IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Malicious URL Detection Using Machine Learning

Koteswara Rao Velpula[1], Assistant Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

Kataru Gayathri Priya[2], Kushwanth Kumar Jammula[3], Krishna Sruthi Velaga[4], Praveen Kumar Kongara[5]
[2,3,4,5] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
koteswararao@vvit.net[1], gpkataru2001@gmail.com[2], jkk07july@gmail.com[3],
ksvelaga@gmail.com[4], praveenkongara123@gmail.com[5]

DOI:10.48047/IJFANS/V11/I12/218

Abstract

Throughout the years, internet usage has grown significantly. The internet continues to transform how we interact with people, organise the flow of goods, and communicate knowledge all across the world. The attackers have used this popularity to their advantage to participate in illegal activities that would lead to monetary advantage. There has been a rise in malicious websites that launch client-side attacks over time, which cannot be identified effectively by existing approaches such as blacklisting. As a result, an efficient solution to detect these malicious websites is required. In this study, we used the random forest method to develop a machine learning model while integrating lexical features, host-based features, and content-based features. The model has an accuracy of 94.7%.

Keywords: Internet Usage, Malicious Websites, Client-Side Attacks, Machine Learning, Random Forest.

Introduction

It is now impossible to imagine a world without the internet. The internet is a key component of today's information society, connecting billions of people worldwide. There were 5.16 billion internet users worldwide as of January 2023, accounting for 64.4% of the global population [1]. This widespread use of the internet has turned into one of the primary channels for malware distribution by attackers. Phishing attacks targeted 323,972 internet users worldwide in 2021. This suggests that half of all victims of cybercrime were deceived by a phishing

attack. This is despite Google's cyber security measures blocking 99.9% of phishing attempts from reaching users [2]. The use of malicious URLs has become increasingly prevalent as a means to carry out criminal activities on the internet, including drive-by-downloads, spamming, and phishing. These websites are frequently targeted by attackers who try to steal identities or spread dangerous software [3]. While security components used today attempt to identify and block such malicious sites and web addresses, attackers are constantly evolving their techniques and finding ways to evade detection. One of the most popular techniques is the blacklist technique, which filters incoming URLs using a database of known harmful URLs. Nevertheless, blacklists have limitations, and this

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

strategy is ineffective for new malicious sites that are constantly being created [4]. According to the Website Report by SiteLock, search engines often miss widespread malware infections, leaving owners and visitors vulnerable to attacks. Almost 92% of infected websites are not flagged or blacklisted by the common search engines [6]. Therefore, it can be inferred that there is a need for an effective solution to identify malicious websites.

Literature Survey

This section provides a summary of relevant work in the identification of malicious URLs. Ferhat et al.[4] developed a machine learning model that detects malicious URLs using supervised learning techniques. They employed random forest and gradient boosting techniques, as well as utilised lexical and host-based characteristics. According to their results, the model attained an accuracy of 98.6% for Random Forest and 96.5% for Gradient Boosting. Sandra et al.[7] detected malicious URLs using a data mining technique known as Classification Based on Association (CBA). They obtained a 95.8% accuracy by utilising both URL and web page content elements. They discovered that CBA was equally effective at identifying malicious URLs as other benchmark classification methods. Christian et al.[8] combined a classification method and a high-interaction client honeypot to develop a hybrid system for detecting malicious URLs. URLs were first classified using the classification approach, and then they were forwarded to the honeypot for final classification. They obtained a false positive rate of 5.88% and a false negative rate of 46.15% in their studies on a sample of 61,000 URLs. Dharmaraj et al.[9] surveyed several approaches and features used to detect malicious web pages, highlighting fellow researchers' ongoing efforts to improve detection. They also addressed several forms of attacks, including malware and injection attacks, as well as social engineering attempts. Machine learning methods such as SVM, Random Forest, and CNN were utilised by Deepa et al. [10] to detect malicious URLs. Word vectors that are challenging for attackers to fabricate were taken into account, along with lexical and host-based properties. CNN outperformed SVM and Random Forest with an accuracy of approximately 70%, as per their experimental results. By taking into account lexical features, host-based features, and popularity features, Naresh et al.[11] used machine learning techniques like SVM and Logistic Regression to identify harmful URLs. A 98% accuracy rate was attained by their model. [15-23]

Problem Identification

The World Wide Web has become an essential aspect of many people's lives, with millions using online services like online banking, shopping, and social networking. However, the internet has also become a more dangerous place due to the rise of illegitimate activities aimed at financial gain. Traditional methods like blacklisting are effective against known

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

malicious URLs, but they cannot detect unknown sites. Therefore, there is a need for new automatic detection methods that employ machine learning approaches. This paper proposes a solution that considers content-based, lexical, and host-based features and employs a supervised learning technique called a random forest model to detect attacks such as phishing and drive-by downloads, which have become more common in recent years.

Methodology

Malicious URLs are a significant risk to internet users because they can be used to deliver malware, steal sensitive information, or launch phishing attacks. Because attackers are constantly creating new, unknown sites, traditional blacklisting methods for detecting malicious URLs are no longer effective. As a result, more sophisticated, automated techniques that can detect these threats in real time are required. The methodology section will describe our approach and demonstrate its effectiveness in real-time malicious URL detection.

A. Dataset

We used the publicly available Phishing Websites Dataset from UCI Machine Learning Repository for this paper, which contains more than 11,000 phishing websites [12]. The dataset was collected between 2013 and 2016 and contains a variety of phishing websites that target several sectors, including banking, e-commerce, and social media. The dataset has 30 features altogether, which are a combination of static and dynamic features. Static features are website characteristics that can be determined without visiting the website, such as domain age and URL length. Dynamic features include things like the number of forms and external links that can only be discovered by visiting the website. The dataset is in CSV format [13] and contains several features that can be utilised to train machine learning models for phishing detection. The labels "-1" and "+1" are used to identify each website in the dataset. A website with a "-1" label is a phishing website, while a website with a "+1" label is not a phishing website.

B. Feature Extraction

The features extracted from a URL are used to determine whether or not the URL is malicious. The features we considered are categorised as follows: lexical, host-based, and content-based features. Lexical features from the URL string are used in the approach to detect attacks, based on the distinguishable visual characteristics between malicious and benign URLs, quantified through statistical analysis [14]. Host-based characteristics are derived from the webpage's hostname and provide information about the website's hosting details, such as location, time active, and hosting organisation. These features assist us in determining "who," "where," "when," and "how" a website is hosted [14]. Content-based

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

features are statistics collected from the raw HTML and JavaScript code of a webpage, such as the number of tokens, scripts, characters in scripts, and the percentage of white spaces. These elements aid in the detection of malicious activity and provide insight into the content of the webpage [14].

Lexical Features: url_of_anchor, sub_domain, having_-, links_in_tags, sfh, request_url, url_length, https_token, shortening_service, having_@, abnormal_url, having_//.

Host-based Features: registration_length, age_of_domain, having_ip, google_index, dns_record.

Content-based Features: web_traffic, favicon, redirect, submitting_to_email, statistical_report, mouse_over, iframe, rightclick

C. Random Forest Classifier

Random forest is an ensemble learning method that combines the outputs of several decision trees to make a final prediction. Each decision tree is built with a random subset of features and a random subset of training data, which helps to minimise overfitting and improve model robustness. The random forest approach builds a classification model using a dataset of URLs labelled as malicious or benign to detect malicious URLs. This model is trained to employ many URL features, including lexical, host-based, and content-based features. Once trained, the model can be used to identify new, unknown URLs as malicious or benign. When a new URL is provided to the random forest model, each decision tree produces a prediction based on the URL's features. The individual predictions made by the decision trees are aggregated to make the final prediction. This aggregation process helps to improve the accuracy of the model's predictions and reduce overfitting.

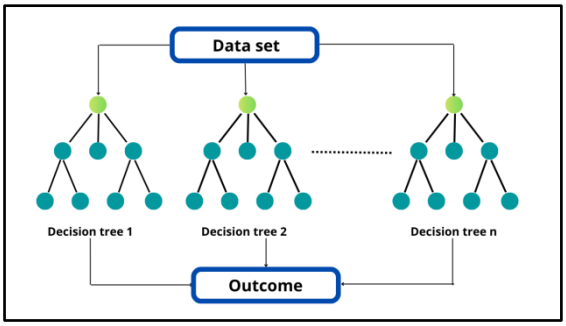

Fig. 1. Working of Random Forest Classifier

By combining the predictions of multiple decision trees, the random forest algorithm can make more reliable predictions for both known and unknown malicious URLs. The use of a random forest model for malicious URL detection has the advantage of reducing false positives. False positives arise when benign URLs are incorrectly identified as malicious, resulting in unnecessary blocks or alerts. The Random forest can significantly reduce false positives and enhance overall accuracy by offering a robust and accurate classification

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

model. Overall, employing a random forest model for malicious URL detection is an efficient method of detecting previously unknown malicious URLs and preventing them from causing damage.

Implementation

We used the Random Forest algorithm which is a popular machine learning algorithm used for classification tasks. The Random Forest algorithm works by training a classification model on a dataset of URLs that have been labelled as either malicious or benign. We considered a total of 25 features for training our model which included lexical, host-based, and content-based features. These features were extracted from the URLs using various techniques such as regular expressions, domain analysis, and webpage content analysis. One challenge we faced during the training of our model was the class imbalance issue, where the number of malicious URLs was much lower than the benign URLs. To address this issue, we used the Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority class (malicious URLs) in our dataset. SMOTE works by creating synthetic samples of the minority class based on its existing samples, thus creating a more balanced dataset. After preprocessing our data and oversampling the minority class using SMOTE, we split our dataset into training and testing sets with an 80-20 ratio. We trained our Random Forest model on the training set and evaluated its performance on the testing set. The model achieved a training accuracy of 97.2% and an overall accuracy of 94.7%. We also evaluated the model using precision, recall, and F1-score metrics and obtained good results. The confusion matrix showed that the model was able to correctly classify most of the URLs, with a few false positives and false negatives. Overall, our implementation of the Random Forest algorithm with SMOTE oversampling and 25 features proved to be effective in detecting malicious URLs with a high level of accuracy.

Results & Conclusion

The model achieved a training accuracy of 97.2% and an overall accuracy of 94.7%. The confusion matrix shows that out of 2,171 samples, 1,094 were correctly classified as benign and 986 were correctly classified as malicious. The precision and recall for class -1 (benign) were 0.95 and 0.93 respectively, with an f1-score of 0.94. Similarly, the precision and recall for class 1 (malicious) were 0.95 and 0.96 respectively, with an f1-score of 0.95.

The confusion matrix offers a more detailed view, showing not only how well a predictive model performs but also which classes are properly and mistakenly predicted as well as the kind of errors that are being produced.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

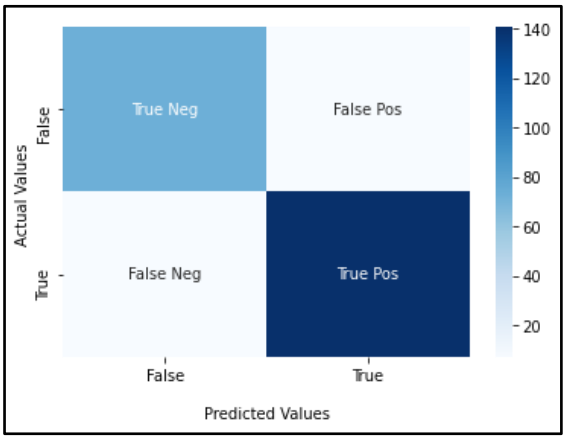

Fig. 2. Confusion Matrix

Precision is a classifier's ability to avoid labelling a negative occurrence as positive. It is defined for each class as the ratio of true positives to the sum of true positives and false positives.

$$Precision = \frac{TP}{TP + FP}$$

Recall is a classifier's capacity to find all positive occurrences. For each class, it is defined as the ratio of true positives to the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP + FN}$$

The F1 score is a weighted harmonic mean of accuracy and recall, with 1.0 being the highest and 0.0 being the lowest. Because F1 scores incorporate precision and recall into their computation, they are lower than accuracy measurements.

$$F1\,Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Accuracy is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0.

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

	Precision	Recall	F1-Score	Support
-1	0.95	0.93	0.94	1001
1	0.95	0.96	0.95	1210
Accuracy			0.95	2211
Macro Avg	0.95	0.95	0.95	2211
Weighted Avg	0.95	0.95	0.95	2211

Fig. 3. Classification Report

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022**

In conclusion, malicious URLs pose a significant threat to internet security, and detecting them accurately and efficiently is crucial. In this study, we developed a classification model based on the random forest algorithm to detect malicious URLs. We considered 25 features, including lexical, host-based, and content-based features, and used oversampling techniques with SMOTE to address the class imbalance issue. Our primary focus in developing the malicious URL detection model was on maximising its effectiveness in detecting attacks, rather than on speed. This was deemed necessary to ensure that the model can accurately identify a wide range of malicious URLs. Therefore, while the current implementation may take up to 30 seconds to provide an output, we believe that this balance is necessary to ensure maximum detection capability.

Future Works

Although the developed model has shown impressive results in detecting malicious URLs, there is still a need for improvement. One area for future work is to address the issue of the model's slow prediction time. Currently, the model takes a minimum of 30 seconds to provide output, which is not ideal for real-time detection. Another area for improvement is to explore ways to handle situations where the page is down or permission to access them is denied, as the current model cannot predict whether a website is malicious under such circumstances. This can be achieved by incorporating additional features or exploring new techniques such as deep learning to improve the model's robustness and accuracy.

References

[1] [1] Ani Petrosyan, “Internet And Social Media Users in the World 2023”. Available online: https://www.statista.com/statistics/617136/digital-population-worldwide/ (accessed on 1 March 2023).

[2] [2] “The Latest Phishing Statistics (updated February 2023): AAG IT Support”. Available online: https://aag-it.com/the-latest-phishing-statistics/ (accessed on 1 March 2023).

[3] [3] Min-Sheng, Chien-Yi, Yuh-Jye, Hsing-Kuo, “Malicious URL Filtering — A Big Data Application” in 2013 IEEE International Conference on Big Data. doi: 10.1109/bigdata.2013.6691627 (accessed on 1 March 2023).

[4] [4] Ashish Kumar Luhach, Atilla Elçi, “Artificial Intelligence Paradigms For Smart Cyber-Physical Systems”. Available online:https://www.researchgate.net/publication/345762406_Artificial_Intelligence_Paradigms_for_Smart_Cyber-Physical_Systems.

[5] [5] Sectigo, “Cybersecurity Statistics Report 2022”. Available online: https://www.sitelock.com/resources/security-report/ (accessed on 1 March 2023)

[6] [6] “Cost of a Data Breach 2022”. Available online: https://www.ibm.com/reports/data-breach (accessed on 1 March 2023).

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

[7] [7] Sandra Kumi, ChaeHo Lim, Sang-Gon Lee, "Malicious URL Detection Based on Associative Classification". doi: 10.3390/e23020182.

[8] [8] Christian Seifert, Ian Welch, Peter Komisarczuk, "Identification Of Malicious Web Pages With Static Heuristics". doi: 10.1109/atnac.2008.4783302.

[9] [9] Dharmaraj Rajaram Patil and J. B. Patil, "Survey on Malicious Web Pages Detection Techniques". doi: 10.14257/ijunesst.2015.8.5.18.

[10] [10] Deepa Mary Vargheese, Sreelakshmi N R, "Phishing Website Detection Using Machine Learning Techniques And CNN". Available online: https://www.ijert.org/phishing-website-detection-using-machine-learning-techniques-and-cnn (accessed on 3 March 2023).

[11] [11] R. Naresh, Ayon Gupta, Sanghamitra Giri, "Malicious URL Detection System using Combined SVM And Logistic Regression Model". Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3598024 (accessed on 3 March 2023).

[12] [12] UCI Machine Learning Repository: Phishing Websites Dataset. Available online:https://archive.ics.uci.edu/ml/datasets/phishing+websites

[13] [13] Akash Kumar, "Phishing Website Dataset". Available online: https://www.kaggle.com/datasets/akashkr/phishing-website-dataset.

[14] [14] Ruth Eneyi Ikwu, "Extracting Feature Vectors From URL Strings For Malicious URL Detection". Available online: https://towardsdatascience.com/extracting-feature-vectors-from-url-strings-for-malicious-url-detection-cbafc24737a (accessed on 3 March 2023).

[15] Sri Hari Nallamala, et al., "A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 729 – 732, March 2018.

[16] Sri Hari Nallamala, et.al., "An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 542 – 545, March 2018.

[17] Sri Hari Nallamala, et.al, "Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems", International Journal of Advanced Trends in Computer Science and Engineering, (IJATCSE), ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2, Page No: 259 – 264, March / April 2019.

[18] Sri Hari Nallamala, et.al, "Breast Cancer Detection using Machine Learning Way", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S3, Page No: 1402 – 1405, July 2019.

[19] Sri Hari Nallamala, et.al, "Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in the Breast Cancer Treatment", International Journal of Scientific and

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Technology Research, (IJSTR), ISSN: 2277-8616, Vol. 8, Issue 11, Page No: 2168 – 2173, November 2019.

[20] Kolla Bhanu Prakash, Sri Hari Nallamala, et al., "Accurate Hand Gesture Recognition using CNN and RNN Approaches" International Journal of Advanced Trends in Computer Science and Engineering, 9(3), May – June 2020, 3216 – 3222.

[21] Sri Hari Nallamala, et al., "A Review on 'Applications, Early Successes & Challenges of Big Data in Modern Healthcare Management'", Vol.83, May - June 2020 ISSN: 0193-4120 Page No. 11117 – 11121.

[22] Nallamala, S.H., et al., "A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems", IOP Conference Series: Materials Science and Engineering, 2020, 981(2), 022008.

[23] Nallamala, S.H., Mishra, P., Koneru, S.V., "Breast cancer detection using machine learning approaches", International Journal of Recent Technology and Engineering, 2019, 7(5), pp. 478–481.