IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, May 2024

Hazard Identification and Detection Using Machine Learning Approach

Saraswati P, Sainath G, Mohammed Rahiel
Department of Computer Science and Engineering
Rao Bahadur Y Mahabaleswarappa Engineering College, Bellary, Karnataka, 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, 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% in comparison to other machine learning classifiers*

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.

"The rest of the research work is planned as follows: Section II presents related work, the methodology is discussed in section III, experimental result analysis is depicted in Section IV and Section V contains the conclusion of the research work and suggests some future work".

II. LITERATURE SURVEY:

[1] Altay, Betel, Tansel Dokeroglu, and Ahmet Cosar. "Context-sensitive and keyword density-based supervised machine learning techniques for malicious webpage detection." Soft Computing 23, no. 12 (2019): 4177-4191.

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

DOI: 10.48175/IJARSCT-18025

which consists of one hundred thousand webpages. Experimental results show that the proposed method can detect malicious webpages with an accuracy of 98.24%, which is a significant improvement compared to state-of-the-art approaches.

[2] Kim, Sungjin, Jinkook Kim, Seokwoo Nam, and Dohoon Kim. "WebMon: ML-and YARA-based malicious webpage detection." Computer Networks 137 (2018): 119-131.T
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. Most importantly, Web Môn's focus on extracting malicious paths in a domain is a novel approach that has not been explored in previous studies.

III. METHODOLOGY SECTION

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 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.

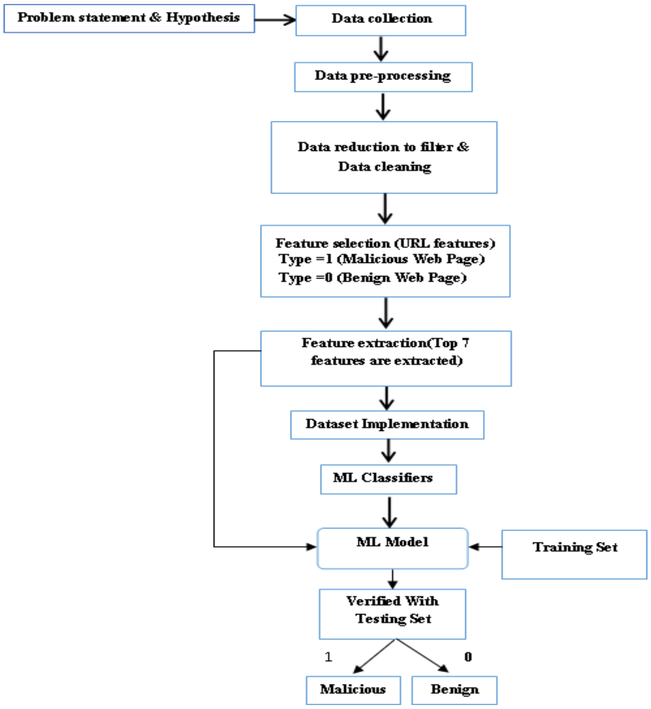
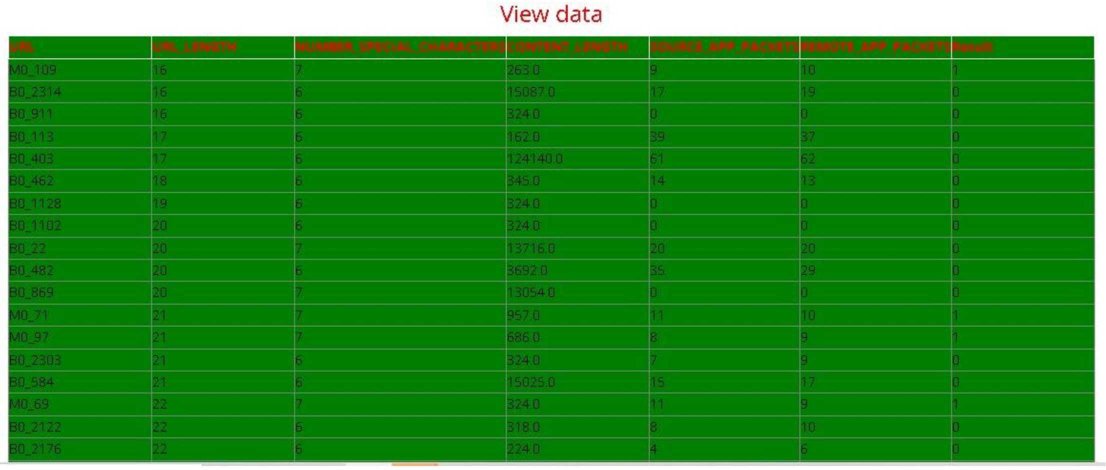


Fig.1 Proposed Approach for Malicious Web Page Detection

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 benignweb site else it is a malicious website.

A. Dataset:

The selection of datasets has a great influence on the quality of classification. We require to select an appropriate dataset as well as possible So we fill this gap by collecting a URL dataset from the Kaggle database [14] which is composed of both malicious and benign websites and it has 1782 records and 21 features. Out of 1782 records, 812 records are used. A snapshot of our dataset is depicted in Fig. 2.



View data

URL	URL_LENGTH	NUMBER_SPECIAL_CHARACTERS	CONTENT_LENGTH	SOURCE_APP_PACKETS	REMOTE_APP_PACKETS	Result
M0_109	16	7	263.0	9	10	1
B0_2314	16	6	15087.0	17	19	0
B0_911	16	6	324.0	0	0	0
B0_113	17	6	162.0	39	37	0
B0_403	17	6	124140.0	61	62	0
B0_462	18	6	345.0	14	13	0
B0_1128	19	6	324.0	0	0	0
B0_1102	20	6	324.0	0	0	0
B0_22	20	7	13716.0	20	20	0
B0_482	20	6	3692.0	35	29	0
B0_869	20	7	13054.0	0	0	0
M0_71	21	7	957.0	11	10	1
M0_97	21	7	686.0	8	9	1
B0_2303	21	6	324.0	7	9	0
B0_584	21	6	15025.0	15	17	0
M0_69	22	7	324.0	11	9	1
B0_2122	22	6	318.0	8	10	0
B0_2176	22	6	224.0	4	6	0

Fig.2 A snapshot of our dataset.

B. Machine Learning Classifiers:

There are a lot of methods for realizing the classifiers. We select four machine learning algorithms to build our classifiers.

“Logistic Regression is a supervised machine learning technique. **Random Forest** is an ensemble learning method **Gaussian Naïve Bayes** is a simple, effective, and commonly used machine learning classifier. **Support Vector Machine** is a training algorithm for learning classification and regression rules fromdata”.

IV. EXPERIMENTAL RESULTS

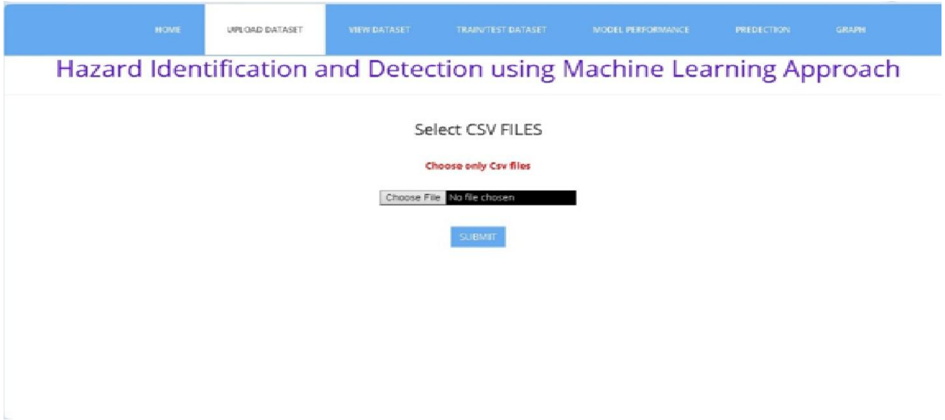
Various experiments have been carried out by implementing the classification algorithms such as logistic regression, random forest, Gaussian Naïve Bayes, and support vector machine. All the experiments were coded and tested in Jupyter Notebook [13] which is an interactive python environment for data 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.

We then compare the performance of the four machine learning classifiers. 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.

DOI: 10.48175/IJARSCT-18025

Hazard Identification and Detection using Machine Learning Approach

Fig.3 Home page



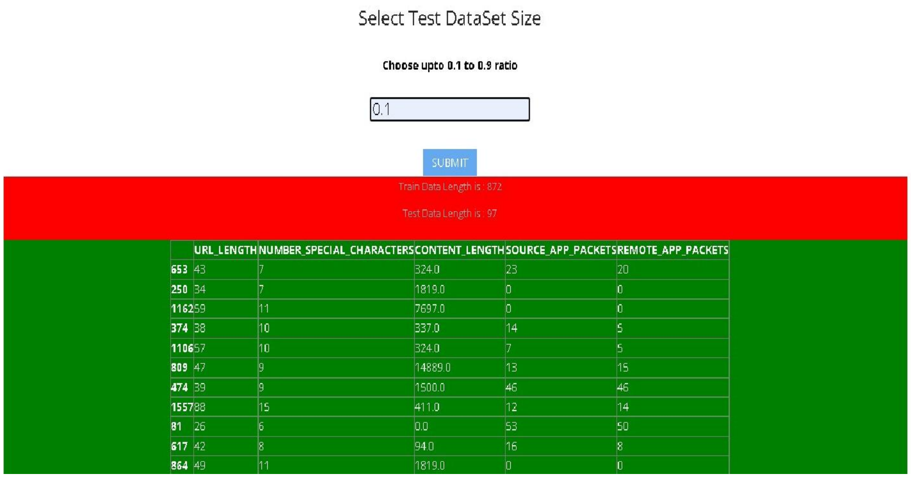
Hazard Identification and Detection using Machine Learning Approach

Select CSV FILES

Choose only Csv files

No file chosen

Fig.4 Uploading Dataset



Select Test DataSet Size

Choose upto 0.1 to 0.9 ratio

0.1

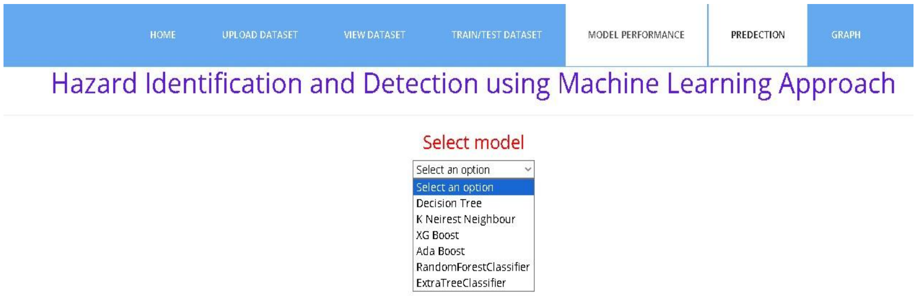
Train Data Length is : 872

Test Data Length is : 97

	URL_LENGTH	NUMBER_SPECIAL_CHARACTERS	CONTENT_LENGTH	SOURCE_APP_PACKETS	REMOTE_APP_PACKETS
653	43	7	324.0	23	20
250	34	7	1819.0	0	0
1162	59	11	7697.0	0	0
374	38	10	337.0	14	5
1106	57	10	324.0	7	5
809	47	9	14889.0	13	15
474	39	9	1500.0	46	46
1557	88	15	411.0	12	14
81	26	6	0.0	53	50
617	42	8	94.0	16	8
864	49	11	1819.0	0	0

Fig.5 Train/Test Data

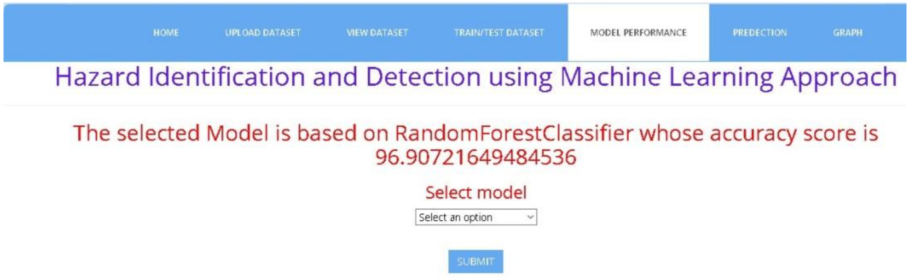
DOI: 10.48175/IJARSCT-18025



Hazard Identification and Detection using Machine Learning Approach

Select model

Fig.6 Selecting the Model

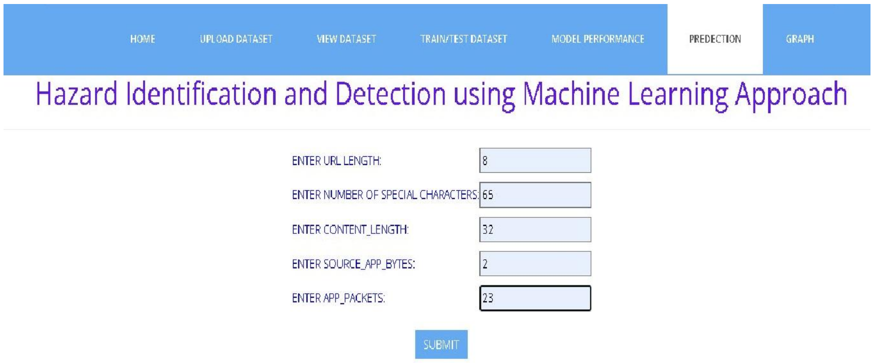


Hazard Identification and Detection using Machine Learning Approach

The selected Model is based on RandomForestClassifier whose accuracy score is 96.90721649484536

Select model

Fig.7 After Training/Testing the Models



Hazard Identification and Detection using Machine Learning Approach

ENTER URL LENGTH: 8

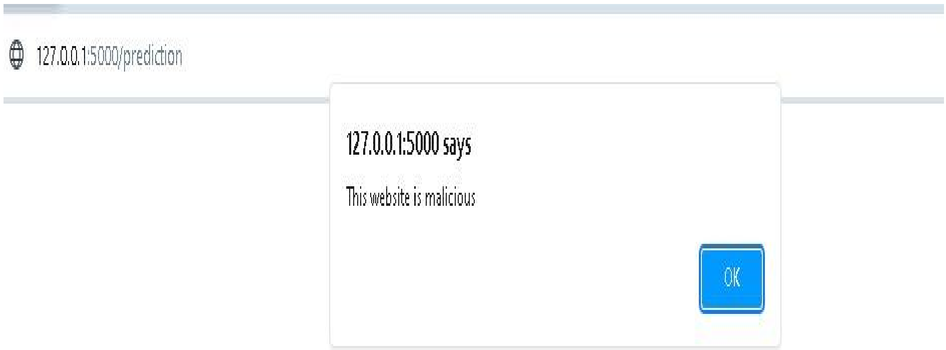
ENTER NUMBER OF SPECIAL CHARACTERS: 65

ENTER CONTENT_LENGTH: 32

ENTER SOURCE_APP_BYTES: 2

ENTER APP_PACKETS: 23

Fig.8 Prediction based on features



127.0.0.1:5000 says

This website is malicious

Fig.9 Predicting Whether the Website is Malicious or Benign

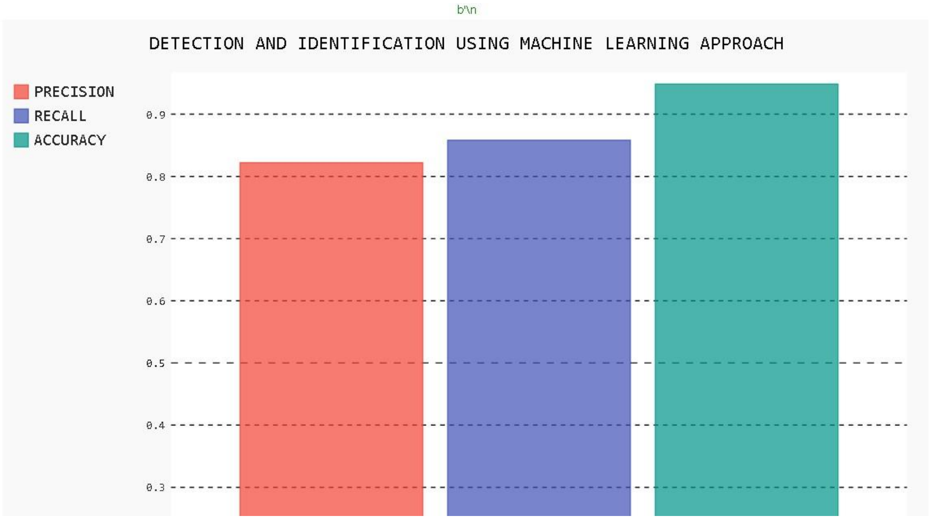


Fig.10 Graph

V. CONCLUSION

Malicious web page identification is an emerging topic in cyber security. 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 RandomForest(RF) achieves a higher accuracy of 95%. The experimental results have shown that our method can perform effectively for detecting the malicious web page. In future work, it has been planned to expand the feature sets and analysis using various sources of data to enhance the classifier performance.

REFERENCES

[1]. Tao, Wang, Yu Shunzheng, and Xie Bailin. "A novel framework for learning to detect malicious web pages." In 2020 International Forum on Information Technology and Applications, vol. 2, pp. 353-357. Ieee, 2020.

[2]. Eshete, Birhanu, Adolfo Villafiorita, and Komminist Weldemariam. "Malicious website detection: Effectivenessand efficiency issues." In 2021 First SysSec Workshop, pp. 123-126. IEEE, 2021..

[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 (2022): 128-133..

[4]. 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 (2023): 395-404.

[5]. Yue, Tao, Jianhua Sun, and Hao Chen. "Fine-grained mining and classification of malicious Web pages." In 2023 Fourth International Conference on Digital Manufacturing & Automation, pp. 616-619. IEEE, 2023

[6]. Krishnaveni, S., and K. Sathiyakumari. "SpiderNet: An interaction tool for predicting malicious web pages." In International Conference on Information Communication and Embedded Systems (ICICES2014), pp. 1-6. IEEE, 2021.