IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12223

# MALICIOUS URL DETECTION USING MACHINE LEARNING AND DEEP LEARNING

**Prof. Dhanraj S[1],**

**Nivas Bhagavath B K[2], Nisha A Jain[3], Manisha M[4], Keerthana M[5]**

Assistant Professor, Computer Science, East West Institute of Technology, Bengaluru, India[1]

Student, Computer Science, East West Institute of Technology, Bengaluru, India[2-5]

**Abstract**: The scope and severity of network concerns about information security have increased over time. Nowadays, most hacking methods target technology from start to finish while also exploiting human frailties. Pharming, phishing, and social engineering are just a few of these methods. One of the aspects of these Using damaging URRs to fool consumers is considered an assault (URLs). This is why identifying malicious Content is a hot issue right now. A variety of academic research has demonstrated several ways to identify malicious URLs using machine learning and deep learning technologies. Based on our hypothesized URL behaviours and characteristics, we provide a machine learning-based solution for detecting malicious URLs in this work. Furthermore, big data technology is applied to enhance the ability to appreciate fraudulent URLs based on aberrant activity. The proposed detection approach consists of a novel set of URL attributes and behaviours, a machine learning algorithm, and big data technologies. The results of the experiment indicate that the stated URL characteristics and behaviour can improve overall ability to identify risky URLs. This implies that consumers may successfully detect risky URLs using the suggested methods.

**Keywords:** phishing, machine learning, malicious URL detection.

## I. INTRODUCTION

Uniform Resource Reference points are used to identify Internet sources (URL). Protocol identifier, which advises which the two fundamental elements of the URL that were initially created in [1] are protocols to use and assistance name, which provides the IP address or the area designating where the help is placed. It is said that each URL has a unique structure and format. Attackers commonly attempt to modify one or more components of the URL shape. in an effort to trick users into sharing their malicious URLs. Links with negative user effects are known as malicious URLs. Consumers will be directed to specific resources or webpages via specific URLs in order for hackers to install malware on users' computers and reroute them to unwanted websites, harmful websites, virus downloads or other misleading websites. Moreover, dangerous URLs can be hidden in downloads that are considered secure, and they can swiftly propagate through file and message interchange in shared networks. Attack methods may make use of malicious Websites. Drive-by downloads, phishing, spam, and defacement.

Malicious URLs represent a significant risk to internet safety since they may be used in scams that cause victims to lose money, personal information, and accounts. Blacklists are the most popular method of recognising and thwarting these threats, but when combined with new URLs, this technique has a number of drawbacks. As a reason, Our study especially addresses methods for machine learning since that is where we are concentrating our attention. The amount and seriousness of the harm brought on by a lack of trust in public information are both rapidly expanding. At the moment, hackers mostly employ methods that aim at all technological knowledge while also takes advantage of individual weaknesses.

These strategies include social engineering, phishing, and pharming, among others. One step in carrying out these attacks is the use of malicious Universal Resource Locators to deceive users (URLs). Finding virus Websites has become a common past time as a result. Using the use of deep artificial intelligence and computer-based educational methodologies, several scientific studies have discovered a range of techniques for identifying bogus URLs. One of the main draws of modern social media platforms like Twitter and Facebook is URL sharing. According to recent study, these components comprise URLs in around 25% of all popularity notifications, or hundreds of thousands of URLs sent each day. There is a problem with this likelihood however, from unethical users looking to spread malware, phishing, and other low-quality

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023
DOI: 10.17148/IJARCCE.2023.12223

material. The issue of spam URLs in social networks eventually degrading the high-quality data available on these platforms has been recognised by a few recent initiatives. The Covid 19 has greatly influenced the development of internet enterprises, including e-commerce and social networking, and e-banking. Regrettably, technology advancements come with a variety of exploitation techniques for consumers. Such attacks frequently use rogue websites that collect a wide range of personal data that a hacker may use.

Internet resources are characterized with their URLs (Uniform Resource Locators). The attributes of URLs and two basic bits were introduced in:

The name of the resource indicating the protocol identity, which denotes the protocol to use, and the Port number or regional, which identifies the existence of the resource, are also obligatory fields. You can assume, because the format and shape of each URL is different. In an effort to trick users and spread dangerous URLs, attackers frequently try to change one or more of her URL form parts. Links that hurt users are recognised as malicious URLs. These URLs reroute users to resource or pages that provide attackers accessibility to the user's computer to run code, send users to undesirable, harmful, or other phishing websites, or let people download malware. To do On shared networks, malicious URLs might well be quickly distributed through the exchange of information and messages by masking yourself in seemingly secure downloads. Malicious URLs may be employed in drive-by downloads, phishing, spam, and many other attack methods.

Malicious URLs represent a significant risk to internet safety since they may be used in scams that cause victims to lose money, personal information, and accounts. The usage of watchlists is the most popular method for identifying and thwarting these threats. We are progressively focused on computer learning techniques, which is exactly the topic of our research, as this method has a number of issues when counterbalancing new URLs. The extent and severity of the harm caused by a lack of confidence in public information are both growing quickly. Currently, hackers mostly use techniques that target whole technological systems while also taking advantage of vulnerabilities in people.

Pharming, phishing, social engineering, and other strategies are some of these. One step in carrying out these attacks is to fool the people into browsing a dangerous URL (Uniform Resource Locator).Because of this, detecting harmful URLs is a popular past time today. Multiple malicious URL detection equipment, a lot of which are based on a knowledge of laptops and an in mastery methodologies, have been validated by numerous scientific inquiry. A major perk of contemporary social media platforms like Twitter and Facebook is sharing URLs. According to recent research, URLs are present in around 25% of all popular stories in these frameworks. This translates to daily URL exchange of hundreds of thousands. This potential, meanwhile, is exacerbated by dishonest users who do want to install malware, phishing, and other uninformed. Numerous recent initiatives have recognized the problem of spam URLs eventually. The data which is available in these systems is of questionable quality. The evolution of tech startups, including social networking, e-banking, and e-commerce, has been greatly affected by the Covid 19 Unfortunately, technology advancements include a number of exploitation techniques of individuals. Such attacks frequently use rogue websites that collect a variety of individual information that an hacker may use.

## APPLICATIONS

Malicious URL detection serves a variety of functions across a wide range of industries, including safeguarding users' financial information, enhancing internet security, warning users of fraudulent websites, and reducing the volume of spam emails that make it to their inboxes thanks to highly specialised junk mail filters.

We just provide model that categorises good and undesirable URLs, displays the assessment outcome considering accuracy metrics, and analyzes the obtained outcome for expert opinions and determines which is better.

## II. LITERATURE SURVEY

The following are the contributions we have made to this work:

(1) This work suggests a DCNN-based model for identifying malicious URLs. The remarkable multilayer convolution structure receives a new folding layer from the dynamic convolution method. The k-max-pooling layer is used in place of the pooling layer. The vector entrance dimension determines the breadth of the core layer of the dynamic convolution process has an unique mapping. Additionally, In order to obtain more in-depth points all over a wide swath, the depth of the current convolution layer and the size of the URL entered both affect how the pooling layer values are adjusted.

(2) The elements are gathered from the URL sequence throughout the function extraction and presentation procedure. In order to assess the classification model, we discovered variables which are integrated into a vector that the convolutional

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023
DOI: 10.17148/IJARCCE.2023.12223

neural network processes directly. In addition to streamlining the function extraction process and eliminating the need for manually point extract, this technique combines the advantages of personality embedding and phrase embedding. While personality embedding cannot be used to collect phrase sequence data, word embedding can. Unusual words and characters in the URL may be processed using character embedding. The dictionary and vector dimensions are also no longer too huge. In addition, the aggregate can accurately specify the URL and retain a memory area that will aid in URL information extraction.

(3) To determine the viability of the mannequin used in this study, we conducted a large number of comparison experiments. We carried out three different tests to illustrate that phrase embedding using personality embedding generates more accuracy than phrase embedding With persona embedding. To demonstrate how using a customised shape made up of a DCNN and certain variables retrieved from the URL may have a higher influence, we also perform three comparison trials.

Currently, there are two broad categories of approaches for detecting malicious URLs: standard methods that only rely on blacklisting and methods that heavily rely on machine learning. The primary list detection technique is first described in literature. Although this strategy as simple and effective, it is restricted and isn't able to realise freshly produced dangerous URLs. The literature demonstrates that. Currently, there still are two broad categories of techniques for identifying malicious URLs: standard methods that just rely on blacklisting and methods that heavily rely on machine learning. The primary list detection technique is first described in literature. Although this strategy as simple and effective, it is restricted and isn't able to recognize freshly produced dangerous URLs. The literature emphasizes that point.

Currently, there still are two broad groups of techniques for identifying malicious URLs: standard methods that just rely on blacklisting and systems that heavily rely on machine learning. The primary list detection technique is first discussed in the literature. Although this methodology as simple and effective, it is restricted and isn't able to materialize freshly produced malware URLs. The literature makes that point Several methods employing deep learning models have been proposed by numerous scientists to identify malicious URLs and assess the hazard of a URL just based on the strings available in the URL. These methods can consistently extract trustworthy information from the URL. For instance, literature uses the personality stage of the cyclic neural community mannequin to categorise URLs given by DGA. The research suggests using a heavy laptop workload to find dangerous URLs. Literature uses character-level semantic factors in tandem with the n-gram model with deep learning to determine whether or not DGA creates the URL. The literature lists a variety of deep learning techniques for detecting bad URLs. It comprises the deep convolution structure, the bidirectional Autoencoder, the blended CNN and LSTM designs, and the single continuous temporary memory (LSTM) styles.To avoid confusion, the family name must be written as the last part of each author name (e.g. John A.K. Smith). Each affiliation must include, at the very least, the name of the company and the name of the country where the author is based (e.g. Causal Productions Pty Ltd, Australia).

## III. PROPOSED METHODOLOGY

CNNs and RNNs have been included into neural network structures to fulfil this purpose. The diagram appears as follows: RNNs and LSTMs are examples of sequence generator architectures that can begin by translating a feature vector with a set length from either a picture. To prepare a list of labels or terms for your image, follow this procedure ResNet50 was the encoder chosen for this project. Using pre-trained algorithms, the ImageNet dataset's millions of images were divided into 1000 categories. To utilise this networks, remove the top layer, which contains 1000 neurons and is solely used for ImageNet classification, and substitute it with a linear layer that contains twice as many neurons. This is due to the network's weights being altered to recognise a number of traits that are prevalent in nature. With LSTM, many neurons are created. Long Short-Term Memory (LSTM) cells make up an RNN, which is used to recursively generate captions from the input pictures. To recall knowledge from earlier phases, these cells employ the principles of repetition and gates. If you want more details, you may watch or read this. The output layer receives the combined output from the encoder and decoder and generally guesses the next word based on the visual and current sequence.

The following is the proposed system:

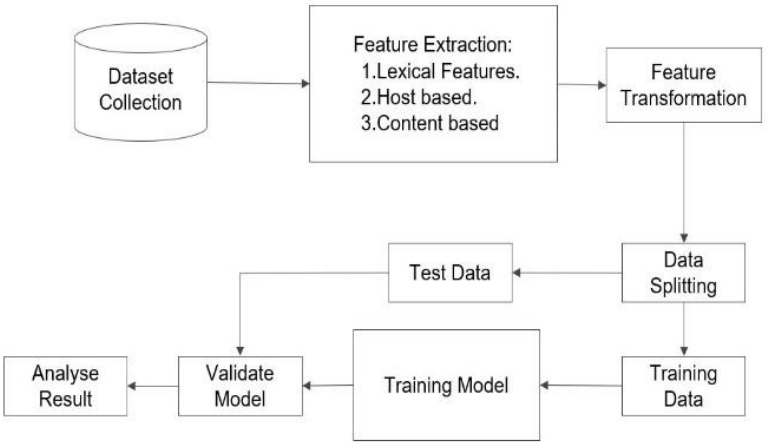
- The graphical user interface is the first alternative (GUI). At this phase, the user alters the system.
- The user must sign in or register if it is their first first-time visitor.
- At that point, the user can upload an image and obtain a description.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12223

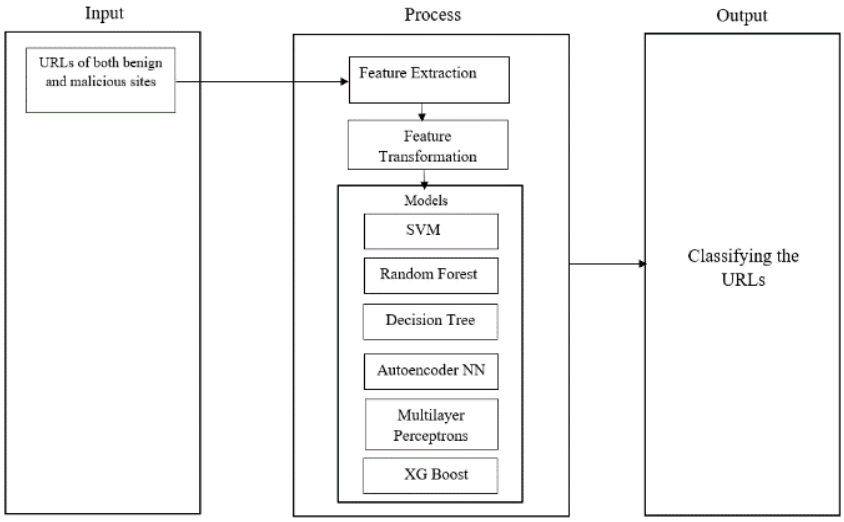



## IV. MODULE DECOMPOSITION

- Importing libraries and extracting datasets are among the modules.
- For record URL records, use the.csv format. Next, import the different libraries that the other modules require. Pandas data frames are constructed from CSV files. Give the URL to be used for feature extraction.
- Extracting features:
- Create, change, and manipulate IPv4 and IPv6 addresses as well as networks using IP addresses. re In order to extract the elements, a regular expression is employed. who:
- Develop a simple Python module that can be exported and generates parsed WHOIS information for a given domain. A collection of URL manipulation components. URL lib, a programme to read and open a URL, is available in the package urlib.
- Functional transition
- Depending upon the circumstances, feature values are assigned as 0 (good) or 1 (poor).
- Considering every one of the features:
- During a feature transformation stage, features taken from diverse sources are integrated for further processing.
- Data sets with feature vectors shared:
- Create training and test data sets from entire data set. supplying several models.
- He builds his six models using deep-learning and machine learning methods.
- Evaluating the construct models:

The Accuracy score measure is used to score models. It represents the proportion of accurate predictions to all input samples. According on training and testing accuracy, contrast all models.

## V. BLOCK DIAGRAM



International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023
DOI: 10.17148/IJARCCE.2023.12223

## VI. CONCLUSION

Its use of malicious websites that appear to be legitimate web pages and URLs is a typical social engineering tactic. In order to predict harmful websites, The goal of this project is to apply the available data sets to train deep neural networks and machine learning models. The targeted URLs and the website's content-based functionality are collected from a dataset containing both the harmful and benign URLs of the website. Measure each model's performance level and make comparisons. In order to more correctly identify dangerous URLs.

## REFERENCES

[1] D. J. Lemay, R. B. Basnet, and T. Doleck, "Examining the relationship between threat and coping appraisal in phishing detection among college students," Journal of Internet Services and Information Security, vol. 10, no. 1, pp. 38–49, 2020.
[2] H. Kim, "5G core network security issues and attack classification from a network protocol perspective," Journal of Internet Services and Information Security, vol. 10, no. 2, pp. 1–15, 2020.
[3] K. Aram and J. O. SoK, "A systematic review of insider threat detection," Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), vol. 10, no. 4, pp. 46–67, 2019.
[4] R. B. Basnet and R. Shash, "Towards detecting and classifying network intrusion traffic using deep learning frameworks," Journal of Internet Services and Information Security, vol. 9, no. 4, pp. 1–17, 2019.
[5] F. Valenza and M. Cheminod, "An optimized firewall anomaly resolution," Journal of Internet Services and Information Security, vol. 10, pp. 22–37, 2020.
[6] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.
[7] C. Seifert, I. Welch, and P. Komisarczuk, "Identification of malicious web pages with static heuristics," in Conference on Telecom Networks and Applications, 2008. ATNAC 2008. Australasian. IEEE, 2008, pp. 91–96.
[8] S. Sinha, M. Bailey, and F. Jahanian, "Shades of grey: On the effectiveness of reputation-based "blacklists"," in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 57–64.
[9] Leo Breiman.: Random Forests. Machine Learning 45 (1), pp. 5- 32, (2001).
[10] D. Sahoo, C. Liu, S.C.H. Hoi, "Malicious URL Detection using Machine Learning: A Survey". CoRR, abs/1701.07179, 2017.