**Malicious URL Detection**

**Interim Progress Report**

**Student Number:**

**Student Name:**

**Course:**

**Supervised by:**

Contents

[Introduction 2](#_Toc100422882)

[Feature Selection and Extraction: 2](#_Toc100422883)

[Data Set Details 3](#_Toc100422884)

[Data Description 3](#_Toc100422885)

[Literature Review 5](#_Toc100422886)

[Significance of the Research 6](#_Toc100422887)

[Aim of project 7](#_Toc100422888)

[Objectives of project 7](#_Toc100422889)

[Research Questions 8](#_Toc100422890)

[Social Issues, Ethical Issues, and Legal Issues 8](#_Toc100422891)

[Technical Work 8](#_Toc100422892)

[Progress to Date 9](#_Toc100422893)

[Planned Work 9](#_Toc100422894)

[Appendices 9](#_Toc100422895)

[References 13](#_Toc100422896)

# Introduction

Since any files on the computer will be available as a file name, in addition to tracking the website, resource tenants or uniform URLs are used. You can restore a domain name by entering the URL into the browser address bar or clicking the appropriate URL to access the popular website. Increasing Internet usage for these purposes will increase the level of Cyber-fraud business. When connecting customers and base customers to develop, there is equal growth of attackers. Government, companies, and individuals are the poorest. Future threats and their nature are difficult to predict and nearly impossible to deal with. Malware or malicious websites emerged as one of the most important online security threats. Although malicious URLs, in particular, have become an online security threat. Harmful URLs are a common and serious threat to Internet security. Malicious URLs contain malicious content including spam, phishing scams, user abuse, and more (Kotoju, and Vijaya Lakshmi, 2021).

Attackers send attack vectors often in the form of e-mails, consultations, web posts, etc., containing a link (URL) to a hosted malicious website to obtain patient statistics public. We are interested in creating a URL analyzer and importing it to detect attacks by cybercriminals. URL parsing is better to maintain distance between attacker and patient, than going to a website and finding work there. Malicious websites include many types of illegal businesses that can be dangerous to visit, that's why different types of malicious websites pose different threats to customers. Once this type of risk is known, it is easier to test these types of risks independently and keep their functions so that it can be profitable to detect a malicious website online. and find solutions against a particular type of risk. Experimental operations show that the learning algorithms are very useful in detecting harmful URLs. For example, the Naive Bayes category assumes that the presence or absence of a selected class element is not related to the presence or absence of another function, which greatly improves the situation encountered. Malicious URLs (Sayamber, and Dixit, 2014).

# Feature Selection and Extraction:

Skill selection is an important and powerful segment in which a useful database can be very large. This makes finding patterns and finding relationships between elements very difficult to compute. In Machine Education, the characteristic of a measurable property or function, or the characteristic of a particular object. Choosing independent, distinct, and informative tasks is an important step for efficient algorithms. Flexible selection and feature selection are marked with function outputs. Contains a selection of skills that are not required for teaching gender translation. This difficulty requires that the URL be classified as invalid or malicious. So, to solve the problem, we will design a model and educate it about using the features extracted from the transcript. The follow-up after data collection provides useful and informative power for interpreting URLs and can be mathematically interpreted to teach how to use machine learning (Patgiri, et al., 2019). Simply put, using a URL will not immediately allow best-class progress. Therefore, it is important to choose the right skills purely based on a few principles or imaginative ideas to get the best functionality from a set of URLs. Thus, the fines for actions extracted from the URL are broken down into the early stages of the results of the malicious URL category model. URL parsing functions can fall into several categories that can be classified by lexical capabilities, purely host-based capabilities, and web content capabilities. During our research, we found that one way to visualize a malicious website is to analyze its software layer and public domain features, to find them, the idea is to use flexible testing. active and static.

# Data Set Details

This is a vital topic as well as one of the most difficult issues to follow, along with other research studies and other open help, we have used 3 blacklists:

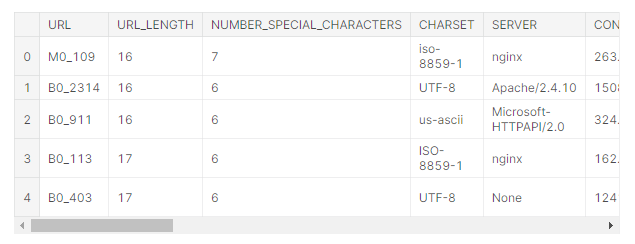
* machinelearning.inginf.units.it/data-andtools/hidden-fraudulent-urls-dataset
* zeuztacker.abuse.ch
* malwaredomainlist.com

From them we found approximately 185181 URLs, which we thought were malevolent according to their statistics, recommending within the next research step to verify them or all other security resources, as well as Virus Total.

We got the benign URLs (345000) from the following link given below: https://github.com/urcuqui/WhiteHat/tree/master/Research/Web%20security.

# Data Description

* URL: Anonymous text URLs analyzed.
* URL\_LENGTH: Font number within a URL.
* NUMBERSPECIALCHARACTERS: Other different fonts or characters identified within the URL, including, "/", "%", "#", "&", ". "," = ".
* CHARSET: Its defined characteristic set.
* SERVER: A straightforward value and that means the active server gadget is given a package response.
* CONTENT\_LENGTH: The content size of the URL.
* WHOIS\_COUNTRY: The values ​​of which countries we have obtained from server response to the Whois API implementation.
* WHOIS\_STATEPRO: Category variation, whose morals ​​are the conditions we received from server reply to the Whois API usage.
* WHOIS\_REGDATE: Registration date of the server in date format.
* WHOISUPDATEDDATE: With Whois, we have received the last replacement date from the revised server.
* TCPCONVERSATIONEXCHANGE: This varies the number of TCP packs traded among client and server.
* DISTREMOTETCP\_PORT: far from the wide range of gaps found and specialized from TCP.
* REMOTE\_IPS: The most diverse of honeypot-related IPs.
* APP\_BYTES: The number of bytes passed.
* SOURCEAPPPACKETS: Sachets sent from honeypot to server.
* REMOTEAPPPACKETS: Sachets found on the server.
* APP\_PACKETS: The entire range of IP packages generated periodically in the exchange of words between client and server.
* DNSQUERYTIMES: The number of DNS packages produced during the conversation between the server and the honeypot.
* TYPE: A clear variable, its values ​​include the type of content page analyzed, in particular, ‘1’- for malicious website URLs as well as, ‘0’- for non-malicious website URLs.



# Literature Review

(Yuan, et al., 2021) proposed an incorporated set of neural policies for the analysis and discovery of the URL. By finding and getting to know malicious URL capabilities, semantic as well as, visual information is extracted. First, a set of preview policies is used to come across what the URL mapping should appear to be in a grayscale picture with texture elements. Second, the lexical as well as, practical element of the URL individual is extracted as well as, moreover processed by way of vector technology. These output possibilities are converted into lexical embedding vectors as well as, character embedding vectors. To integrate the sensory and textual features, a consistent neural network consisting of a tablet computer network (CapsNet) and recurrent neural network (RNN) was used to acquire vectors for a text. series of corresponding semantic and visual records. The rest of the class uses an optical approach to further clarify the in-depth features taken from the conventional network and even focuses on effective operations to improve class accuracy, learning, and looking for malicious URLs. Based on the test results, proved that the neural-network-based algorithm has the best accuracy compared to the traditional algorithms (Yuan, et al., 2021).

(Sayamber and Dixit, 2014) recommends the Naïve Bayes category for default and detection of harmful URLs. The proposed model is entirely based on Naive Bayes backed by mergers and acquisitions strategies. Otherwise, they cannot be used for the study of possible fashions and the definition is often used to measure under conditional and lateral distributions. The designs provided on paper show that, given a series of reference data sets, Naive Bayes models were obtained using a random version with better accuracy than the Machine learning model of SVM.

(Patiri, et al., 2019) used to teach a database of the most beautiful and most impressive URLs. The database's URL is divided into forming and watching facts at 60:40, 70:30, as well as, 80:20. The accuracy of random forests and SVM is calculated in many instances on the type ladder. As a result, the rate of 80: 20 is found to be the most accurate department and the average accuracy of random woodland is more than SVM. SVM got better performance over random forests.

(Mentumumwa, et al., 2020) compared the performance of the following new participants: Gradient Extreme XgBoost, AdaBoost, Lightgbm, and Catboost. These include introductory features like the KullbackLeibler Divergence (KL Divergence), a bag of phrase examples, and various phrase-based functions. The results showed that their characteristics were significantly increased compared to the trials (Manyumwa, et al., 2020) performed beyond their capabilities. These algorithms can be used at 126,983 URLs from the benchmark database and four beginners with typical accuracy above 95%. They target three common URL attacks that can be spam, malware, and phishing, and can be used as a supplementary tool in anti-spam detection platforms, against new or existing spam and cybercrime.

(Li, et al., 2020) used 33,122 URLs with 62 objects collected to authenticate projected development techniques. The consequences showed that the proposed techniques significantly improved the general efficiency as well as, the performance of excellent separators, as well as aid, close-sensory, and vector networks. Dangerous neighbor URL detection rate increased from sixty-eight percent to 86%, Support Linker range price increased from 58% to 81%, and Perceptron multilayer rate increased from 63-82%.

# Significance of the Research

The challenge is to discover layer modes that rarely predict malicious sites, based on resource levels as well as, network characteristics. Facts were found on how to use guaranteed content of malicious and unsafe URL types, in a collaborative client honeypot to separate website visitors from the network. Although the security features used today are aimed at finding malicious websites and online addresses, these additions are still avoided by using various tactics that attackers use (Bhagwat, et al., 2019). Researchers have learned how to offset the hassle of a malicious URL. One of the most popular techniques is the blocking off approach which uses seen malicious URL records to filter incoming URLs. However, there are a few downsides to listing blockchains, and this approach is futile for brand new malicious websites which might be constantly created. Security improvements have all started to use continuous device packages for AI-based cognitive and inferential fashions to cope with this, during the last few a long time. We pick smart prediction of device intelligence and overall performance as opposed to assisting signature fetching URLs. Machine getting to know methods tend to use fixed URLs as school statistics and learn the heuristics to distinguish whether or not a URL is malicious or innocent. This method permits them to bring together new URLs, in contrast to narrow search engine optimization techniques. Soon, these solutions will be implemented to network bodily structures (CPS), and another place will be the search for malicious websites and URLs (Vanhoenshoven, et al., 2016). Therefore, it may be said that the anti-malware device based on synthetic intelligence will assist detect recent malware attacks, and boom the scanning engine.

# Aim of project

Concerning the problem of finding malicious URLs, the two most important factors in finding the wrong URLs are mainly based entirely on symbols or sets of rules, and finding malicious URLs is based primarily on analytical behavior. How to find malicious URLs is based on unmodified tags or guidelines that can meet malicious URLs quickly and accurately. However, this method cannot meet new bad URLs that do not match the predefined set of terms or guidelines. The method of finding malicious URLs is based entirely on behavioral testing techniques that use various system algorithms to classify URLs mainly based entirely on their behavior.

In our study, the principle of device algorithms was used to measure URLs based primarily on URL characteristics and behavior. Actions are taken from static and dynamic URLs and are new to the text. Our recommended way to pursue analysis and classification of URLs is accurate or scary to visit them. All in all, our goal is to analyze malicious URLs using algorithms to read a specific gadget and compare the accuracy obtained with the result (Kazemian, and Ahmed, 2015).

# Objectives of project

The objectives of the project are:

* To analyze the dataset of malicious URLs
* To apply machine learning algorithms to detect malicious behavior by opening URLs.
* To compare the performance of different algorithms for detecting malicious URLs.

# Research Questions

The research questions are:

1. How do they recognize malicious URLs before opening them in the web browser?
2. How do differentiate malicious URLs from safe URLs?
3. What is the accuracy rate of detecting malicious URLs by various machine learning algorithms?
4. What is the accuracy rate of detecting malicious URLs by deep learning algorithms?

# Social Issues, Ethical Issues, and Legal Issues

Unfortunately, advances in generation include safety troubles that can be used to tune and close down users. These assaults encompass illegal websites that sell ads, and fraudulent commercial fraud by tricking customers into sharing touchy profiles for the reason of stealing cash or assets or making an investment in malicious code, and malware, in the consumer's tool. Given the type of possible assaults, as well as the numerous conditions in which such assaults can be completed, it's far hard to establish strong operations for cybercrime consciousness. The barriers to conventional security management technologies are deepening and responding to this significant growth in stressful security situations with new developments in modern computing technology (Manjeri, et al., 2019). However, there is an alarming shortage of security professionals capable of dealing with this difficult situation. The extent of these attack tactics is determined by maximizing broken URLs. Ethical matters are at the heart of cybersecurity practices as this performance is increasingly needed to stabilize and protect the individual entities and agencies to live happily ever after.

# Technical Work

From the various machine learning algorithm, we chose to apply certain algorithms to identify malicious URLs. We have collected information from the literature to get knowledge about the best algorithms from the research (Harikrishnan, et al., 2019). As we declared in the proposal, using python programming language, and jupyter notebook tool the analysis part will be done. The dataset will be analyzed through various visualization and techniques like the correlation of variables. We have selected, Naïve Bayes, decision tree, random forest, Knn, SVM, catboost, xgboost, and neural network to detect malicious URLs. Moreover, during neural network application, we have evaluated different hidden layers optimization to get a better result.

# Progress to Date

To date, we have examined data set data in the table in addition to accumulating statistical data in the database. During the pre-processing, the URL column was found to be a different identifier so we decided to remove that. There are always empty values ​​in the DNS\_QUERY\_TIMES and SERVER columns, so you should be able to downgrade these realities/realms at a whole new price without compromising on the details. The CONTENT\_LENGTH column is part of a larger concern; we cannot have enough records to discard multiple statistics (about half of the database) and aggregation may distort data. As there are many different skills. So, we decided to drop the column again. Next, we examined the correlation between other variables. We have found that there are amazingly linked skills out there. In addition, we have found many long URLs that will probably contain different characters. Therefore, we have removed some of the closest features. We will then describe the many machines that acquire knowledge of algorithms to solve this problem and check their accuracy performance result.

# Planned Work

This section describes approaches to detect malicious URLs from the selected dataset. This approach involves:

1) collecting data set information,

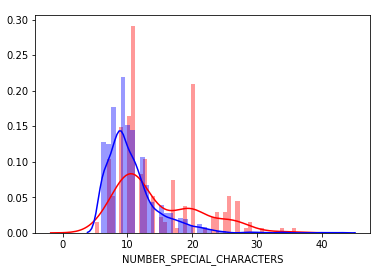
2) extracting features from the dataset,

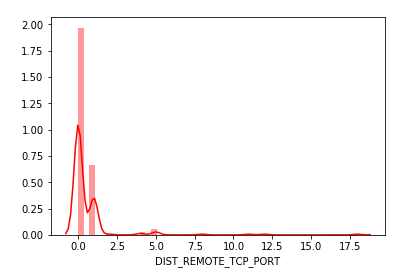
3) generating graphs based on malicious or benign websites,

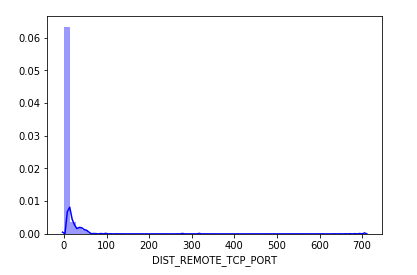
4) Analyzing the dataset through various machine learning algorithms and comparing the accuracies.

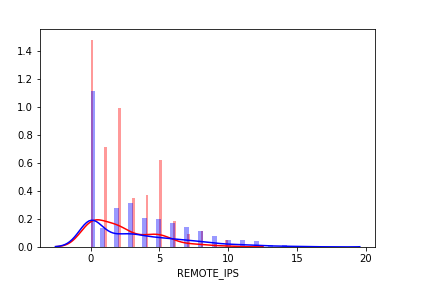
# Appendices

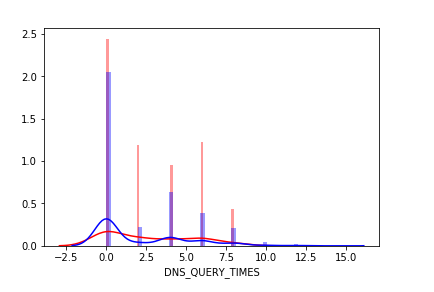
Some screen snippets have given in the following:

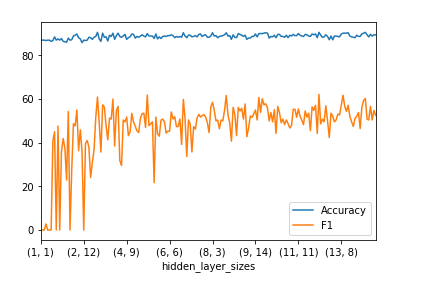












# References

Bhagwat, A., Lodhi, K., Dalvi, S. and Kulkarni, U., 2019, March. An Implementation of a Mechanism for Malicious URLs Detection. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1008-1013). IEEE.

Harikrishnan, N.B., Vinayakumar, R., Soman, K.P. and Poornachandran, P., 2019. Time split based pre-processing with a data-driven approach for malicious URL detection. In *Cybersecurity and secure information systems* (pp. 43-65). Springer, Cham.

Kazemian, H.B. and Ahmed, S., 2015. Comparisons of machine learning techniques for detecting malicious webpages. *Expert Systems with Applications*, *42*(3), pp.1166-1177.

Kotoju, R. and Vijaya Lakshmi, D., 2021. Study of Comparison on Efficient Malicious URL Detection System Using Data Mining Algorithms. In *Data Engineering and Communication Technology* (pp. 523-534). Springer, Singapore.

Li, T., Kou, G. and Peng, Y., 2020. Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods. *Information Systems*, *91*, p.101494.

Manjeri, A.S., Kaushik, R., Ajay, M.N.V. and Nair, P.C., 2019, June. A machine learning approach for detecting malicious websites using URL features. In *2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 555-561). IEEE.

Manyumwa, T., Chapita, P.F., Wu, H. and Ji, S., 2020, December. Towards Fighting Cybercrime: Malicious URL Attack Type Detection using Multiclass Classification. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 1813-1822). IEEE.

Patgiri, R., Katari, H., Kumar, R., and Sharma, D., 2019, January. An empirical study on malicious URL detection using machine learning. In *International Conference on Distributed Computing and Internet Technology* (pp. 380-388). Springer, Cham.

Sayamber, A.B. and Dixit, A.M., 2014. Malicious URL detection and identification. International Journal of Computer Applications, 99(17), pp.17-23.

Vanhoenshoven, F., Nápoles, G., Falcon, R., Vanhoof, K. and Köppen, M., 2016, December. Detecting malicious URLs using machine learning techniques. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1-8). IEEE.

Yuan, J., Chen, G., Tian, S. and Pei, X., 2021. Malicious URL detection based on a parallel neural joint model. *IEEE Access*, *9*, pp.9464-9472.