International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

SaleSurf -Strengthening the Web's Armour, One URL at a Time - Malicious URL Detection using Machine Learning Models

Sunidhi Rathod[1], Nehal Panchal[2], Atul Sarowa[3], Prof. Vricha Chavhan[4]

[1,2,3,4]Author, K.J. Somaiya Institute of Technology Mumbai, India

Abstract

One of the most common cybersecurity vulnerabilities involves malicious websites or URLs. Every year, individuals and organizations suffer major financial losses from using harmful content such as spam, malware, inappropriate advertising, and scams that encourage visitors to cheat. These malicious URLs are often promoted through emails, advertisements, web search results, or links to other websites. Considering how many users click on these malicious URLs, there is an urgent need for a reliable system that can classify and identify dangerous URLs; In particular, phishing, spam and malware attacks are increasing. Data volume, updated attack models and strategies, correlation between URL features, lack of data, inconsistent data and the presence of outliers make the division of labor very difficult. In our research, we focus on negative URL search to gain more insight. Our information is divided into four main categories: phishing, harmless (safe), tampering, and malware. We have collected a large database of 651,191 URLs to support our application. To achieve the goal of identifying and identifying malicious URLs, we use three machine learning algorithms: Random Forest, LightGBM, XGBoost, Logistic Regression, CNN and Ensemble Model.

Keywords: URL discovery, network security, machine learning, URL isolation, phishing, benign, tampering, malware.

INTRODUCTION

Hackers often use popular content and videos to spread hate speech or links to phishing websites, malfunction users' computers, or deceive users by obtaining personal information from the Internet. Most attacks involve hate websites or feeds luring victims, such as general email hate attacks, SQL injections, distributed denial of service (DDoS) attack, or direct attack. Attacking the defense is different. Increasing awareness of prevention and care can improve freedom of truth. Data protection includes three aspects: confidentiality, legality and emergency. Data protection is necessary for the planned role in the Internet, the operation of the Internet and the Internet of Things. Among these, the most common attacks across the Internet are indirect attacks,

including indirect downloads, security attacks, phishing sites, DDoS and SQL injections. When a customer accesses a malicious website, the program on the website will look for malicious customer orders to try to attack. If the attack is successful, the final solution will definitely be to download and eliminate the malware or bug.

At this point the developer developed a plugin for the botnet operator. There are many new studies using

machine intelligence methods to detect malicious URLs. The research continues based on various datasets and various distinctive location detection methods using various distinctive location profiles, removal techniques, machine learning intelligent algorithms, interactive cluster configuration and network traffic strategies. In this study, we combined these studies to detect hateful websites in four ways: focusing on network link, URL keyword, website news, and network content. In this study, machine learning was used to ensure the authenticity of the website host, to create the content of the website, and to detect and block malicious URLs. It causes the blacklist to disappear, which means identifying many unknown facts and identifying many bad URLs.

LITERATURE SURVEY

Blacklists provide a way to identify bad (bad) URLs by creating and maintaining a list of URLs that fall into the threat category. Each time a new URL is discovered, an optimal site search is performed and if the URL appears on the blacklist, it is considered dangerous; If not, it is considered benign or harmless. Since creating new URLs is not a complex task, the blacklisting process cannot control and amplify all dangerous URLs on a large and large scale, thus failing to define speed, privacy, and security [5]. This is particularly concerning and dangerous because attackers can create new URLs through a process that bypasses all blacklists.

Cui et al. proposed a search method that uses machine learning techniques to automatically classify URLs as malicious or malicious websites. [1] The authors use statistical analysis methods, including gradient learning and subtraction, to select the main features of the sigmoid level. The accuracy of its findings is 98.7%.

Altaher [2] proposed a hybrid scheme to identify websites as legitimate, suspicious or phishing. This method is the result of a two-stage combination of two machine learning methods: the support vector machine (SVM) algorithm and the K nearest neighbor (KNN) algorithm. Tests using the scheme achieved an accuracy of 90.04%. Using the HTTP protocol, Taoet al. [3] proposed an automatic method to identify fake websites. This approach considers both domain-based and HTTP session header-based features. The authors developed a classification system that uses machine learning to identify malicious websites. This method classified 92.2% of bad websites.

Wang [4] developed a hybrid detection method that combines static and dynamic detection methods. Static properties fall into three categories: URLs, HTML files, and JavaScript. They also divided websites into three categories: unknown, problematic, and harmless.

Using the Receiver Operating Procedure (RoC) curve to analyze the performance of various features, the accuracy of the URL feature was found to be 93%, which is higher than other features (JavaScript and HTML).

Şahingöz et al. He proposed a method to detect phishing URLs. [5] used various machine learning algorithms. including language vectors, composition, and natural language processing (NLP) features. When NLP and word vector features are used simultaneously, the performance of NLP-based features and word vectors increases by 2.24% and 13.14%, respectively. NLP features of the function according to language vector properties.

To identify fraudulent websites, Sirgeldin et al. [6] It uses machine learning techniques based on URL message and page content features. Artificial Neural Networks (ANN) provide the best features to achieve the best performance, which leads to false alarms.

Liu et al. used machine learning in their experimental study on dangerous URL detection [7]. To identify

the key characteristics of bad URLs, the authors focused on the character frequency and structure of the URLs. The results show that using the random forest method with the extracted URL results is the best and most effective method.

Two methods were adopted by Zhao et al. [8] used machine learning models to classify various problematic URLs. GRU (or Gated Recurrent Neural Network) was compared to 2) Random Forest (RF) classifier (using carefully selected features). six sorts of URLs (XSS injection, SQL To compare the two strategies, several attacks (such as injection, directory traversal, legitimate, sensitive file attacks, and others) are employed.

The outcomes demonstrated that the GRU model outperformed the RF model by 2.1% and reached an accuracy of 98%.

A solution that use machine learning algorithms to identify fraudulent URLs was proposed by Patgiri et al. [9]. They calculated the classification accuracy of the classifiers by dividing the collected dataset into training and test data in three different ratios. The comparison results show that the 80:20 ratio is more accurate, the average accuracy of Random Forest (RF) is higher than Support Vector Machine (SVM), and the true standard deviation of RF is greater.

< br> Novel Coronavirus (COVID-19) has affected people's behavior and decisions worldwide, causing physical and mental harm. COVID-19-related cyberattacks have increased as many services have moved online due to the pandemic. This study draws on psychological and traditional crime theories to explore online fraud and online security issues in the COVID-19 era.

METHODOLOGY

1. **Data Collection:** We collected data from the Kaggle repository listing problematic, malicious, tampered and malware URLs. It is worth noting that this data was carefully selected to ensure that it does not contain blank or empty cells.
2. **Data Cleaning:** Our data preparation process consists of several steps. We handle missing data, remove redundant features, normalize numeric values, code categorical variables, and normalize data for similarity.
3. **Model training:** We use a variety of machine learning techniques to build our predictive models. Specifically, we used the Random Forest classifier, the Light GBM classifier, and the XGBoost classifier. We use the Sklearn Python library to train this model using 80% of the data.
4. **Model Validation and Optimization:** The remaining 20% of the data is kept for validation purposes. We performed hyperparameter tuning to optimize the model, aiming to improve precision, F1 score, recall, and accuracy.
5. **Comparison model:** We evaluated and compared the performance of the learning classification system using appropriate metrics. This step allows us to choose the model that performs best according to our criteria.

Classification Technique:

Classification is a machine learning process in which data is fed into a model with matching labels. This allows the model to learn from previous information contained in the training data. Classification can be used for both structured and unstructured data. Steps in the classification process include collecting preliminary data, training the model, and using the model to classify new content.

In classification, the different categories or groups we focus on to provide data points are often called

classes, collections, or categories. While the category represents the set of posts across the entire dataset, tags are specific to individual content.

There are two main types of classification:

1. **Binomial classification:** In this type, items are divided into two groups, usually representing a binary value such as spam or not spam.
2. **Multi-category classification:** In contrast, multi-category classification will provide data for more than two different categories. This is often used for tasks such as tweet sentiment analysis, where tweets can have positive, negative or neutral sentiments, or for classifying various images such as fruits, animals or insects.

Classification plays an important role in many applications, including classifying emails as spam or not spam, document analysis, and many applications found in patches such as biology and zoology. Content distribution. Botanical.

Random Forest Algorithm: This versatile algorithm can be used to solve both regression and classification problems. It can be used for regression work, where the goal is to estimate constant numbers, as well as for classification work, which involves assigning scores to different classes or groups.

Light GBM Classifier: Light GBM Classifier is a decision tree based gradient boosting framework. This classifier uses gradient boosting to make the prediction model stronger by combining multiple weak models (usually decision trees). It optimizes the use of memory and computing resources, making it ideal for large files and limited space. It is a decision tree-based gradient boosting framework that performs well on training models and reduces memory, making it useful for many tasks of machine learning.

XGBoost Classifier: The XGBoost classifier is an idea derived from the research work of the University of Washington. It is used as a C++ library and is designed to improve the training process in gradient support. This classification brings improvements to the gradient boosting technique, making it efficient and effective for many tasks of machine learning. XGBoost is widely recognized for its robustness, robustness, and ability to solve back-and-forth problems.

CNN: The neural network approach for identifying malicious URLs consists of two main stages: online training and online testing. Initially, simple features are employed for a preliminary screening process before being applied to URLs that remained indistinguishable. Subsequently, a CNN-based method is utilized to discern malicious URLs based on webpage content.

Logistic Regression: Logistic regression is a statistical method used for binary classification tasks. It predicts the probability of an outcome based on input features by applying the logistic function to a linear combination of those features. Unlike linear regression, it models the probability of the outcome being true rather than predicting a continuous value. It's widely used in fields like medicine, finance, and marketing for its simplicity, interpretability, and effectiveness in handling linearly separable data.

Ensemble Model: An ensemble model combines predictions from multiple individual models to improve overall performance. Each model might specialize in different aspects of the data, resulting in a more accurate and robust final prediction. Techniques like bagging (Bootstrap Aggregating) and boosting (training models sequentially, emphasizing the mistakes of previous models) are common in ensemble methods. By leveraging diverse perspectives, ensemble models often achieve higher accuracy than any single model alone.

Data Visualization:

We use publicly available data from 651,191 websites in the Kaggle repository to train and evaluate machine learning models.

The database contains four categories of URLs: benign (428,103), tampered (96,457), phishing (94,111), and malware (32,520). We use the word cloud module in Python to find the most frequently occurring words in a particular directory.

The most frequently used words are found in malicious URLs, modified URLs, phishing URLs and malware URLs. The results show that words such as "Wikipedia", ".org", "Youtube" and "Facebook" occupy prominent positions in the URLs. Similarly, the WordCloud module in Python is used to find words used in other types of URLs in documents.

The WordCloud module in Python is a tool for showing word frequency and importance in specific data. It works on threshold principle and words beyond the threshold will appear in the output. Big words have more importance, while small words have more importance.

A. Feature Extraction:

Feature extraction is the process of changing or improving features to improve the performance of learning models and help make better decisions. It also speeds up the process by reducing waste. Two commonly used methods in extraction are PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis).

Since machine learning models cannot interpret text directly, categorical features such as IP usage, anomalous URLs, and Google indexes are encoded as numerical values. In this case, the encoder is used as the coding process. The Target column assigns a value of 0 to harmless websites and a value of 1 to malicious websites.

B. Feature Scaling:

Feature scaling is a technique used to transform data into a fixed space and is often used when the data is first used to control different data. When scaling information is removed, machine learning models tend to return higher values than lower weights. The two most common measurement methods are standardization and normalization.

C. Normalization:

This model rescales the numerical value in the range 0 to 1. Normalization is used to improve data organization to remove outliers and improve repair efforts.

Performance Metrics:

The data is partitioned into an 80:20 ratio for training and testing purposes. Various metrics are used for evaluation, including:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FP)

F-Score = 2 *[(Precision * Recall) / (Precision+Recall)]

RESULTS AND DISCUSSIONS

A. Confusion Matrix:

This matrix shows the effectiveness of the machine learning classification model used to predict the classification form for each example. Essentially, it calculates the probability of true positives (TP), negatives (TN), negatives (FP) and negatives (FN) produced by the classification model when describing

the dataset. The figure below represents the binary classification matrix:

This study evaluates the effectiveness of Random Forest, LightGBM, XGBoost, CNN individually and when used in the Ensembler model using logistic regression in detecting malicious URLs. The combined approach using different datasets outperforms the classification method alone, achieving better accuracy, sensitivity, and specificity. Random Forest, LightGBM, and XGBoost perform well in identifying complex malicious URLs, while logistic regression provides identification, performance and integration of all individual classifiers in the Ensembler Model. CNN has a good effect on the connection model. The integrated model increases the reliability of malicious URL detection by demonstrating the ability to protect against changing threats. This research highlights the importance of an integrated approach to cybersecurity measures, providing insight into effective strategies to reduce cyber risks.

According to the diagram. As shown in Figures 2 and 3, the accuracy of the model is 92.5% .Additionally, precision is 89%.

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Fig.1. Confusion matrix

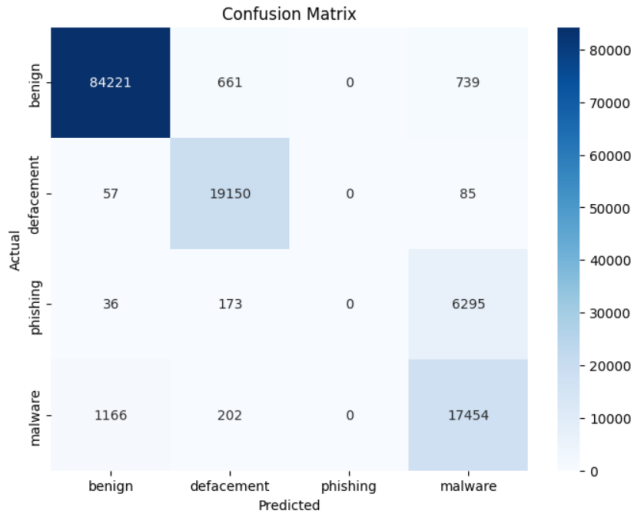

Fig 2. Confusion Matrix of Model

Ensemble Model Accuracy: 0.9245924799791153

	precision	recall	f1-score	support
benign	0.98	0.98	0.98	85621
defacement	0.94	0.98	0.96	19292
phishing	0.00	0.00	0.00	6504
malware	0.71	0.93	0.81	18822
accuracy			0.92	130239
macro avg	0.66	0.72	0.69	130239
weighted avg	0.89	0.92	0.90	130239

Fig 3. Performance Metrics of Model

CONCLUSION

This project addressed the pressing issue of malicious URL detection through the application of machine learning methodologies. The primary aim was to devise a robust system capable of accurately discerning various forms of malicious URLs, encompassing defacement, phishing, and malware. To accomplish this objective, an ensemble of machine learning algorithms, including Random Forest, XGBoost, Logistic Regression, convolutional neural network (CNN) model and LightGBM alongside a Ensemble model was deployed. These models underwent training on a dataset comprising 22 lexical features extracted from raw URLs. The experimental outcomes underscored the efficacy of the ensemble strategy in augmenting detection performance. While individual models exhibited commendable predictive capabilities, the ensemble model excelled, achieving an impressive accuracy rate of 92.5%. Furthermore, a feature importance analysis unveiled critical insights into malicious URL traits. Features such as hostname length, directory count, and the presence of suspicious terms emerged as pivotal indicators for detection. Notwithstanding the promising outcomes, certain limitations warrant consideration. The dataset employed for training and evaluation may not fully encapsulate the breadth and intricacy of real-world malicious URLs. Additionally, model performance may fluctuate contingent upon URL characteristics and evolving cyber threats. Future research avenues may entail exploring additional feature engineering techniques, harnessing advanced deep learning architectures, or integrating dynamic analysis methods to further enhance detection capabilities. Moreover, concerted efforts to amass larger, more diverse datasets will be pivotal for bolstering model robustness and generalization.

In conclusion, this project underscores the potential of machine learning in combating cyber threats and underscores the imperative for ongoing research and development. By leveraging advanced methodologies and fostering interdisciplinary collaboration, strides can be made in fortifying cybersecurity and safeguarding the digital landscape against malicious incursions.

REFERENCES

1. M. A. Waheed, B. Gadgay, S. DC, V. P and Q. U. Ain, "A Machine Learning approach for Detecting Malicious URL using different algorithms and NLP techniques," 2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon), Vijaypur, India, 2022, pp. 1-5, doi: 10.1109/NKCon56289.2022.10126798.
2. R. Chiramdasu, G. Srivastava, S. Bhattacharya, P. K. Reddy and T. Reddy Gadekallu, "Malicious URL Detection using Logistic Regression," 2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS), Barcelona, Spain, 2021, pp. 1-6, doi: 10.1109/COINS51742.2021.9524269.
3. P. Rastogi, E. Singh, V. Malik, A. Gupta and S. Vijh, "Detection of Malicious Cyber Fraud using Machine Learning Techniques," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 520-524, doi: 10.1109/Confluence52989.2022.9734181.
4. A. Saxena, A. Arora, S. Saxena and A. Kumar, "Detection of web attacks using machine learning based URL classification techniques," 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-13, doi: 10.1109/CONIT55038.2022.9847838.
5. B. Yogesh and G. S. Reddy, "Detection of Malware in the Network Using Machine Learning Techniques," 2022 International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC), Hyderabad, India, 2022, pp.204-211, doi: 10.1109/ICMACC54824.2022.10093525.

6. T. Kavitha, S. Hemalatha, R. Mounica, V. Niveda and Y. Kumar, "A Visionary Approach to Detect Spoofing Website using Machine Learning Algorithms," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-6, doi: 10.1109/ICCCI56745.2023.10128396.
7. C. C, P. K. Pareek, V. H. Costa de Albuquerque, A. Khanna and D. Gupta, "Improved Domain Generation Algorithm To Detect Cyber-Attack With Deep Learning Techniques," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-8, doi: 10.1109/MysuruCon55714.2022.9972526.
8. S. R. A, M. R, R. N, S. L and A. N, "Survey on Malicious URL Detection Techniques," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 778-781, doi: 10.1109/ICOEI53556.2022.9777221.
9. M. Mehndiratta, N. Jain, A. Malhotra, I. Gupta and R. Narula, "Malicious URL: Analysis and Detection using Machine Learning," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 1461-1465.
10. Y. -C. Chen, Y. -W. Ma and J. -L. Chen, "Intelligent Malicious URL Detection with Feature Analysis," 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 2020, pp. 1-5, doi: 10.1109/ISCC50000.2020.9219637.
11. Shantanu, B. Janet and R. Joshua Arul Kumar, "Malicious URL Detection: A Comparative Study," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 1147-1151, doi: 10.1109/ICAIS50930.2021.9396014.
12. A. Mantravadi, K. Indurkhya, S. M. Buchi, S. Y. Ganda, S. K. Kumar Reddy and A. R. Kaveeshwar, "A new method for the Detection of Anomalies in Streaming Data Malicious URL Segmentation and Classification using Machine Learning Techniques," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2022, pp. 258-262, doi: 10.1109/ICAISS55157.2022.10011007