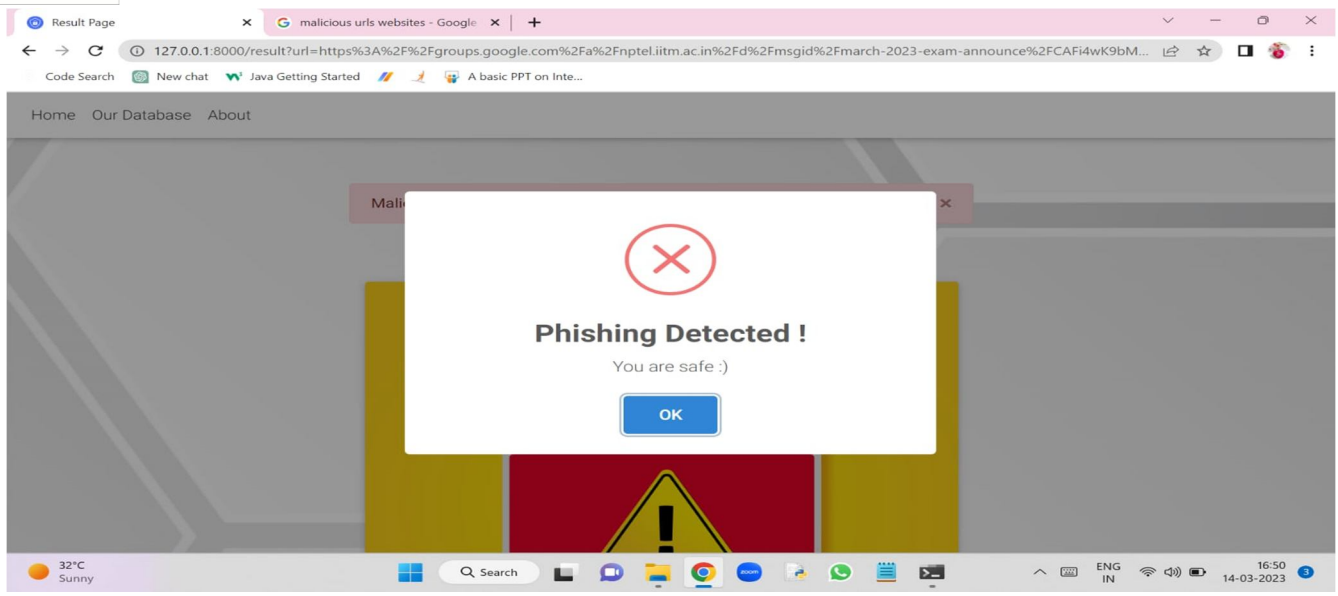
IV. RESULTS AND ANALYSIS

The last chapter is dedicated to analysis and evaluation of the results of classification malicious URLs with proposed methods. We decided to focus on 4 different issues, specifically study the impact of selection different features, set of data, learning method or different amount of sample data. For evaluation of models, we used two metrics, accuracy and loss. Accuracy is the ratio of the correct predictions to the total number of test samples.

For binary classification, accuracy can be calculated as follows:

$$\frac{TP+TN}{TP+TN+FP+FN}$$





V. CONCLUSION

With the development of technology and computer systems, people exchange information on the Internet attracts people because of the convenience of the services they provide daily and moreover, they do many things other related to daily life. During these processes, the user is provided with important information and information such as a descriptive username and password.

In this section, we have presented a large and organized study on the detection of malicious URLs using machine learning techniques. represent new features and design new learning algorithms to identify malicious URL detection tasks**.** We also identified the requirements and challenges to develop Malicious URL Detection as a true cybersecurity application service.

VI. FUTURE WORK

Due to the use of static datasets for testing and analysis, we could not include the host or rank properties in the list of used features. These properties may change over time, and thus their extraction from the URL after a long time may not reflect the real state at the time of collection. As future work, a tool or way for dynamic data reception can be added, allowing to extract and use more features. Even though the machine-learning technologies we used achieved very good results, comparing multiple models and settings can bring more interesting insights and results.

REFERENCES

[1] Bannur, S. N., Saul, L. K., & Savage, S. (2011). Judging a site by its content: learning the textual, structural, and visual features of malicious web pages. Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence. 10.1145/2046684.2046686

[2] S. Purkait, “Phishing counter measures and their effectiveness– literature review,” Information Management & Computer Security, vol. 20, no. 5, pp. 382–420, 2012.

[3] M. Khonji, Y. Iraqi, and A. Jones, “Phishing detection: a literature survey,” IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.

[4] H. B. Kazemian and S. Ahmed, “Comparisons of machine learning techniques for detecting malicious webpages,” Expert Syst. Appl., vol. 42, no. 3, pp. 1166–1177, 2015.

[5] R. Heartfield and G. Loukas, “A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks,” ACM Computing Surveys (CSUR), vol. 48, no. 3, p. 37, 2015.

[6] A. Firdaus, N. B. Anuar, M. F. A. Razak, and A. K. Sangaiah, “Bio-inspired computational paradigm for feature investigation and malware detection: interactive analytics,” Multimed. Tools Appl., 2017 investigation and malware detection: interactive analytics,” Multimed. Tools Appl., 2017

[7] S. G. Selvaganapathy, M. Nivaashini, and H. P. Natarajan, “Deep belief network-based detection and categorization of malicious URLs,” Inf. Secur. J., vol. 27, no. 3, pp. 145–161, 2018.

[8] D. R. Patil and J. B. Patil, “Feature-based Malicious URL and Attack Type Detection Using Multi-class Classification,” Int. J. Inf. Secur., vol. 10, no. 2, pp. 141–162, 2018.

[9] M. Hazim, N. B. Anuar, M. F. Ab Razak, and N. A. Abdullah, “Detection opinion spams through supervised boosting approach,” PLoS One, vol. 13, no. 6, pp. 1–23, 2018.

BIOGRAPHIES

Bethapudi Haritha M. Tech, Asst. Professor, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India.

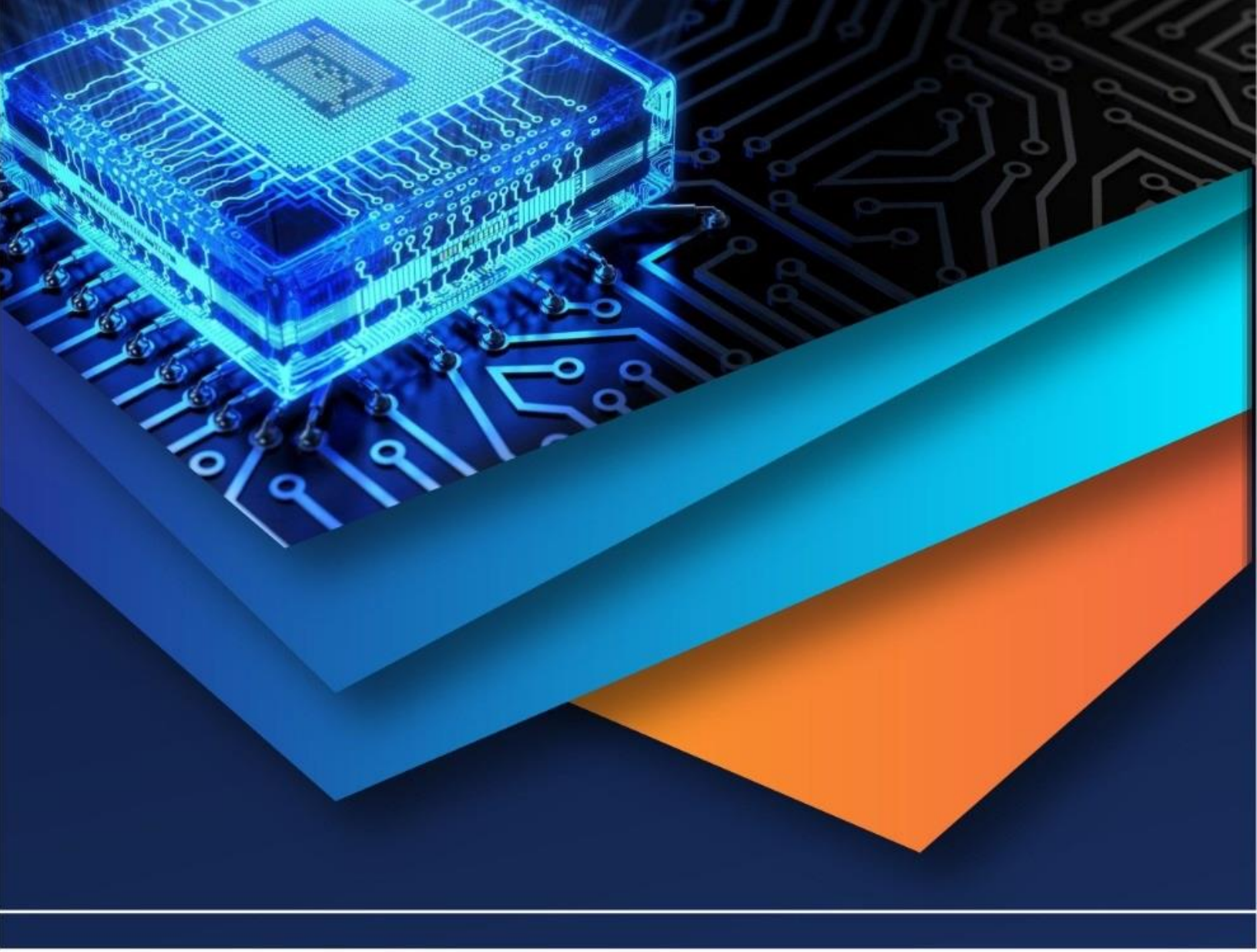
Koppula Mounika [B. Tech], Student, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India

Tunuguntla Naga Harshitha [B. Tech], Student, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India

Shaik Kashipha [B. Tech], Student, Dept of Computer Science and Technology, BWEC, Andhra Pradesh, India.

Vadithe Dhana Lakshmi Bai [B. Tech], Student, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India

Batthula Venkata Lakshmi [B. Tech], Student, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY