IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204261 |

Hazard Detection of Webpages Using Machine Learning

Mrs. Y. Suneetha, M. Tech., (Ph.D.,) G Ranga Swamy, C Jyothi, DS Janani, K Sri Manjunath

Assistant Professor, Dept. of CSE, JNTUA University, Kuppam Engineering College, Andhra Pradesh, India

UG Students, Dept. of CSE., JNTUA University, Kuppam Engineering College, Andhra Pradesh, India

ABSTRACT: Internet surfing has become a vital part of our daily life. So to catch the attention of the users' different browser vendors compete to set up the new functionality and advanced features that become the source of attacks for the intruder and the websites are put at hazard. However, the existing approaches are not adequate to protect the surfers which require an expeditious and precise model that can be able to distinguish between the begnign or malicious webpages. In this research article, we design a new classification system to analyze and detect the malicious web pages using machine learning classifiers such as,gradient boosting algorithm, random forest, support vector machine. naïve Bayes, logistic regression and Some special URL (Uniform Resource Locator) based on extricated features the classifiers are trained to predict the malicious web pages. The experimental results have shown that the performance of the random forest classifier achieves better accuracy of 95% and gradient boosting algorithm achieves better accuracy of 97%.

KEYWORDS: malicious web page, machine learning, detection, URL, malicious websites.

I. INTRODUCTION

With the rapid development of the web, more and more services like internet banking, e-commerce, social networking, shopping, making a bill payment, e-learning, etc. are available to users and they are surfing the internet via browsers or web application. As the browsers are come up with different advanced features and functionalities which leads to risk by losing their personal and sensitive information . As the naïve users are not aware of the different malware so they are easily trapped by the intruder by just a single click on the malicious web sites which allows the invaders to detect the vulnerabilities on the web page and inject the payloads to get remote access to victim's web page. Therefore, the precise identification of web pages in an ever-growing web environment is very important. Blacklisting services were embedded in the browsers to face the challenges but it has several disadvantages like incorrect listing. In this article, we explore a self-learning approach to classify the web page based on a small feature set. We use four machine learning classifiers to classify the web site into two classes benign and malicious web pages.

II. RELATED WORK

The implementation of the project is based on detection of hazard webpages using machine learning algorithms. We have divided into two be two classes like benign or malicious webpages. We have used some classifiers are random forest, support vector machine, naive Bayes and logistic regression. The implementation of this project is based on Machine Learning algorithms like Support Vector Machine (SVM) because machine learning models can be used to predict the result in less amount of time. Support Vector Machine (SVM) is one of the prominent algorithms that can be used for classification purpose. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X. Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. Bayes Theorem provides a way that we can calculate the probability of a piece of data belonging to a given class, given our prior knowledge. The language we have used for the development of the project is Python due to it is more supportable for developing the Machine Learning model due to its built-in libraries. The development of the project has done using PyCharm tool and uses high-level Python web framework that enables rapid development of

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204261 |

secure and maintainable websites. This project consists of many files regarding to the detection of hazard web pages. The completion time for the project is approximately 4 to 6 weeks.

III. LITERATURE SURVEY

Context-sensitive and keyword density-based supervised machine learning techniques for malicious webpage detection.

Authors: Altay, Betel, Tansel Dokeroglu, and Ahmet Cosar. Conventional malicious webpage detection methods use blacklists in order to decide whether a webpage is malicious or not. The blacklists are generally maintained by third-party organizations. However, keeping a list of all malicious Web sites and updating this list regularly is not an easy task for the frequently changing and rapidly growing number of webpages on the web. In this study, we propose a novel context-sensitive and keyword density-based method for the classification of webpages by using three supervised machine learning techniques, support vector machine, maximum entropy, and extreme learning machine. Features (words) of webpages are obtained from HTML contents and information is extracted by using feature extraction methods: existence of words, keyword frequencies, and keyword density techniques. The performance of proposed machine learning models is evaluated by using a benchmark data set which consists of one hundred thousand webpages.

WebMon: ML-and YARA-based malicious webpage detection

Authors: Kim, Sungjin, Jinkook Kim, Seokwoo Nam, and Dohoon Kim. Attackers use the openness of the Internet to facilitate the dissemination of malware. Their attempts to infect target systems via the Web have increased with time and are unlikely to abate. In response to this threat, we present an automated, low-interaction malicious webpage detector, WebMon, that identifies invasive roots in Web resources loaded from WebKit2-based browsers using machine learning and YARA signatures. WebMon effectively detects hidden exploit codes by tracing linked URLs to confirm whether the relevant websites are malicious. WebMon detects a variety of attacks by running 250 containers simultaneously. In this configuration, the proposed model yields a detection rate of 98%, and is 7.6 times faster (with a container) than previously proposed models.

IV. METHODOLOGY

In this section, we provide a detailed discussion about our proposed approach to identifying the malicious web page. To address the drawback of previous studies we design a new web site classification system based on the URL features to identify malicious websites which are shown in Fig.1. In step 1 according to our requirements, we have collected a dataset from the internet source [14] contains both the malicious and benign web sites. In step 2 data is reduced to filter and data cleaning by selecting a few relevant attributes out of 21 attributes in total from the dataset. In step 3 we have designed our dataset consisting of 7 URL features and 1782 records. Then we manually divide the dataset into two sets; one is for training set made up of 812 records and another is for testing set consists of 970 records. In step 4 machine learning classifiers are trained to create a Machine Learning (ML) model with the help of the training set. In the final step, the ML model is verified with the testing set to obtain our required result. If the Type attribute value contains 0 means the inputted URL is a benign web site else it is a malicious web site. The following subsection explains the three basic components of our approach: the dataset, feature extraction, and machine learning classifiers elaborately.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204261 |

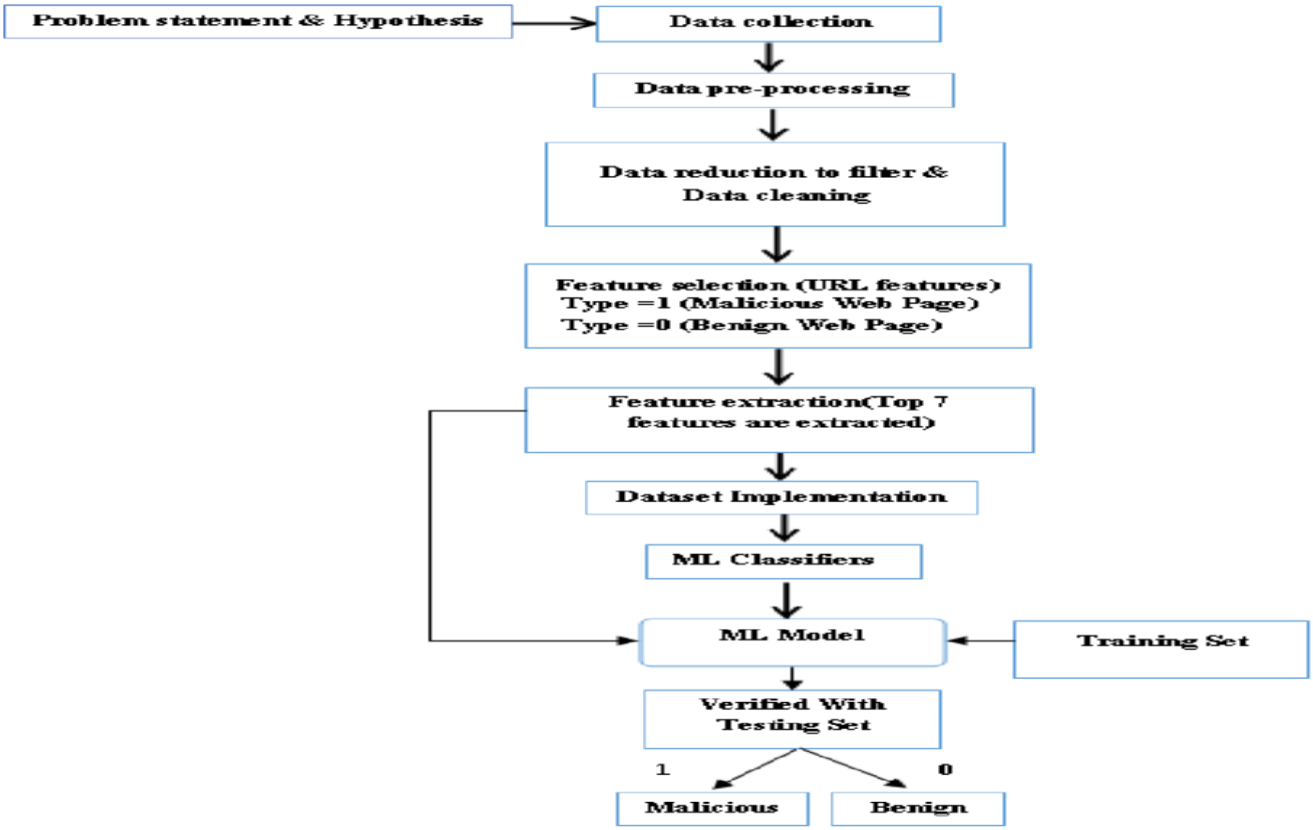

Fig.1 Proposed Approach for Malicious Web Page Detection

Sl. No	Features	Descriptions
1	URL	It is a reference to a web resource that specifies it's location on a system.
2	URL-LENGTH	The maximum length of a URL is 2048 characters. An URL doesn't exceed 2000 characters. So that it can parse with a browser.
3	NUMBER-SPECIAL-CHARACTERS	Special characters appear in the URL. Example: "%","&","#" , "." etc.
4	CONTENT-LENGTH	Represents size of total characters in URL
5	SOURCE-APP-PACKETS	Indicates the packets sent from the Client.
6	REMOTE-APP-PACKETS	Indicates the packets received from the server.
7	TYPE	Represents a target attribute. It takes value '1' for malicious websites and '0' for benign websites.

Fig.2 Effect of Source-App and Remote-App packets feature in malicious web page detection .

V. MODEL PHASES

1. Data Collection
2. Building Dataset
3. Building the models
4. Select the model for training
5. Evaluation the model

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204261 |

1. Data Collection:
This is the first real step towards the real development of a machine learning model, collecting data. This is a first and critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform. We have to select the data which is suitable for the fast development of the model and dataset that consists of a smaller number of parameters. For the developing of the project, we have collected a lot of raw data and we have collected different types of URLS from various resources. These raw data consist of many details regarding the URLS. We have taken the features from the gathered data. The collected data can be used for the training the model.

2. Building Dataset:
The dataset we have used for the development of this project is .CSV file which consists of various features which are extracted from the various resources. In the proposed dataset, we have different types of parameters which is taken as features for URLS. These features are used to extract the data it removes the outliers, null values and gives the data.

3. Building the models:
After the collection of data and creation of dataset the next step is developing a model for the detection of either URLS are malicious or benign. For building the models we have used a machine learning algorithm like SVM, Logistic regression, Random Forest, Naive Bayes Algorithm,Gradient Boosting Algorithm. To get the final result we are going to build a model using the available dataset. While building a model for the implementation of project we have to perform the following actions: 1. Import all the modules which are necessary for developing the project 2. Building the architecture required for the training purpose 3. Selection of data type like Images/Text/CSV .

4. Select the model for training:
After all, doing all the necessary activities for building the model, we have to do the further implementations. While building a model for dataset we have to make sure that data should be relevant, uniform, representative, diverse. We are going to build a model using Support Vector Machine (SVM). As Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers' detection. We have to train the given data by selecting one of the developed algorithm. We have to train the model based up on their features. The training for each model can be given by selecting the respective machine learning algorithm.

5. Evaluation of model:
The last step is to evaluate the model whether our trained model is working properly for the user input. For the evaluation purpose of the proposed model, we have done some experimental analysis on some evaluation metrices like confusion matrix, precision and recall, F1 score, classification accuracy.

VI. EXPERIMENTAL RESULTS ANALYSIS

Various experiments have been carried out by implementing the classification algorithms such as logistic regression, random forest, Gaussian Naïve Bayes, and support vector machine. All the experiments were coded and tested in Jupyter Notebook which is an interactive python environment for data science. With it's integrated support for Pandas, Scikit-Learn, Matplotlib, markup language, plots, and tables, a much more appealing and understandable presentation of the flow of the code can be made.

The performance of the four machine learning classifiers. Table III. Lists the performance. We have used the performance metric, accuracy to evaluate the detection performance because it correctly labels a web page. So to obtain the best results, accuracy performance metric plays a vital role. We notice that the machine learning classifier RF obtains higher accuracy of 95%, whose performance is better than the other classifiers on malicious web page detection. The experimental result shows that our method achieves superior performance even by selecting a small set of URL based features. The performance comparison of different classifiers is depicted in Table III.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204261 |

Table II Performance Comparison of different Classifiers.

Classifiers	Evaluation Criteria(Accuracy)
GBA	97%
SVM	89%
LR	91%
RF	95%

VII. CONCLUSION AND FUTURE WORK

Malicious web page identification is an emerging topic in cybersecurity. Though several research studies have been performed relating to the issues of malicious web page detection these are very costly as they consume more time and resources. In this research article, we employed a new web site classification system based on URL features to predict the web pages as malicious or benign using machine learning algorithms. The machine learning classifiers Random Forest (RF) and Gradient Boosting Algorithm achieves a higher accuracy of 95% and 97% . The experimental results have shown that our method can perform effectively for detecting the malicious web page.

REFERENCES

1. Tao, Wang, Yu Shunzheng, and Xie Bailin. "A novel framework for learning to detect malicious web pages." In 2010 International Forum on Information Technology and Applications, vol. 2, pp. 353-357. Ieee, 2010.
2. Eshete, Birhanu, Adolfo Villafiorita, and Komminist Weldemariam. "Malicious website detection: Effectiveness and efficiency issues." In 2011 First SysSec Workshop, pp. 123-126. IEEE, 2011..
3. Aldwairi, Monther, and Rami Alsalman. "Malurls: A lightweight malicious website classification based on url features." Journal of Emerging Technologies in Web Intelligence 4, no. 2 (2012): 128-133..
4. Yoo, Suyeon, Sehun Kim, Anil Choudhary, O. P. Roy, and T. Tuithung. "Two-phase malicious web page detection scheme using misuse and anomaly detection." International Journal of Reliable Information and Assurance 2, no. 1 (2014): 1-9.
5. Hwang, Young Sup, Jin Baek Kwon, Jae Chan Moon, and Seong Je Cho. "Classifying malicious web pages by using an adaptive support vector machine." Journal of Information Processing Systems 9, no. 3 (2013): 395-404. .

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details