**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# An Innovative Machine Learning based Approach for Detecting Spam Websites

***[1]D.Yaswanth Reddy, [2]D. S. Thousif Basha, [3]Yaswanth Reddy, [4]S. Alibasha***

[1,2,3,4] CSE Students, Santhiram Engineering College, Nandyal, Kurnool Dist. A. P. India

**ABSTRACT:**

Nowadays, it is normal practice to do all of our searches online, and this aspect of daily life is crucial. And since we frequently use browsers to search for items, browser producers are constantly attempting to offer new functionalities and capabilities, which creates a risk for fraudulent activity and puts us at risk if we visit those sites. Although there are tools available to help us determine if a website is malicious or benign, they do not offer protection from fraudulent activity. Therefore, we need a model that can determine whether a website is dangerous or not. We classify the websites using machine learning techniques in order to achieve this.

**Keywords:** malicious, website, online, risk, model.

## 1. Introduction:

As browsers evolve, they have access to more advanced features and functionalities, putting their personal and sensitive information at danger. Because inexperienced customers believe of the various types of malware, they are readily captured by intruders with just a single click on malicious web sites, allowing intruders to discover weaknesses on the page and inject payloads to gain remote access to the victim's web page. As a result, in an ever-growing web environment, precise identification of web pages is critical. To meet the issues, blacklisting services were included in browsers, although they had various drawbacks, such as erroneous listing. We look at a conscience strategy to classifying web pages based on a small collection of features in this post. To categorize the data, we employ four classification techniques. Research practitioners recommend three main ways to identify dangerous online pages: blacklisting, static analysis, and dynamic analysis. Each method has a certain goal in mind, and we've gone over a few of them in order. Using supervised machine learning algorithms, Tao et al. introduced a unique framework for automatically determining whether a web page is harmful or benign. Feature extraction techniques, the web pages were classified as malicious or not. As from sample, innocuous web sites were gathered. Adware et al. proposed a new lightweight self-learning technique to categorizing harmful web pages based on classified properties. A framework called MALURL (Malicious Uniform Resource Locator) was created that employed the Genetic Algorithm (GA) to train classifiers that can detect malicious web sites. For benign web sites, the dataset alexa was used, and for harmful web sites, the dataset Phis Tank was used. The average system precision was found to be 87 percent, and the machine learning technique used was adaptive Support Vector Machine(SVM).Because of its adaptable capacity, the SVM can deal with new training data.

## 2. Literature Review

Web page features (words) and information are extracted utilizing feature extraction algorithms from HTML (Hyper Text Markup Language) content: Keyword counts, keyword analysis strategies, and the existence of words. The performance of proposed machine learning models is assessed using a reference data set consisting of 100,000 WebPages. Experimental results demonstrated that the suggested method can detect fraudulent WebPages with a 98.24 percent accuracy, which is a major improvement over existing methods. Attackers take advantage of the Internet's openness to spread malware more easily. Their attempts to infect target systems via the Internet have grown in frequency over time and are unlikely to slow down. In approach to this issue, we propose Web on, an automatic, low-interaction malicious webpage detector that uses machine learning and YARA signatures to detect invasive roots in Web resources loaded from WebKit2-based browsers. The suggested model achieves a detection rate of 98 percent in this configuration and is 7.6 times faster (with a container) than previously proposed models. Most notably, Web Moon’s focus on extracting harmful pathways in a domain is a novel approach not before investigated. In recent years, malicious web sites have become a more serious danger to web security. In this research, we offer a new detection method for malicious web sites that combines static and dynamic assessments. Static analysis employs machine learning classification methods to distinguish between benign and malicious web pages. Dynamic analysis is used in conjunction with static analysis to evaluate whether unknown web pages include dangerous shell codes while they are being executed. The suggested detection approach provides good performance by combining static and dynamic analysis, and it is low weight and simple to use[3]. Access to websites with dangerous content is a risk on the Internet because they can serve as entry points for criminality or as a means for downloading things that can harm organizations, people, and the environment. Furthermore, website attack logs have been included in hack reports in recent years; this information covers attacks perpetrated by current hazards

discovered in new technologies, such as the Internet of Things. Due to the complexity of computer security, researchers have been experimenting with using machine learning algorithms to detect dangerous web content. In order to classify a website, this paper examines the implementation of a data analysis method using a framework that incorporates dynamic, static analysis, updated websites, and a low interaction client honeypot. It also assesses the categorization power of four machine learning algorithms using the data analyzed. For the detection of harmful web sites, the misuse detection method and anomaly detection approach are extensively employed. Machine learning is used in both cases. Vulnerability assessment can detect dangerous web pages that have already been identified, but it cannot detect new ones.

## 3. Proposed System

Textual qualities, link topologies, webpage information, DNS (Domain Name System) metadata, and network traffic are among the discriminative features used by our method. Several of these characteristics are unique and quite useful. Our experiments with 40,000 benign URLs and 32,000 harmful URLs taken from real-world Internet sources reveal that our technology outperforms the competition: the reliability in identifying fraudulent URLs was over 98 percent, and the accuracy in identifying attack types was over 93 percent.

The Support Vector Machine (SVM) technique is a simple and effective Supervised Machine Learning algorithm that may be used to create both predictive modelling models. Both linearly separable and non-linearly separable datasets can benefit from the SVM method. Even with a small amount of data, the support vector machine method performs admirably.

Step 1: Using Pandas, load the Pandas library and the dataset.

Step 2: Define the characteristics and the goal

Step 3: Before developing the SVM algorithm model, split the dataset into train and test using sklearn.

Step 4: From the Sklearn SVM module, import the support vector classifier (SVC) function. Using the SVC function, create a Support Vector Machine model.

Step 5: Use the SVM algorithm model to predict values.

Step 6: Assess the Support Vector Machine (SVM) model

## 4. Conclusion

The detection of malicious web pages is a new topic in cyber security. Several research studies linked to the identification of malicious web pages have been conducted, but they are quite expensive because they need more time and resources. In this study, we used machine learning techniques to predict whether online sites were harmful or benign using a novel web site classification system based on URL attributes. Random Forest(RF), a machine learning classifier, achieves a greater accuracy of 95%.

**References**

1. Tao, Wang, Yu Shunzheng, and Xie Bailin.” A novel framework for learning to detect malicious web sites”, International Forum on Information Technology and Applications, vol.2, pp.353-357, 2020.
2. Eshete, Birhanu, Adolfo, “Effectiveness and efficiency difficulties in malicious website identification, IEEE conference on security, pp. 123-126.2011.
3. Rami Alsalman, Aldwairi, and Monther, "Malurls is a simple way to classify dangerous websites based on their url properties”, Journal of Emerging Technologies in Web Intelligence, vol. 4, no. 2, pp.128-133, 2012. .