IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 6, June 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106062 |

The Phishing Sleuth Using Machine Learning

Chiranjeevi Chandrasekhar K[1], Dr. Prabhudeva S[2]

PG Student, Dept. of Master of Computer Applications, Jawaharlal Nehru New College of Engineering, Shivamogga, India

Director and professor, Dept. of Master of Computer Applications, Jawaharlal Nehru New College of Engineering, Shivamogga, India

ABSTRACT: The internet has integrated seamlessly into our daily lives, but it has also made it possible to carry out harmful acts like phishing invisibly. Phishers attempt to target their victims by employing social engineering techniques or building fake websites in order to obtain information such as account IDs, usernames, and passwords from people and businesses. Although several strategies have been put out to identify phishing websites, phishers have developed ways to circumvent these strategies. Machine learning is one of the best techniques for spotting these dangerous behaviours. This paper presents a comprehensive study on phishing URL detection using machine learning techniques. The research compares the performance of various algorithms in identifying phishing URLs using a dataset containing legitimate and phishing URLs. The study employs evaluation metrics such as accuracy, precision, recall and F1 score. The experiments include popular algorithms like logistic regression, support vector machines, random forests, naive bayes classifier, k-nearest neighbors, gradient boosting classifier, catboost classifier, decision tree, xgboost classifier and multi-layer perceptron. Overall, this research offers insightful information about the use of machine learning methods for phishing URL detection. Researchers and cyber security experts may use the information offered here to create stronger, more effective defences against phishing attempts and systems to shield people and businesses from online dangers.

KEYWORDS: Website Classification, Machine learning, Spam, Phishing URLs, Cyber Security.

I. INTRODUCTION

Having a computer and internet connection facilitates both our personal and professional life in numerous ways easier. It makes it possible for us to conduct business and carry out activities in a variety of sectors, including commerce, health, education, research, engineering, entertainment, and public services. With the advent of wireless and mobile technology, individuals that need access to a local network may now connect to the Internet whenever and wherever they are assaults can be carried out by pirates, hacktivists, non-malicious (capped) assaults, cybercriminals, and others.

The Morris Worm began operating in 1988 and has continued to this day. The object information in the computer assaults the data it contains. Here are only a few instances include fraud, forgeries, coercion, shakedowns, and hacking. Illegal digital content has serious issues, such as malware programmes and illegitimate digital content. People using the internet must thus take steps to protect themselves from possible online threats in addition to social engineering Attackers intend to contact a high number of target users in order to collect a lot of data or money. The average cost of an assault in 2019 is between $ 108K and $ 1.4 billion, according to Kaspersky statistics [10]. Additionally, 124 billion dollars are spent globally on security-related goods and services. Attacks that are classified as "phishing attacks" are the most prevalent and harmful ones. In this kind of attack, cybercriminals frequently use social networking sites or email as their communication medium. Attackers deceive users into believing the message came from a reliable source, such a bank, an e-commerce site, or something similar. Make an effort to obtain private information as a consequence.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106062 |

II. RELATED WORK

Here we have selected few key literatures after exhaustive literature survey and listed as below:

J. Shad et al, [1] proposed a machine learning-based approach for detecting phishing websites using various features and attributes, including URL properties, HTML content, visual elements, and page structure. This robust feature set is used to train a machine learning model in the titled paper- "A Novel Machine Learning Approach to Detect Phishing Websites"

Bireswar Banik et al [2], in their study discuss a phishing URL detection system using SVM, addressing the growing threat of attacks. The system's performance is compared with various kernel functions, demonstrating its robustness in distinguishing legitimate URLs.

S. Sheng et al [3], examines the use of phishing blacklists as a protective measure in email systems, web browsers, and security applications. It compares a large dataset of phishing URLs against blacklist entries, assessing their performance through precision, recall, and false positive rate.

Jayveer Singh et al [4], this survey explores machine learning techniques in intrusion detection systems, examining algorithms, effectiveness, and emerging trends like deep learning, feature selection, and hybrid techniques.

M. Khonji et al [5], made a literature survey on phishing detection aims to analyze techniques and approaches, addressing challenges like zero-day detection, adaptability to evolving techniques, and real-time response.

W. Fadhilel et al [6], worked on feature selection techniques for predicting phishing websites, emphasizing the importance of selecting relevant, discriminative features to differentiate legitimate and phishing websites.

M. Karabakh et al [7], focuses on evaluating the performance of classifiers on a reduced Forensics Secure Website dataset. The study aims to compare the effectiveness of different classifiers in detecting and classifying secure and insecure websites in the field of digital forensics.

S. Parekh et al [8], the publication titled "A New Method for Detection of Phishing Websites: URL Detection" Researchers develop a new method for detecting phishing websites using URL detection, focusing on URL properties to prevent attacks and extract sensitive information.

K. Shima et al [9], introduced a Bag of Bytes approach for efficiently classifying URL bit streams in cloud computing, internet, and networks, leveraging characteristics of URL streams.

T. Eisenberg et al [10], The paper discusses the findings and recommendations of the "Cornell Commission," which was a group formed at Cornell University to investigate the Morris worm and its impact. The Morris worm, created by Robert Tappan Morris, had caused significant disruption and highlighted the vulnerabilities in computer systems at the time.

III. PROBLEM STATEMENT

Drawbacks of Traditional Methods Human-based detection methods are slow, expensive, and prone to error. Signature-based methods, adversarial attack, limited and imbalanced data, limited adaptability, dependency of human expertise, can only detect known phishing URLs and fail to identify new ones. This is where Machine Learning algorithms come in, providing a faster and more accurate detection of Phishing URLs.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106062 |

Taking phishing detection to the next level machine learning algorithms offers a smarter way to detect phishing URLs. By analyzing patterns and behaviours, these algorithms can distinguish between legitimate and suspicious URLs.

IV. DESIGN AND IMPLEMENTATION

The collection of data and selection of the most crucial attributes is the first step in the system's operation. The necessary data is pre-processed into the necessary format. After that, the data is split into training and testing data. The machine learning techniques are employed and further the training data is used to train the model. By testing the system with test data, the correctness of the system is determined. Using the following modules, this system is put into action.

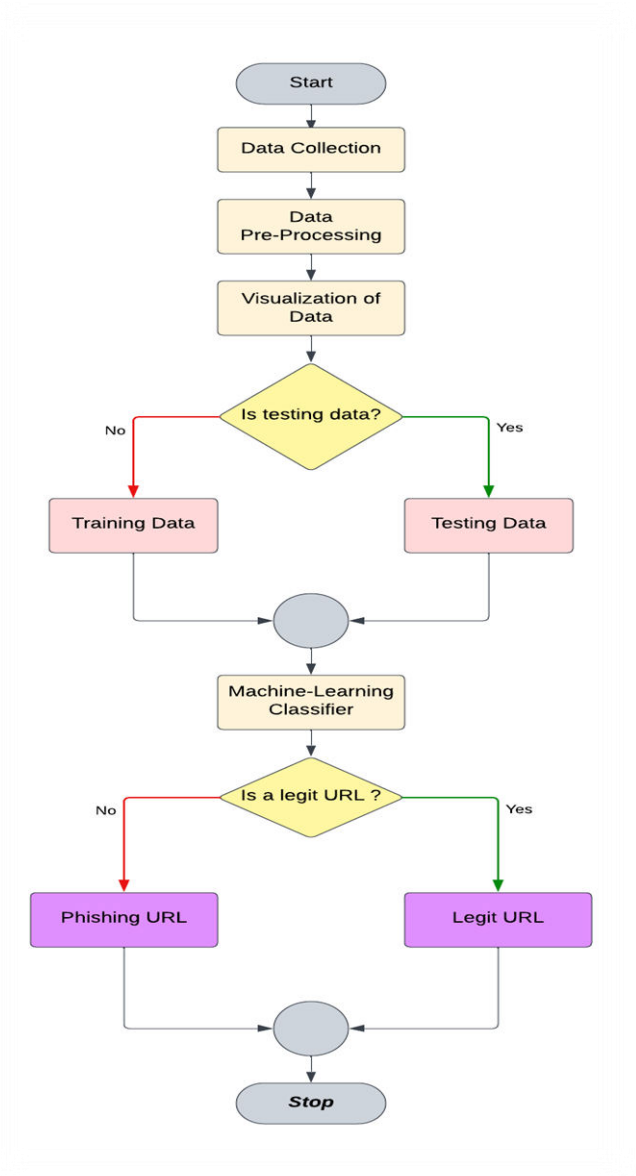

Figure 1: Flow chart of the system

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106062 |

The Fig.1 shows the flowchart of the Phishing URL detection using machine learning techniques. Data required for the prediction is collected using open resources.

Data collection: It is a crucial phase since the quality and volume of the data we collect for the suggested system will directly affect how well the predictive model can perform. The phishing-website-detector dataset is taken from Kaggle (https://www.kaggle.com/eswarchandt/). There is a list of more than 11000 websites' URLs. Each sample comprises 30 website parameters and a class label designating whether it is a phishing website or not (either 1 or -1). This dataset's summary states that it has 11054 samples and 32 features.

Data visualization: The depiction of data or information in a visual or graphical format is referred to as data visualisation. It entails producing visual graphics or visuals to present complex facts or data in a clear and understandable manner. Data of many kinds, including monetary, textual, spatial, and temporal data, can be represented visually. By presenting data in an easily interpreted and analysed format, visualisation seeks to increase data accessibility and comprehension.

Splitting of data: The act of dividing a dataset into training and testing is known as data splitting. Training receives 80% of the data, while testing receives the remaining 20%. The performance of the model is assessed using the ML method by testing the data. The best model is chosen based on the testing and training data. Data used for testing and training are not the same.

Model building & training: Building and training models is one of the most successful and widely applied forms of machine learning is supervised machine learning. When we need to predict a specific outcome or label from a set of supplied features and we have instances of feature-label pairs, we utilise supervised learning. These features-label pairings, which make up our training set, are used to create a machine learning model. For fresh, unheard-of data, our objective is to create precise forecasts. The terms "classification" and "regression" refer to the two main categories of supervised machine learning issues. Since the detection of phishing urls is a continuous number, or in programming language, a floating-point number, our data set falls within the category of regression problem. Accuracy & F1 score are the measures used to assess the model's performance.

Result Analysis

	ML Model	Accuracy	f1_score	Recall	Precision
0	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
1	CatBoost Classifier	0.972	0.975	0.994	0.989
2	Random Forest	0.966	0.969	0.996	0.987
3	Multi-layer Perceptron	0.965	0.969	0.988	0.986
4	Support Vector Machine	0.964	0.968	0.980	0.965
5	Decision Tree	0.957	0.962	0.991	0.993
6	K-Nearest Neighbors	0.956	0.961	0.991	0.989
7	Logistic Regression	0.934	0.941	0.943	0.927
8	Naive Bayes Classifier	0.605	0.454	0.292	0.997

Fig 2: Results of Trained Models

In the above Figure 2 it shows the accuracy, precision and F1 score of all the machine learning models. In the above 9 ML models Gradient Boosting classifier as the highest accuracy and F1 score.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106062 |

V. RESULTS AND DISCUSSION

After using a machine learning approach for both training and testing, we discover that gradient boosting classifiers have higher accuracy than other approaches. The gradient boosting algorithm is a machine learning technique that combines a number of ineffective learning models to produce a powerful predictive model. Decision trees are frequently used for gradient enhancement. In order to manage the bias variance trade-off, boosting techniques are necessary. Boosting algorithms are thought to be more effective than bagging algorithms since they regulate both bias and variance in a model, as opposed to bagging algorithms, which solely control for high variance.

Table 1: Results of Gradient boosting classifier

Parameter	precision	recall	f1- score	support
-1	0.99	0.96	0.97	976
1	0.97	0.99	0.98	1235
Accuracy			0.97	2211
Macro avg	0.98		0.97	2211
Weighted avg	0.97		0.97	2211

The above table 1 shows the testing and trained values of gradient boosting algorithm, where result accuracy of trained data is 0.98, 0.97 is for testing data and f1 score of trained data is 0.99 and 0.97 for testing data.

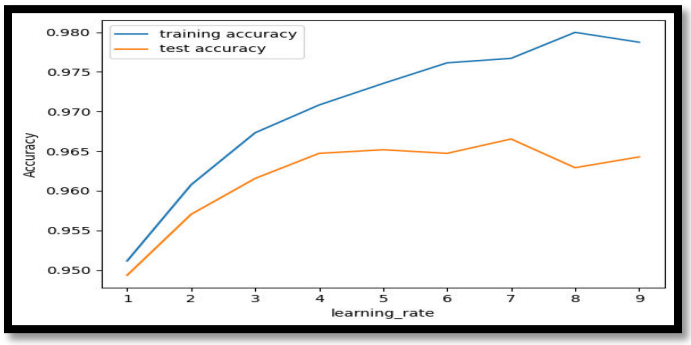

Figure 3: Testing and Training accuracy of Gradient boosting algorithm

The above fig 3 shows the graphical representation of trained data values of Gradient boosting algorithm.

Snapshots of User Interface

During the work we have designed the following web pages for supporting users to detect phishing url and displaying the message that is safe or unsafe. The following Fig 4 shows the results of Instagram which is a legitimate website with a 99% safe website notification**.**

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106062 |

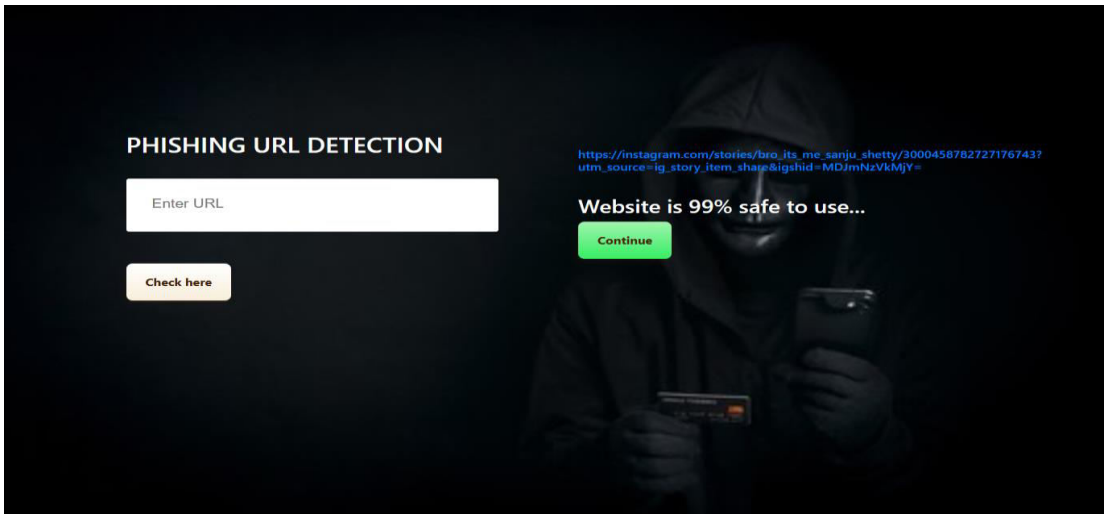

Figure 4: User Interface after predicting a safe or legitimate url

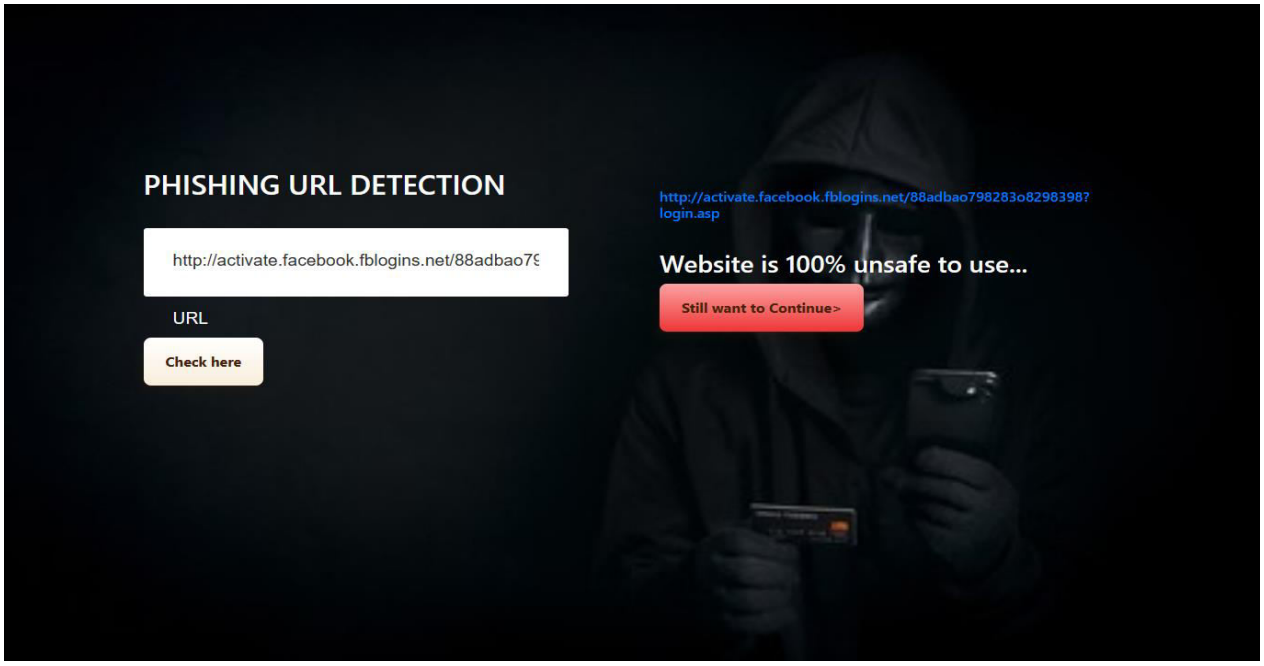

Figure 5: User Interface after predicting an unsafe or malicious url

The above fig 5 shows the results of a fake Facebook website with a 100% unsafe website notification.

VI. CONCLUSION AND FUTURE WORK

Machine learning algorithms have shown great promise in detecting phishing URLs, enabling efficient systems to prevent users from accessing malicious websites. These algorithms analyse large volumes of data, learn patterns,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106062 |

and make accurate predictions. Supervised learning algorithms, like random forests, support vector machines, and gradient boosting and other specified models can be trained on labelled datasets containing examples of legitimate and phishing URLs.

The effectiveness of machine learning-based phishing URL detection systems depends on the quality and diversity of training data. Regular updates and retraining are necessary to ensure their effectiveness against new and emerging threats. Combining machine learning approaches with other security measures, such as user education, multi-factor authentication, and real-time threat intelligence, is essential for comprehensive cyber security. In this paper, essentially the study is focussed on detection and classification of URLs.

REFERENCES

1. J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites ", Jaypee Institute of Information Technology, pp.425430, 2018.
2. Bireswar Banik, Abhijit sarma "Phishing URL detection system based on URL features using SVM,". International journal of electronics and applied Research vol.5, issue 2, Dec 2018.
3. 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.
4. Jayveer Singh and Manisha J Nene. "A survey on machine learning techniques for intrusion detection systems". International Journal of Advanced Research in Computer and Communication Engineering, 2013.
5. M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: a literature survey," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.
6. W. Fadhilel, M. Abusharkh, and I. Abdel-Qader, On Feature Selection for the Prediction of Phishing Websites, 2017 IEEE Securent Conf Dependable, AutonSecureur. Compute. 15th Intell Conf Pervasive Intell. Compute. 3rd Intl Conf Big Data Intell. Compute. Cyber Sci. Technol. Congr., pp. 871876, 2017.
7. M. Karabakh and T. Mustafa, Performance comparison of classifiers on reducing Forensics Secure website dataset, 6thInt. Symp. Digital Forensics Securer. ISDFS 2018 –Proceeding, vol.2018 Juana, pp. 15, 2018.
8. S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, A New Method for Detection of Phishing Websites: URL Detection, in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949952.
9. K. Shima et al., Classification of URL bit streams using bagi of bytes, in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 15.
10. T. Eisenberg, D. Gries, J. Hartmains, D. Holcomb, M. S. Lynn, T. Santoro et al, "The Cornell Commission: on Morris worm", Communications of ACM, vol. 32, issue 6, Jun 1989, pp 706-709.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details