Detection of Malicious URLs

Dr S L Aruna Rao *, A. Meher Gayatri Devi †, T.G.S.N Sai Chandana ‡, J. Anuhya§
BVRIT HYDERABAD College of Engineering for Women, Hyderabad, Telangana, India 500090

***Abstract*—The World Wide Web has evolved into the web-based application with a greater prone to cyber intrusions. As technology evolves digitally, it becomes much more vulnerable to malicious data that could lead to cyber-attacks. Automated malevolent software, usually known as malware, pose a significant threat to computers and information security linked to the Web with the expansion of the underground digital economy. Malicious URLs are widely used in extortion, malware, as well as other types of cyber-attacks. It's essential to easily determine phishing URLs. Previous studies used blacklisting, regular expressions, and signature matching approaches. These approaches do not affect detecting existing malicious URL variants or entirely newly found URLs. This problem can be mitigated by proposing a machine learning-based solution. The detection approach uses machine learning algorithms such as Logistic Regression, Naive Bayes, Decision tree, SVM classifier, Random Forest, and Extreme Gradient Boosting to validate the accuracy and efficiency of this method. Finally, experimental results show that this method has good detection performance.**

***Index Terms*—cyber-attacks, malware, phishing, spamming, black listing, Logistic Regression, KNN, SVM, Random Forest, Decision tree, Extreme Gradient Boosting Algorithm.**

I. Introduction

Users can benefit from malicious URL detection, since it prevents users from accessing malicious links and alerts them about the risks involved. In order to trick users into browsing malicious URLs and installing trojans on their systems or exposing sensitive information, hackers frequently utilise spam and phishing. Malicious URL detection software can enhance users in detecting and safeguarding from cybersecurity threats.

One of the most important security breaches on the Internet is the one with web application security. It is absolutely important to do a comprehensive and systematic evaluation of encryption and authentication. URLs are text strings that can be read by users, but client applications cannot directly use them. The browser converts a URL into directions on how to find the server hosting the site and where the site or resource is located inside that host through a multi-step resolution process.

There are three common ways of malicious URL detection: content-based detection, signature-based detection and anomaly-based detection. Content-based detection is a method that compares the content of a URL against a database of known malicious URLs. If the content of a URL matches that of a known malicious URL, the URL is flagged as being malicious. Signature-based detection is a method that compares the digital signature of a URL against a database of known malicious URLs. If the digital signature of a URL matches that of a known malicious URL, the URL is flagged as being malicious. Anomaly-based detection is a method that compares the behaviour of a URL against a database of known malicious URLs. If the behaviour of a URL is abnormal, the URL is flagged as being malicious.

Most of the business enterprises rely on the Internet services available on the WWW. These services are however, susceptible to cyberattacks. Most cyber-attacks usually occur when users click on malicious URLs. URLs are used on WWW to access legitimate resources. When they are used for other reasons, they show a threaten about data availability, controllability, confidentiality, and integrity. There is a continuous surge in web attacks launched through phishing, spam, and malware-ridden URLs. Malicious URL Detection and Filtering Techniques. URL filtering is inspecting, classifying, and blocking web URLs. Malicious URL detection and filtering techniques prevent users from accessing malicious websites. Various methods are used for detecting malicious URLs, including blacklists, heuristics, and machine learning.

The World Wide Web has become an application on the Internet that poses a significant risk of cyber attacks. As technology grows over the Internet, it is exposed to malicious data that can lead to cyber-attacks. With the rise of the underground Internet economy, automated malicious programs commonly known as malware, pose a significant threat. Phishing URLs are usually used to gain unauthorized access to personal information, for example banking details, passwords, and names. When attackers obtain this information, they proceed to steal money or log into private institutions like government websites, school websites, and hospital medical report databases. Spam URLs usually intend to perform unauthorized advertisements, especially for well-known brands like Amazon, Taobao, Alibaba, and PayPal. Malware URLs usually spread malicious software that, once downloaded, will infect the host. Malicious software can launch silent attacks on a user's computer by installing malicious programs.

An attacker can create a custom malicious domain to avoid exploiting legitimate websites to host malware. A custom malicious domain is registered by an unknown attacker to evade detection and remains active for a short period. This design is primarily used for widespread infections, not targeted infections. However, attacks are more targeted due to changes in the attack patterns used in drive-by downloads. The context of malware infection remains the same, but the approach is different.

Malicious URLs can harm a person in various ways, sending through emails being one of them. The sender's email address should be verified because cloned emails are standard in fraudulent messages. Despite the sender's email address possibly revealing their malevolent intentions, the email may appear to be from a real organization. The sender's email address might not even belong to the organization the message claims to be from. The sender's display name does not correspond to the email address. You should not comply with any requests or demands made in the email because this is a clear sign that it is possibly malicious.

II. ARCHITECTURE

Malicious URL Detection using machine learning approach tries to analyze the information of a URL and its corresponding websites or webpages, by extracting good feature representations of URLs like general features, count features, binned features and ratio features. In the UI an user enters the URL and the results will display the type of URL i.e, benign or malware.

Proposed architecture contains three main phases, they are Feature extraction phase, Detection phase and Interface phase. In feature extraction phase general features and additional features of an URL are extracted. In detection phase, the data is splitted into test and train and machine learning algorithms are used to train the data. In interface phase, UI is implemented using flask framework.

First, the data is given as input and is preprocessed. The preprocessing procedure makes sure the data are in a usable format and transfers the values to vector features. Feature extraction is the core of the processing; it identifies the essential characteristics of malicious URLs. Labeled URL data is collected from the Feature Analysis, and the features are fed into an algorithm. The machine learning algorithms like Decision Tree, Logistic Regression, LightGBM, XGB, and SVM are used. The machine learning module contains two independent processes. The training process uses the machine learning algorithm to build the detection model. The testing process loads the detection model to test unknown URLs and outputs the test result. The evaluation module is devoted to estimating whether the given URL is malicious or not, and the result is updated to the model.

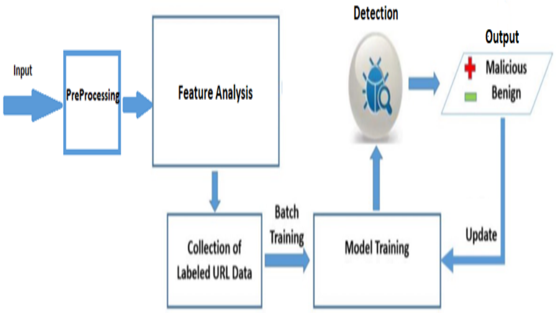


Fig. 1. Framework for Malicious URL Detection

III. MODULES

The modules included in Detection of Malicious URLs are

A. Feature Extraction

This module is engineering oriented, which aims to collect relevant information about the URL.The unstructured information about the URL (e.g. textual description) is appropriately formatted, and converted to a numerical vector so that it can be fed into machine learning algorithms. For example, the numerical information can be used as is, and Bag-of-words is often used for representing textual or lexical content. , an automatic feature extraction algorithm is proposed to validate the effectiveness of each feature. This module proposes a new method to evaluate features automatically and extract more effective features to detect malicious URLs.The features are extracted from the URLs and categorised based on background information.

B. Detection

This detection framework can be divided into three main modules: the processing procedure, the machine learning procedure and the evaluation procedure. The processing procedure makes sure the data are in a usable format and transfers the values to vector features. The machine learning module is where the core algorithms reside; depending on the algorithm, it feeds the algorithm valid clean data from the processing module and extracts knowledge or information. The evaluation module is to estimate whether the preceding processing can result in good performance and be put to use. Different algorithms are adopted at analysing different data types. To obtain the best performance, multiple classification models are tested to obtain the most appropriate algorithm.

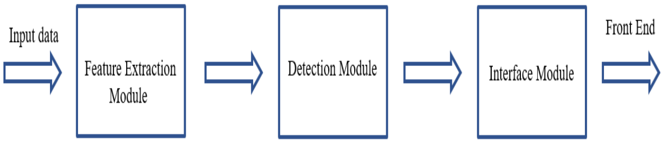


Fig. 2. Modules

C. Interface

In this module we introduce a UI using flask framework where the user can enter the URL to check if it is a malicious or not. If the URL is legitimate or benign, the URL will redirect to the entered site. If the URL is malicious, phishing or defacement, an alert message is displayed.

IV. Dataset

The dataset used here is **Malicious_phish.csv**. It has 6,51,191 records consisting of two columns namely URL and type. It contains four different types of URLs such as Benign - 66%, Defacement - 15%, Phishing and Malware - 19%.

The safest URLs are those that are benign. Benign URLs were acquired from prominent Ranking websites. To retrieve the URLs, the websites were run through a Heritrix web crawler.

Defacement is an attack where malevolent parties hack into a website and substitute its information with their own contents. Malware, often referred as ”malicious code,” is a file or piece of software that can do almost anything a hacker wants it to do—infect, discover, pilfer, or accomplish other undesirable actions. Furthermore, there are countless ways to exploit computer systems due to the wide variety of viruses.

Malicious hackers exploit spoofing, a form of social engineering tactic, to mislead victims into revealing confidential material or acquiring malware.

V. Algorithms

The basic aspect behind detection is to enhance the detection performance by feeding accurate, clean data to the machine learning classification model. The most essential factor is figuring out the best way to analyse the online data because many machine learning categorization models have good performance and might be used. SVM, Decision Tree, Random Forest, XGBoost, Logistic Regression, KNN employing cosine metric, and other classifiers are utilised in the classification framework. The initial application of these boosting ensembles was for binary classification.

A. Logistic Regression

Basically, supervised classification is what logistic regression performs. For a specific set of characteristics (or intakes), X, the target output (or outcome), y, can only take distinct values in a classification problem. Contrary to popular perception, a regression analysis is a logistic regression. Even though it is recognized as logistic regression and utilizes the concept of predictive analysis as regression, logistic regression mainly utilizes classification models to identify and classify observations.

B. KNN Using Cosine Metric

Both classification and regression issues can be addressed using the KNN algorithm, a fundamental, quick and easy supervised machine learning approach. Machine learning models predict expected output using a set of input data. One of the most essential types of machine learning algorithms, KNN is mostly used for classification. The classification depends on how the neighbouring data points are labelled.

Detecting whether two vectors are pointing in the same direction requires calculating the cosine of the angle between the two vectors. The distance measure is mostly used to compare and contrast two vectors. As a result, the Cosine Distance's span is 0 to 1, as well. The objects are identical if the cosine distance is zero (0). When the cosine distance matches one, those things are clearly distinct.

C. SVM

An approach for supervised machine learning called SVM can be applied to classification or regression issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method.

D. Decision Tree

The decision tree algorithm is a member of the supervised learning algorithm family. Contrary to other supervised learning methods, the decision tree technique is equipped to handle both classification and regression issues by learning straightforward decision rules derived from previous data. A decision tree is used to build a training model that may be used to predict the class or value of the target attribute (training data). In decision trees, we start at the tree's root when anticipating a record's class label. We evaluate the root attribute's values compared to the value for the record.

E. Random Forest

A popular algorithm for classifying and interpreting data is called ”random forest,” which is supervised machine learning. On samples collected, decision trees are constructed, and the mean decision for classification and regression is dependent on the majority vote of the decision trees. Being able to handle large datasets with both continuous variables, as in regression, and categorical variables, as in classification, is one of the Random Forest Algorithm's key features. For classification issues, random forest performs better.

F. XGBoost

A gradient boosting framework for faster and more effective computation is provided by the XGBoost algorithm. The boosting method offers improved regularisation and the capacity to manage missing values. Due to its capacity for loss minimization, XGBoost is widely employed. Decision trees are generated sequentially in this approach. Weights are significant in XGBoost. Each independent variable is given a weight before being fed into the decision tree that forecasts outcomes. Variables that the tree incorrectly predicted are given more weight before being placed into the second decision tree. These distinct classifiers/predictors are then combined to produce a robust and accurate model. Comparing with all above mentioned machine learning algorithms Extreme Grading Boosting algorithm has resulted with higher accuracy about 96%.

VI. Results

Figure 3 shows the home page. To check with the URL type, click on predict hyperlink on top of home page or the button-let's go at the bottom of the page.

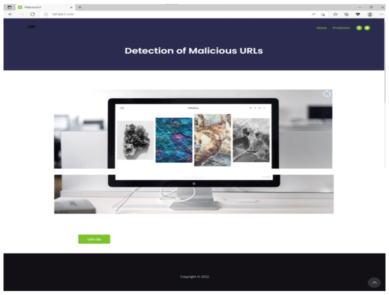


Fig. 3. HomePage

After clicking on the hyperlink or button, the user will be redirected to the page shown in figure 4. In this page the user can check the type by entering the URL in the input field.

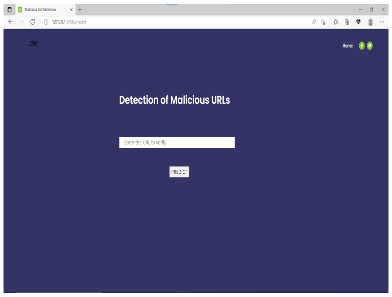
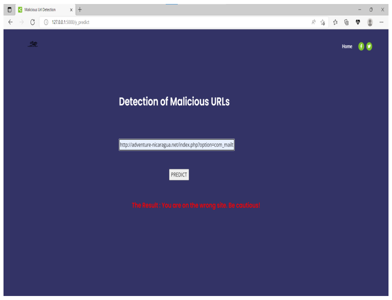


Fig. 4. Detection Page

Figure 5 shows the detection page when the entered URL is a malicious URL, a warning message is shown in red.

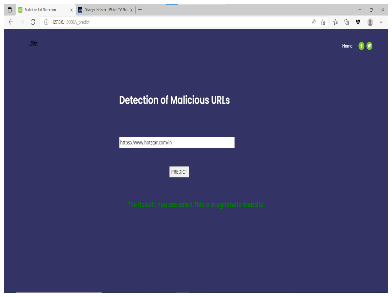


http://adventure-nicaragua.net/index.php?option=com_mailt

The Result : You are on the wrong site. Be cautious!

Fig. 5. Identification of Malicious URL

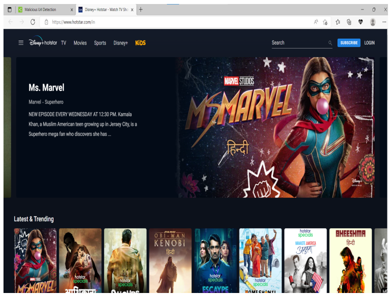
Figure 6 shows detection page when the entered URL is benign. If the entered URL is benign/legitimate, A new tab is opened with the given URL. Since the enter URL is benign the redirected page is shown in figure 7.



https://www.hotstar.com/in

The Result : You are safe!! This is a legitimate Website.

Fig. 6. Identification of Benign URL



NEW EPISODE EVERY WEDNESDAY AT 12:30 PM. Kamala Khan, a Muslim American teen growing up in Jersey City, is a Superhero mega fan who discovers she has ...

Fig. 7. Redirection to Benign Site

VII. Conclusion

A new approach for automatically classifying URLs as either malicious or benign based on machine learning techniques is proposed. By analysing the features extracted from an URL this experiment shows an better approach than blacklist methods, which cannot predict the status of previously unseen malicious patterns. Furthermore, we have compared with different classification algorithms like logistic regression, SVM, KNN, random forest, decision tree and found that the Extreme Boosting algorithm has the best performance. In future work, the detection can be made more specific by analysing the behaviours of malicious URLs. The precision may be improved to 99.99% by perfecting the malicious key words and extracting more features. We would also like to implement chrome extension for easy use of the application. This can be considered as one of the significant directions for future work.

References

[1] A. S. Popescu, D. B. Prelipcean and D. T. Gavrilut, ”A Study on Techniques for Proactively Identifying Malicious URLs,” 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2015, pp.204-211, doi:10.1109/SYNASC.2015.40.

[2] M. N. Feroz and S. Mengel, ”Phishing URL Detection Using URL Ranking,” 2015 IEEE International Congress on Big Data, 2015, pp. 635-638, doi: 10.1109/BigDataCongress.2015.97..

[3] S. B. Rathod and T. M. Pattewar, ”A comparative performance evaluation of content based spam and malicious URL detection in E-mail,” 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), 2015, pp. 49-54, doi: 10.1109/CGVIS.2015.7449891.

[4] R. Kumar, X. Zhang, H. A. Tariq and R. U. Khan, ”Malicious URL detection using multi-layer filtering model,” 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2017, pp. 97-100, doi: 10.1109/ICCWAMTIP.2017.8301457.

[5] T. Manyumwa, P. F. Chapita, H. Wu and S. Ji, ”Towards Fighting Cybercrime: Malicious URL Attack Type Detection using Multiclass Classification,” 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 1813-1822, doi: 10.1109/BigData50022.2020.9378029.

[6] Cui, Baojiang He, Shanshan Yao, Xi Shi, Peilin. (2018). Malicious URL detection with feature extraction based on machine learning. International Journal of High Performance Computing and Networking. 12. 166. 10.1504/IJHPCN.2018.094367.

[7] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. “Bayesian CART Model Search.” Journal of the American Statistical Association, Vol. 93(443), pp 935–948, September 1998.

[8] Fadi Thabtah Maher Aburrous, M.A.Hossain, Keshav Dahal. “Intelligent phishing detection system for ebanking using fuzzy data mining.” Expert Systems with Applications, Vol. 37(12), pp 7913–7921, Dec 2010.

[9] Dan Steinberg and Phillip Colla. “CART: Classification and Regression Trees.” The Top Ten Algorithms in Data Mining, pp 179–201, 2009.

[10] Ying Yang and Geoffrey I. Webb. “Discretization for Naive-Bayes learning: managing a discretization bias and variance.” Machine Learning, Vol. 74(1), pp 39–74, Jan 2009.

[11] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access, 7, 41525-41550.

[12] Alazab, M., Layton, R., Broadhurst, R., Bouhours, B. (2013, November). Malicious spam emails developments and authorship attribution. In Cybercrime and Trustworthy Computing Workshop (CTC), 2013 Fourth (pp. 58-68). IEEE.