IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

# HAZARD IDENTIFICATION AND DETECTION USING MACHINE LEARNING APPROACH

**[1] B.v.s. Pavan Kumar**

[1]M.Tech

[1]Department of IT&CA, [1]Andhra University College of Engineering,

**[2] A.Mary Sowjanya**

[2]Associate Professor

[2]Department of CS&SE, Visakhapatnam, Andhra Pradesh, India

## ABSTRACT

In an ever-evolving technological landscape, the ability to accurately identify and detect hazards is of paramount importance to ensure safety across various domains. Hazards, encompassing potential threats and risks that can lead to harm or adverse events, necessitate proactive measures for timely intervention. This project focuses on harnessing the power of Machine Learning (ML) techniques to enhance hazard identification and detection processes. Internet browsing has seamlessly integrated into our everyday routines, and in order to engage users, different browser vendors constantly strive to introduce new functionalities and advanced features. Unfortunately, these enhancements also create opportunities for attackers to exploit vulnerabilities, posing risks to websites and users. Current security measures are insufficient in protecting surfers, necessitating the development of a fast and accurate model capable of distinguishing between benign and potentially harmful web pages. In this research paper, I developed a novel classification system utilizing machine learning classifiers, including random forest, support vector machine, naïve Bayes, and logistic regression. for the purpose of examining and identifying malicious web pages. By extracting features from special Uniform Resource Locators (URLs), we train the classifiers to predict whether a webpage is malicious or benign. Experimental results demonstrate that the random forest classifier compared to other machine learning classifiers, achieves an accuracy rate of 95%. Achieving a higher classification accuracy is crucial in enhancing web security and protecting users from potential cyber threats.

***Keywords***- *malicious web page, machine learning, detection, URL, cyber threats.*

## INTRODUCTION

As the internet continues to experience rapid growth, an expanding array of services, including Online banking, electronic commerce, social media networking, bill settlement, and digital learning have become readily accessible to users through web browsers and web applications. However, with the advancement of browsers and their numerous features, there is a growing risk of users' personal and sensitive information being compromised. Many users, especially those who are not well-versed in online security, are unaware of various malware threats. Consequently, they can easily fall prey to intruders by simply clicking on malicious websites. These websites enable attackers to detect vulnerabilities and inject harmful code to obtain remote access to the target's webpage. Thus, it is crucial to accurately identify web pages in this ever-expanding web landscape. To address these challenges, blacklisting services have been integrated into browsers. Our work, we introduce a self-learning approach that relies on a streamlined set of features for the categorization of web pages. The

primary objective is to utilize six machine learning classifiers to distinguish web pages into two distinct categories: benign and malicious.

## LITERATURE SURVEY

Researchers have proposed various techniques, for the identification of malicious web pages by employing methods, encompassing techniques such as blacklisting, static analysis, and dynamic analysis.
Tao et al. [1] introduced an innovative framework that leverages supervised machine learning to autonomously determine the nature of a web page, whether it is malicious or benign. Their classification relied on specific features, and they compiled a dataset of benign web pages for this purpose.
Adware and rami et al. [2] put forth a lightweight self-learning methodology for categorizing malicious web pages, utilizing a framework named MALURL. They utilized Genetic Algorithm (GA) for the training of classifiers designed to detect malicious web pages. Their training dataset included benign websites from Alexa and malicious ones from Phish Tank, resulting in an average system precision of 87%.
Hwang et al. [3] employed the Adaptive SVM (SVM) machine learning technique, known for its ability to effectively adapt to new training data, thereby reducing the risk of misclassifying novel web pages.
Yue et al. [4] introduced a method for classifying malicious web pages, utilizing 30 distinctive features with the assistance of machine learning algorithms such as K-NN and SVM. Their research indicated that the K-NN algorithm outperformed SVM. They implemented two classification models for the detection of malicious web pages and specific threat types.
Yoo et al. [5] introduced two distinct detection methods: misuse detection, which aims to identify known malicious web pages, and anomaly detection, designed to detect previously unidentified malicious web pages. In their experiments, using the RafaBot dataset within the WEKA tool, they achieved a notable detection rate of up to 98%. However, it's worth noting that the false positive rate was relatively high, reaching 30.5%.
Kim et al. [6] introduced WebMon, an automated and minimally interactive malicious webpage detection tool. WebMon leverages machine learning and YARA signatures to discern detrimental components within web resources loaded via WebKit2-based browsers.

## 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 imported packages and downloaded a dataset from the internet source contains both the malicious and benign web sites. In step 2 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 3 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 outcome of our experimentation illustrates the performance of our methodology with an impressive accuracy rate, surpassing 98%, in effectively detecting malicious URLs, while also achieving a precision of over 93% in correctly identifying the specific attack types associated with these URLs.
Additionally, we provide comprehensive analyses of the effectiveness of each group of discriminative features incorporated into our methodology. Furthermore, we delve into an exploration of the potential susceptibilities of these features to evasion techniques, fostering a comprehensive comprehension of the system's strengths and constraints.

## ARCHITECTURE

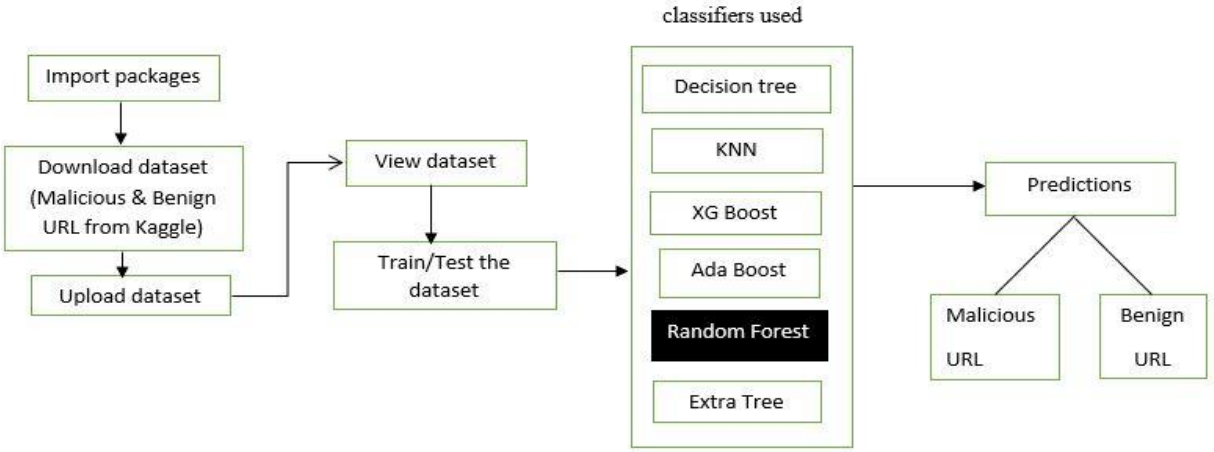
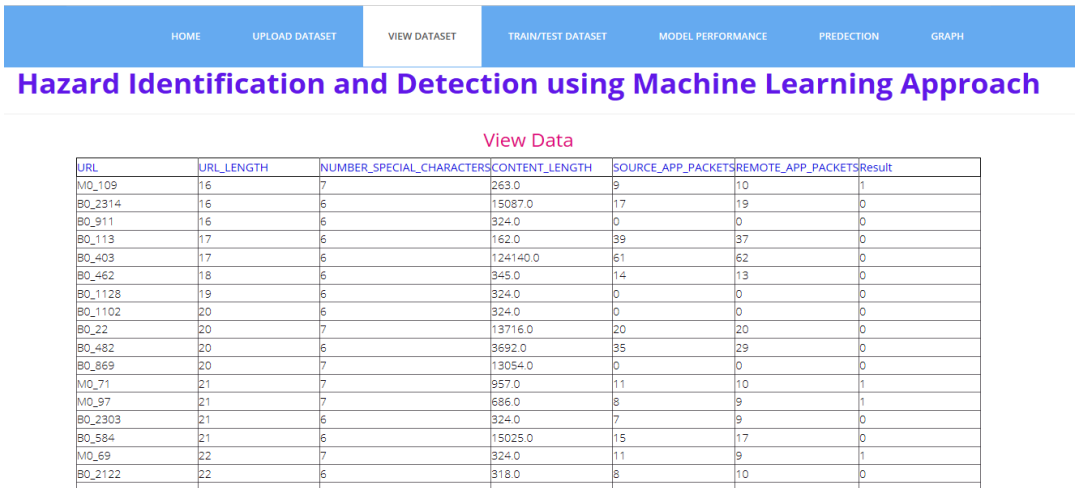



fig 1. Proposed Approach for Malicious Web Page Detection

## RESULT ANALYSIS

### *A. Dataset:*

The provided dataset, sourced from Kaggle database [7], encompasses a collection of attributes characterizing instances of cloud storage access. Each entry comprises features like URL attributes, URL length, the count of special characters, content length, source and remote application packets, culminating in an outcome label with 56 entries, the dataset exhibits a blend of binary classification outcomes (1 or 0), representing public integrity auditing results. While some fields exhibit missing values (NA), the dataset provides valuable insights into factors affecting cloud storage security. Derived from Kaggle, it offers a foundation for research and analysis in identity-based public integrity auditing, contributing to advancements in cloud data protection and privacy preservation.




| 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 |

fig 2. A snapshot of dataset.

## B. *Graph:*

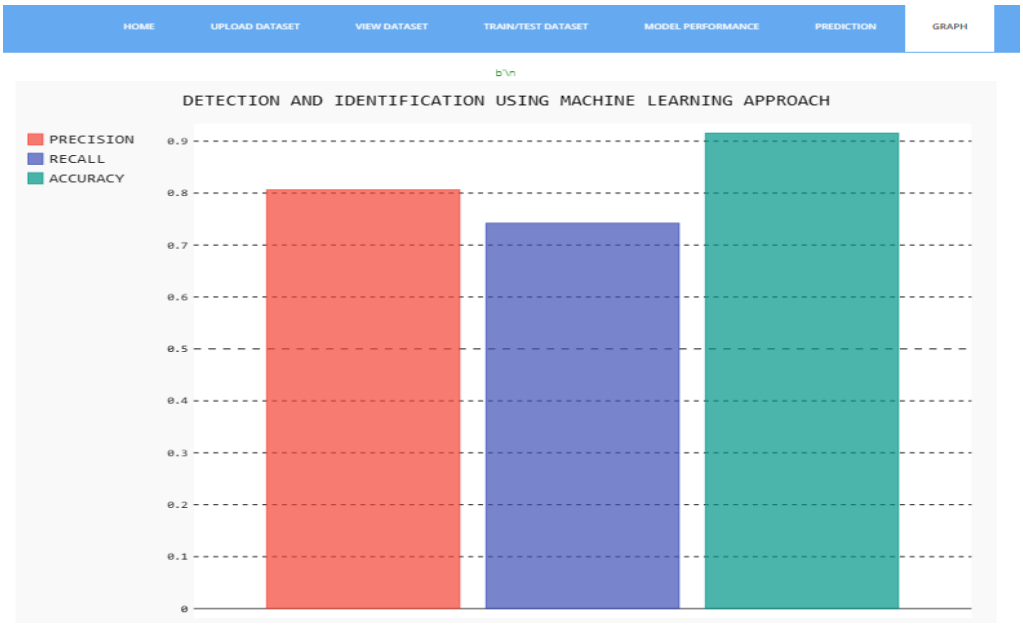



fig 3. Graphical representation of RF approach

## C. *Comparative Analysis of Classifiers Used:*

Various experiments have been carried out by implementing the classification algorithms such as Decision tree, K-Nearest Neighbours (KNN), Extreme Gradient (XG) Boost, Adaptive Boost, Random Forest, Extra Tree Classifiers. All the experiments were coded and tested in PyCharm Software [8] which is an interactive python environment for Machine Learning. With its 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 Six machine learning classifiers. We have used the precision, recall and accuracy to evaluate the detection performance because it correctly labels a webpage. So, to obtain the best results, accuracy performance metric plays a vital role. We notice that the machine learning classifier RF obtains higher accuracy, whose performance is better than the other classifiers on malicious web page detection.

Table: Performance Comparison of different classifiers

| Classifiers | Precision | Recall | Accuracy |
|---|---|---|---|
| Decision Tree | 0.96 | 0.94 | 0.91 |
| K-N Neighbours | 0.93 | 0.98 | 0.92 |
| XG Boost | 0.96 | 0.98 | 0.94 |
| Ada Boost | 0.96 | 0.98 | 0.94 |
| **Random Forest** | 0.95 | 1.00 | **0.95** |
| Extra Tree | 0.94 | 0.98 | 0.93 |

## CONCLUSION

The identification of malicious web pages is a growing concern within the cybersecurity domain. While numerous research endeavours have been committed to addressing the complexities linked with identifying malicious web pages, these initiatives frequently entail substantial time and resource investments. In this research work, we introduce a pioneering website classification system that leverages URL attributes to anticipate the character of web pages, discerning between malign and benign ones by employing machine learning algorithms.

Among the machine learning models employed in our investigation, the Random Forest (RF) classifier distinguishes itself with exceptional performance, achieving an impressive accuracy rate of 95%.

The empirical results gleaned from our experiments unequivocally highlight the effectiveness of our approach in the successful identification of malicious web pages. These findings underscore its potential as an efficient and promising cybersecurity solution, capable of mitigating the risks posed by malicious online content.

## 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] Aldwairi, Monther, and Rami AL Salman. "MalURLs: A lightweight malicious website classification based on URL features." Journal of Emerging Technologies in Web Intelligence 4, no. 2 (2012): 128-133.

[3] 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.

[4] Yue, Tao, Jianhua Sun, and Hao Chen. "Fine-grained mining and classification of malicious Web pages." In 2013 Fourth International Conference on Digital Manufacturing & Automation, pp. 616-619. IEEE, 2013.

[5] Yoo, Suyuan, 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.

[6] Kim, Sungjin, Jinkook Kim, Seok woo Nam, and Dohoon Kim. "WebMon: ML-and YARA-based malicious webpage detection." Computer Networks 137 (2018): 119-131.

[7] https://archieve.ics.uci.edu/ml/dataset/

[8] Website: https://www.jetbrains.com/pycharm/