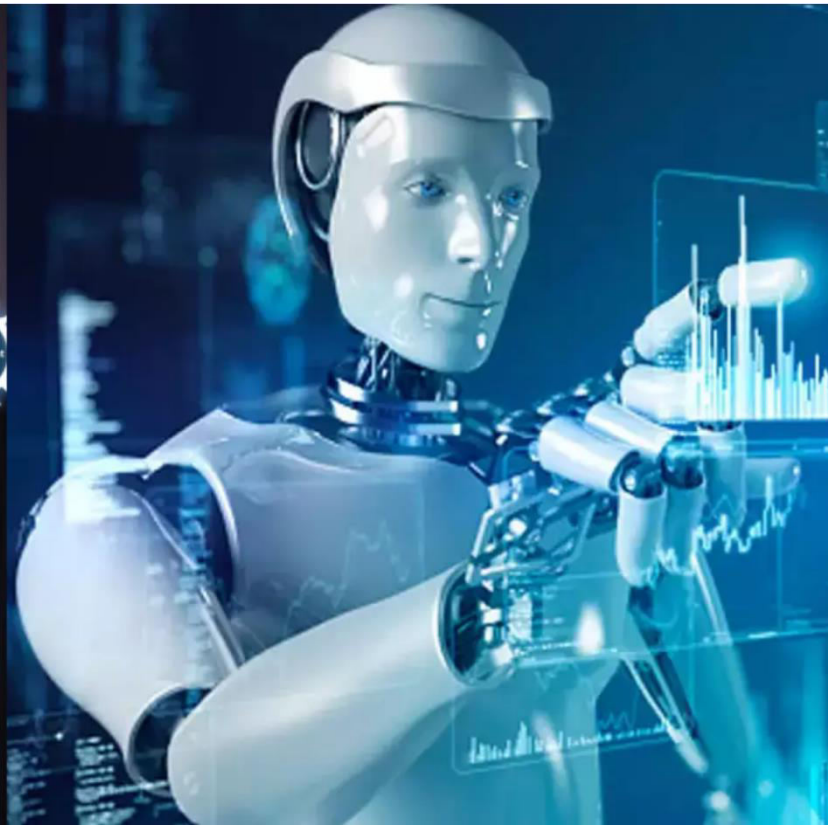
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404503|

Machine Learning-Based Fraudulent and Harmful Link Detection System

S. Sarjun Beevi, Venanka Sai Nithin, Kavati Venkatesh, Chintala Sai Rohit, Vardineni Shiva Teja

Associate Professor, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

B. Tech Students, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

ABSTRACT: Phishing sites which expects to take the victims confidential data by diverting them to surf a fake website page that resembles a honest to goodness one is another type of criminal acts through the internet and its one of the especially concerns toward numerous areas including e-managing an account and retailing. Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, neural system and data mining methods can be a successful mechanism in distinguishing phishing sites. We applied Random Forest (RF), one of the different types of machine learning based algorithms used for detection of Phishing websites. Finally we measured and compared the performance of the classifier in terms of accuracy.

KEYWORDS: Phishing Website Detection, Cybersecurity, Random Forest (RF), Machine Learning.

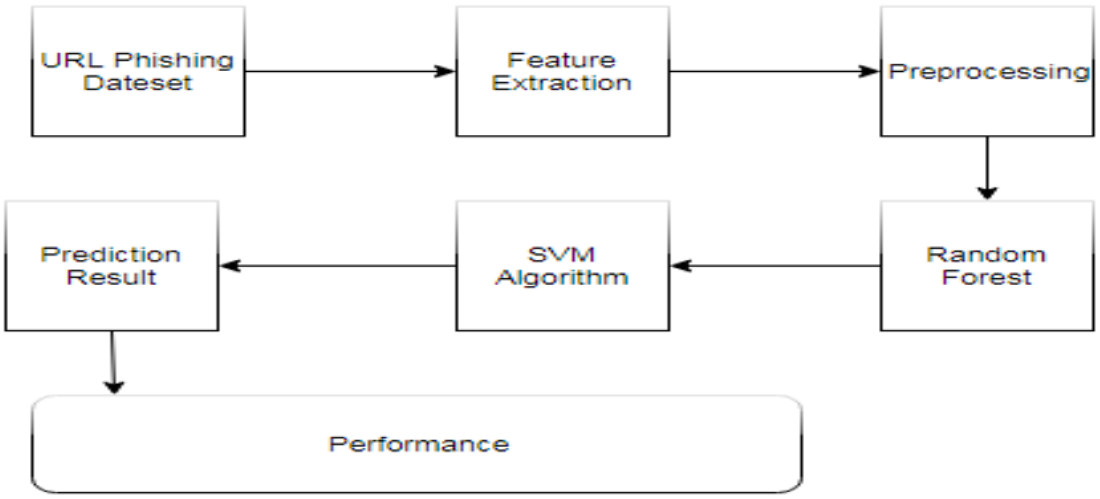


Figure 1: **SYSTEM ARICHITECTURE**

I. INTRODUCTION

Phishing is a type of extensive fraud that happens when a malicious website act like a real one keeping in mind that the end goal to obtain touchy data, for example, passwords, account points of interest, or MasterCard numbers.In spite of the fact that there are a few contrary to phishing programming and methods for distinguishing potential phishing endeavours in messages and identifying phishing substance on sites, phishes think of new and half breed strategies to go around the accessible programming and systems.Phishing is a trickery system that uses a blend of social designing

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404503|

what's more, innovation to assemble delicate and individual data, for example, passwords and charge card subtle elements by taking on the appearance of a dependable individual or business in an electronic correspondence. Phishing makes utilization of spoof messages that are made to look valid and implied to be originating from honest to goodness sources like money related foundations, ecommerce destinations and so forth, to draw clients to visit fake sites through joins gave in the phishing email. The misleading sites are intended to emulate the look of a genuine organization site page.

II. LITERATURE REVIEW

Phishing attacks have become one of the most prevalent threats in the domain of cybersecurity. These attacks typically involve tricking users into revealing sensitive information such as usernames, passwords, and credit card details through websites that closely mimic legitimate ones. As phishing techniques continue to evolve in sophistication, traditional blacklisting methods and rule-based systems have proven inadequate due to their inability to detect newly created or rapidly changing phishing sites.

Numerous studies have explored the application of **machine learning (ML) techniques** for phishing detection, owing to their capability to learn patterns and make decisions based on data rather than fixed rules. Among these, **Random Forest (RF)** has gained popularity due to its robustness, high accuracy, and ability to handle large datasets with numerous features. Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the mode of the classes for classification tasks.

III. METHODOLOGY

This project uses the Random Forest (RF) algorithm to detect phishing websites based on features extracted from URL, domain, and page content. A labeled dataset containing both phishing and legitimate websites is collected and preprocessed by handling missing values, encoding features, and splitting into training and testing sets. Feature selection is performed to improve accuracy and reduce complexity. The Random Forest model is then trained, which builds multiple decision trees and uses majority voting for classification. Finally, the model is evaluated using accuracy, precision, recall, F1-score, and a confusion matrix to assess its effectiveness in identifying phishing sites.

IV. IMPLEMENTATION

Data Set

- One of the challenges faced by our research was the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using data mining techniques have been disseminated these days, no reliable training dataset has been published publically, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features.
- In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we proposed some new features, experimentally assign new rules to some well-known features and update some other features.

Phishing Websites Features

- One of the challenges faced by our research was the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using data mining techniques have been disseminated these days, no reliable training dataset has been published publically, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404503|

- In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we proposed some new features, experimentally assign new rules to some well-known features and update some other features.

Detection Technique

Detection of phishing websites has received a lot of attention recently due to their impact on users' security. Therefore, many techniques have been developed to detect phishing websites varying from communication-oriented techniques, such as authentication protocols, blacklisting, and white-listing, to content-based filtering techniques. The blacklisting and white-listing techniques have not proven though to be sufficiently efficient when used in different domains, and thus they are not commonly used. Meanwhile, the content-based phishing filters have been widely used and have proven to be of high efficiency. In light of this, researches have focused on content-based mechanism and on developing machine learning and data mining techniques based on the header and body of emails.

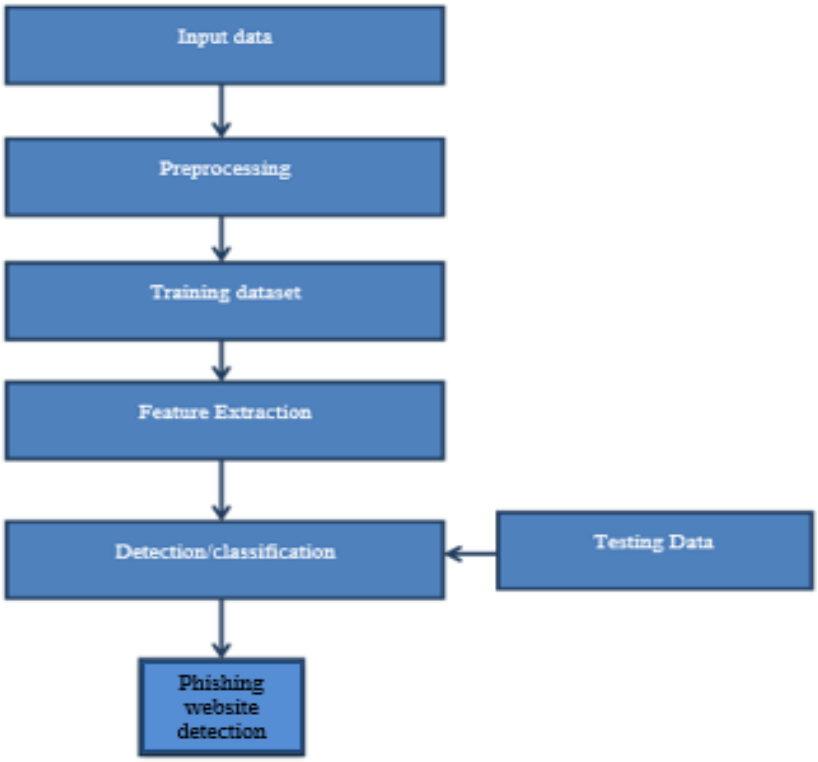


Figure 2: **Data Flow**

V. RESULTS AND CONCLUSION

Phishing is a cyber crime procedure utilizing both social building and specialized deception to take individual sensitive data. Besides, Phishing is considered as another extensive type of fraud. Experimentations against recent dependable phishing data sets utilizing different classification algorithm have been performed which received different learning methods. The base of the experiments is accuracy measure.The aim of this research work is to predict whether a given URL is phishing website or not. It turns out in the given experiment that Random forest based classifiers are the best classifier with great classification accuracy for the given dataset of phishing site. As a future work we might use this

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404503|

model to other Phishing dataset with larger size then now and then testing the performance of those classification algorithm's in terms of classification accuracy.

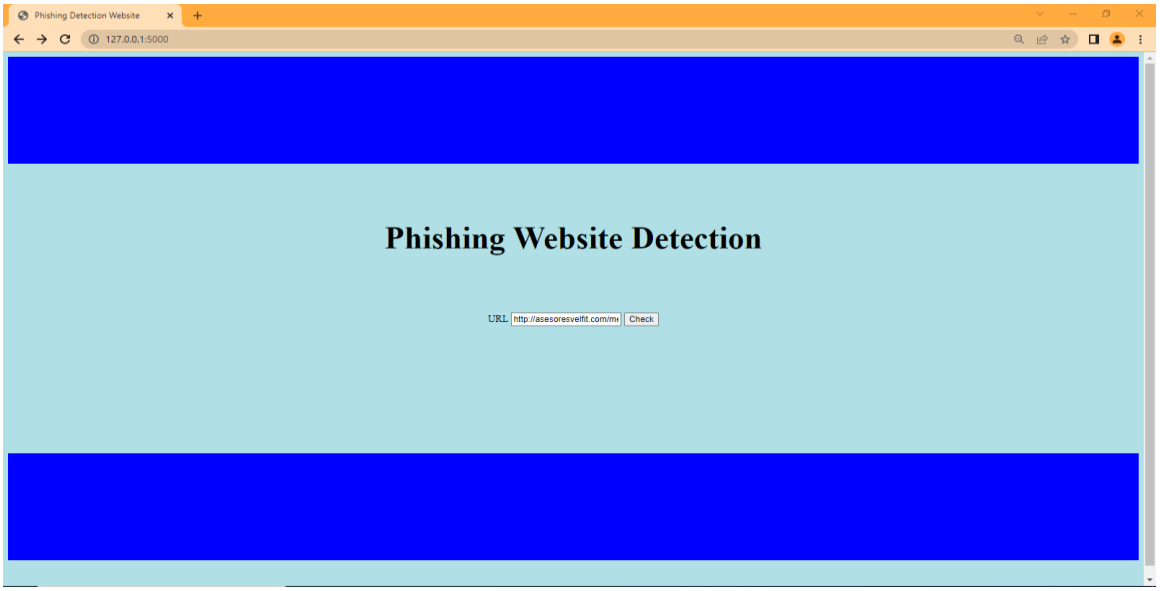
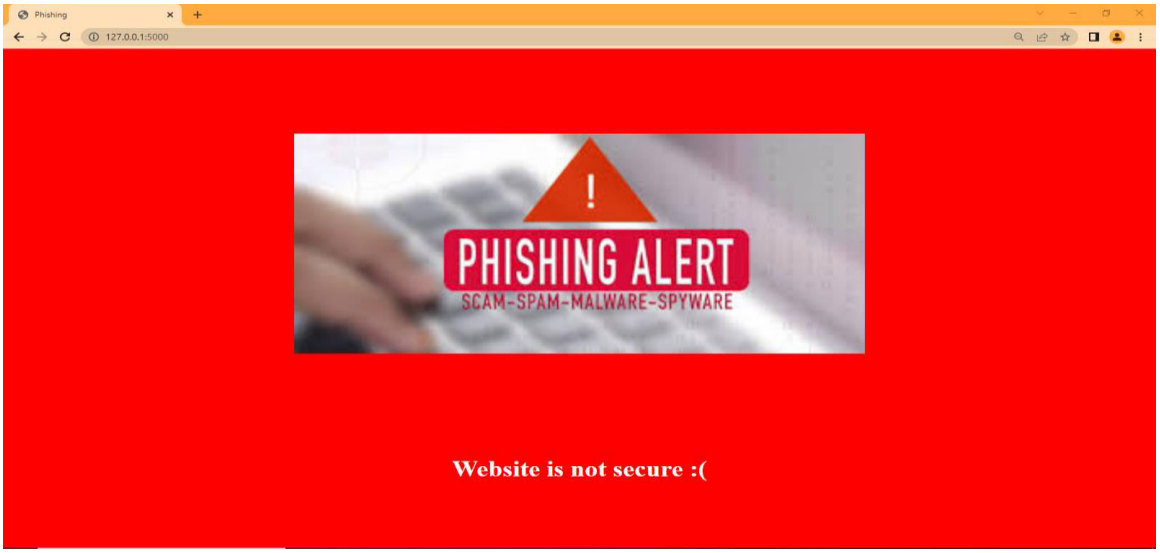


Fig:phishing Page



VI. FUTURE WORK

Future work will aim to develop a system that can learn by itself about new types of phishing attacks by adding a more enhanced feature to the detection process. The scope of this approach not only helps in adding more enhanced features but also updating the existing features to improve its importance level to make detection more efficient and reduce the false positive rate to a large extent. Another further work should include deploying this approach into a web extension to make the detection more robust to the user.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404503|

VII. FUTURE WORK

Future work will aim to develop a system that can learn by itself about new types of phishing attacks by adding a more enhanced feature to the detection process. The scope of this approach not only helps in adding more enhanced features but also updating the existing features to improve its importance level to make detection more efficient and reduce the false positive rate to a large extent. Another further work should include deploying this approach into a web extension to make the **detection more robust to the user.**

REFERENCES

[1] Samuel Marchal, Jérôme François, Radu State, and Thomas Engel, "PhishStorm: Detecting Phishing With Streaming Analytics," IEEE Transactions on Network and Service Management, vol. 11 , issue: 4 , pp. 458-471, December 2014.

[2] Mohammed Nazim Feroz,Susan Mengel, "Phishing URL Detection Using URL Ranking," IEEE International Congress on Big Data, July 2015 .

[3] Mahdieh Zabihimayvan, Derek Doran, "Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection," International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, June 2019 .

[4] Moitrayee Chatterjee,Akbar-Siami Namin, "Detecting Phishing Websites through Deep Reinforcement Learning," IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), July 2019.

[5] Chun-Ying Huang,Shang-Pin Ma,Wei-Lin Yeh,Chia-Yi Lin,ChienTsung Liu, "Mitigate web phishing using site signatures," TENCON 2010-2010 IEEE Region 10 Conference, January 2011 .

[6] Aaron Blum,Brad Wardman,Thamar Solorio,Gary Warner, "Lexical feature based phishing URL detection using online learning," 3rd ACM workshop on Artificial intelligence and security, Chicago, Illinois, USA, pp. 54-60, August 2010 .

[7] Mohammed Al-Janabi,Ed de Quincey,Peter Andras, "Using supervised machine learning algorithms to detect suspicious URLs in online social networks," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, pp. 1104-1111, July 2010 .

[8] Panyaram, S., & Kotte, K. R. (2025). Leveraging AI and Data Analytics for Sustainable Robotic Process Automation (RPA) in Media: Driving Innovation in Green Field Business Process. In Driving Business Success Through Eco-Friendly Strategies (pp. 249-262). IGI Global Scientific Publishing.

[8] Erzhou Zhu,Yuyang Chen,Chengcheng Ye,Xuejun Li,Feng Liu, "OFSNN:An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network," IEEE Access(Volume:7), pp. 73271-73284, June 2019 .

[9] Ankesh Anand,Kshitij Gorde,Joel Ruben Antony Moniz,Noseong Park,Tanmoy Chakraborty,Bei-Tseng Chu, "Phishing URL Detection with Oversampling based on Text Generative Adversarial Networks," IEEE International Conference on Big Data (Big Data), December 2018.

[10] Justin Ma,Lawrence K. Saul,Stefan Savage,Geoffrey M. Voelker, "Learning to detect malicious URLs," ACM Transactions on Intelligent Systems and Technology (TIST) archive Volume 2 Issue 3, April 2011.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details