Report Title: **MALWARE DETECTION OF URLs**

Ali Zarafshani, CSCE, University of North Texas

Sravya Alladi,  
*Computer Science*   
*University of North Texas*Denton, Texas, US  
[sravyaalladi@my.unt.edu](mailto:sravyaalladi@my.unt.edu)

Dikshitha Mamidi,  
*Computer Science*   
*University of North Texas*Denton, Texas, US  
[dikshithamamidi@my.unt.edu](mailto:dikshithamamidi@my.unt.edu)

Sravan Sai Palvaye,  
*Computer Science*   
*University of North Texas*Denton, Texas, US  
palvayesai[sravan@my.unt.edu](mailto:sravan@my.unt.edu)

Rakesh Kommineni,  
*Computer Science*   
*University of North Texas*Denton, Texas, US  
[rakeshkommineni@my.unt.edu](mailto:rakeshkommineni@my.unt.edu)

Harish Kumar Manthoju,  
*Computer Science*   
*University of North Texas*Denton, Texas, US  
[harishkumarmanthoju@my.unt.edu](mailto:harishkumarmanthoju@my.unt.edu)

***Abstract*— Today's smartphones are changing every day, and as a result, security is a major concern. In a world with insufficient protection, security—an essential component of human existence— becomes problematic for smartphone users' safety. Malware is one of the biggest security risks to cellphones. The study conducted a review on malware detection methods to find gaps and create a foundation for developing and effective defense against unidentified Android malware. According to the findings, machine learning offers a more promising method with improved detection accuracy. Future studies should investigate deep learning approaches using a huge dataset in order to improve accuracy.**

Keywords—lexical features, random forest, malicious, malware

# introduction

A case study of identifying dangerous URLs using lexical characteristics and a boosted tree-based machine learning strategy will be covered in this report. Use the enter key to start a new paragraph. The appropriate spacing and indent are automatically applied. Recent years have seen a sharp rise in cybersecurity attacks on many websites around the world, including ransomware, phishing, malware injection, etc. As a result, numerous financial institutions, e-commerce businesses, and people suffered significant financial losses. Since new attack types are being developed daily, controlling a cyber security attack in such a situation is a significant problem for cyber security specialists.

# Malicious url

## WHAT IS URL?

### The Uniform Resource Locator (URL) is a structured, well-defined, and specific address used to visit websites on the World Wide Web (WWW). A legal URL typically consists of these three elements.

### Protocol: The protocol, such as HTTP, HTTPS, etc., is primarily determined by an identity.

### Hostname is another name for the resource name. It includes the domain name or IP address of the location of the actual resource.

### Path: This describes the precise route that leads to the resource.

### Diagram, timeline Description automatically generated

Diagram, timeline

Description automatically generated

1. Components of URL

Malicious URLs are modified, or compromised URLs used in cyber assaults. A malicious URL or website typically includes various trojans, malware, and unsolicited material in the form of phishing, drive-by downloads, and spam. The harmful website's primary goal is to defraud users or steal their personal or financial information. The current COVID-19 pandemic has greatly increased the number of cybercrime instances. Malicious URLs are a frequently employed tactic in cybercrimes, according to the 2019 Symantec Internet Security Threat Report (ISTR).

According to fig.1., another element of the domain name is the top-level domain, which indicates the type of website, such as commercial (.com), educational (.edu), organization (.edu), etc.

In this case study, we discuss the multi-class classification problem of dangerous URL detection. In this case study, we categorize the raw URLs into various class categories, including safe or benign URLs, phishing URLs, malware URLs, and defacement URLs.

# Project design

#### Due to the fact that lexical numeric features must be created from input URLs since machine learning algorithms only accept numeric inputs, Therefore, rather than actual raw URLs, the input to machine learning algorithms will be the numeric lexical characteristics. Therefore, we will be utilizing the three well-known machine learning ensemble classifiers Random Forest, Light GBM, and XGBoost in this case study. Later, in order to determine which features are crucial for predicting dangerous URLs, we will also compare their performances and create an average feature importance plot.

Diagram

Description automatically generated

1. Overflow

## Different types of URLs

Benign URLs: These are safe to browse URLs. Some of the examples of benign URLs are as follows:  
mp3raid.com/music/krizz\_kaliko.html, infinitysw.com, google.co.in

Malware URLs: These types of URLs inject malware into the victim’s system once he/she visits such URLs. Some of the examples of malware URLs are as follows:  
proplast.co.nz,<http://103.112.226.142:36308/Mozi.m>, microencapsulation.readmyweather.com, xo3fhvm5lcvzy92q.download

Defacement URLs: Defacement URLs are generally created by hackers with the intention of breaking into a web server and replacing the hosted website with one of their own.   
Ex:<http://www.vnic.co/khach-hang.html>, <http://www.raci.it/component/user/reset.html>

Phishing URLs: By creating phishing URLs, hackers try to steal sensitive personal or financial information such as login credentials, credit card numbers, internet banking details, etc. Some of the examples of phishing URLs are shown roverslands.net, corporacionrossenditotours.com

## Lexical Features

Since lexical numeric features must be created from input URLs since machine learning algorithms only accept numeric inputs, Therefore, rather than actual raw URLs, the input to machine learning algorithms will be the numeric lexical characteristics. Therefore, we will be utilizing the three well-known machine learning ensemble classifiers Random Forest, Light GBM, and XGBoost in this case study. The following lexical features will be extracted from raw URLs in this step and utilized as input features to train the machine learning model. The ensuing features are produced in the following ways:

having ip address: To disguise the identity of a website, online criminals frequently use an IP address instead of the domain name. This function will determine if the URL contains an IP address or not.

abnormal url: The WHOIS database can be used to extract this feature. For a genuine website, identity is often part of its URL.

Count.: The URLs of phishing or malware websites frequently contain more than two sub-domains. A dot separates each domain (.). Any URL with more than three dot characters (.) raises the risk of a malicious website.

count@: If the URL contains the "@" sign, everything before it is ignored.

Count%: As we all know, spaces are not permitted in URLs. Normal URL encoding substitutes the symbol (%) for spaces. Safe websites typically have less space in their URLs than dangerous websites, which means that they have more spaces overall.

Count digits: Suspicious URLs are typically those that contain numbers. Counting the number of digits in a URL is a key feature for identifying malicious URLs because safe URLs typically do not have digits.

Count letters: Another important factor in recognizing fraudulent URLs is the number of letters in the URL. Attackers typically do this by adding more letters and numbers to the URL to lengthen it and conceal the domain name.

So, now after creating the above 22 features, the dataset looks like the below fig 2.

Table

Description automatically generated with medium confidence

Fig. 2. Components of URL

Next, we eliminate the useless columns, namely URL, Google index, and tld. We have already derived pertinent properties from the URL column that can be utilized as input in machine learning algorithms, so we are deleting it for this reason.

Since we have built a tld column to measure the length of the top-level domain, the tld column has been removed because it is an indirect textual column.

# Training and testing split

The next step is to split the dataset into train and test sets. We have split the dataset into 80:20 ratio i.e., 80% of the data was used to train the machine learning models, and the rest 20% was used to test the model.

As is well known, our dataset is unbalanced. The data contains benign URLs in about 66% of cases, malware in 5%, phishing in 14%, and defacement URLs in 15% of cases. So it's possible that the distribution of distinct categories changed after the dataset was randomly split into train and test, which would have a significant impact on the machine learning model's performance. Therefore, it is necessary to retain the same percentage of the desired variable stratification.

The split created by this stratify parameter ensures that the proportion of values in the sample produced and the proportion of values supplied to the stratify parameter are the same.

# model building

In this step, we will build three tree-based ensemble machine learning models i.e., Light GBM, XGBoost, and Random Forest. A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and utilizes averaging to increase the predicted accuracy and control over-fitting. If bootstrap=True (the default), the size of the sub-sample is determined by the max samples’ argument; otherwise, each tree is constructed using the entire dataset.

# model evaluation and comparision

After fitting the model, as shown below, we have made predictions on the test set. The performances of Light GBM, XGBoost, and Random Forest are shown below. As can be seen from the results above, Random Forest performs best in terms of test accuracy, achieving the maximum accuracy of 96.6% and having a higher detection rate for malware, phishing, and benign content. We have chosen Random Forest as our primary model for identifying malicious URLs based on the performance described above, and in the following phase, we will also plot the feature importance plot.

A screenshot of a computer

Description automatically generated with low confidence

A picture containing text, receipt, screenshot

Description automatically generated

A picture containing text, receipt, screenshot

Description automatically generated

##### conclusion

In this report, we have shown how to use machine learning to find malicious URLs. From raw URLs, we extracted 22 lexical characteristics and trained the XG Boost, Light GBM, and Random Forest machine learning models. In addition, we evaluated the effectiveness of the three machine learning models and discovered that Random Forest performed better than the others, achieving the greatest accuracy of 96.6%. We discovered that the top 5 features for identifying malicious URLs are hostname length, count dir, count-www, fd length, and url length by charting the feature importance of Random Forest. The prediction method for categorizing any raw URL using our saved model, Random Forest, has finally been coded.

##### References

1. D. R. Patil and J. B. Patil, “Survey on malicious web pages detection techniques,” International Journal of U- and E-Service, Science and Technology, vol. 8, no. 5, pp. 195–206, 2015
2. S. Garera, N. Provos, and M. Chew, “A framework for detection and measurement of phishing attacks,” in Proceedings of the ACM Workshop on Recurring Malcode, ACM, New York, NY, USA, November 2007