The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 19(2): S.I (1), 760-766, 2024 www.thebioscan.com

# INTELLIGENT DETECTION DESIGNS OF HTML URL PHISHING ATTACKS

**Sk. Nishanth Anjum[1], Mohammad Rahmat Ali[2], Dr. DvssSubramanyam[3]**

[1]PG Student, Computer Science & Engineering ISL Engineering College, India

[2]Assistant Professor, Department of Computer Science & Engineering, ISL Engineering College, India.

[3] Professor & Vice Principal, Department of Computer Science (Keshav Memorial Engineering College)

**DOI: https://doi.org/10.63001/tbs.2024.v19.i02.S.I(1).pp760-766**

**KEYWORDS**
HTML,
URL,
Phishing attacks.

**Received on:**

**19-09-2024**

**Accepted on:**

**23-12-2024**

**ABSTRACT**

Phishing attacks are a type of cybercrime that has grown in recent years. It is part of social engineering attacks where an attacker deceives users by sending fake messages using social media platforms or emails. Phishing attacks steal users' information or download and install malicious software. They are hard to detect because attackers can design a phishing message that looks legitimate to a user. This message may contain a phishing URL so that even an expert can be a victim. This URL leads the victim to a fake website that steals information, such as login information, payment information, etc. Researchers and engineers work to develop methods to detect phishing attacks without the need for the eyes of experts. Even though many papers discuss HTML and URL-based phishing detection methods, there is no comprehensive survey to discuss these methods. Therefore, this paper comprehensively surveys HTML and URL phishing attacks and detection methods. We review the current state-of-art machine learning models to detect URL-based and hybrid-based phishing attacks in detail. We compare each model based on its data preprocessing, feature extraction, model design, and performance.

## INTRODUCTION

Phishing attacks employ social engineering to acquire personal and financial information from victims. Attackers might approach victims by sending fake emails or messages on social media sites like Twitter and Facebook. Users are susceptible when they enter information or download attachments. Since attackers may reach numerous individuals worldwide with a single message, social media platform attacks have increased in recent years. According to the Anti-Phishing Working Group (APWG), phishing assaults climbed by 250000 in January 2021. In the first quarter of 2021, business compromises rose 56% from the fourth quarter of 2020. Financial institutions, social media, and online emails are 2021's top targets. Base, criminals target financial industry and social media networks to steal victims' financial information or identities. Attackers may deliver malware, ransomware, or other intrusions. Recently increased phishing attempts and cybersecurity risks have become a major problem. Most companies identify these threats using human expertise. Phishing assaults are difficult to spot, even for experts, since authentic and phony communications look same. To identify phishing mails, cybersecurity professionals focus on message attachments like URLs or email IDs. However, attackers use new approaches to develop hard-to-detect phishing assaults. They create phishing URLs and webpages that look like https://www.facebook.com/, https://www.faceb00k.com/, https://www.facebook.edu/, etc. Identifying phishing URLs from benign URLs is crucial. In recent years, researchers have presented high-accuracy phishing detection techniques such blacklist conventional machine learning and Deep Learning. Below is a short examination of each solution. • Blacklists include URLs of potential phishing sites. System blocks all URLs or IPs in this list. However, this strategy has a major downside. A system cannot stop phishing attacks without a URL in the list. • Traditional machine learning methods identify phishing. Traditional machine-learning algorithms need human feature extraction. Extraction of characteristics involves human work and time. These functionalities use URLs. When attackers create new phishing URLs, feature analysis and extraction grow, increasing feature dimension. Even after evaluating vast sets of characteristics and high dimensions, new phishing URLs may target it. • A model collects text and picture information without human intervention, making DL phishing URL detection advantageous. The intelligent design of phishing assaults and the phishing website developed by new DL methods causes various issues. A model learns to recognize lengthy URLs. It cannot identify small URLs. DL needs a large dataset to train, test, and verify models. DL model complexity makes it expensive. Phishing attacks may be detected using URL-based, content-based, or hybrid data. URL-based approaches collect URL data without looking at the content, title, etc. URL-based detection may identify phishing messages without clicking on the URL and installing malware. However, collecting simply URL-based information removes important phishing site elements like page title and code. URL-based analysis of small URLs is difficult. Content-based methods take data from website pictures, JavaScript, text, HTML, etc. However, the content-based solution requires users or systems to access the site and extract material, which might lead to a malware download and installation. The hybrid-based content feature combines URL- and content-based capabilities.

## LITERATURE REVIEW

H. Ma, Y. Zuo, and T. Li., Intelligent Detection Designs of HTML URL Phishing Attacks, Phishing assaults are a kind of cybercrime that has proliferated in recent years. This is a sort of social engineering assault in which an assailant misleads consumers by disseminating fraudulent communications via social media

platforms or email. Phishing attacks expropriate users' information or deploy and install harmful software. They are difficult to identify since perpetrators might craft a phishing message that seems authentic to the victim. This message may include a phishing URL, rendering even an expert susceptible to victimization. This URL directs the victim to a fraudulent website that extracts sensitive information, including login credentials and payment details. Researchers and technologists endeavor to create techniques for identifying phishing assaults autonomously, without professional intervention. Despite several studies addressing HTML and URL-based phishing detection techniques, a full study on these approaches is lacking. This study thoroughly examines HTML and URL phishing attacks along with their detection methodologies. We examine the contemporary state-of-the-art machine learning algorithms for the detection of URL-based and hybrid phishing assaults in detail. We evaluate each model according to its data preparation, feature extraction, architecture, and performance.

S. Salloum, T. Gaber, S. Vadera, and K. Shaalan., Phishing email detection using natural language processing techniques: A literature survey. Phishing is the most common kind of cybercrime that persuades individuals to provide personal information, such as account IDs, passwords, and banking data. Emails, instant messaging, and telephone calls are often used to initiate such cyber-attacks. Notwithstanding the continual enhancement of strategies to prevent such cyber-attacks, the present results remain insufficient. Conversely, phishing emails have surged dramatically in recent years, indicating a need for more effective and sophisticated countermeasures. Various techniques have been developed to filter phishing emails; nonetheless, a comprehensive solution remains elusive. This survey is, to our knowledge, the first study concentrating on the use of Natural Language Processing (NLP) and Machine Learning (ML) methodologies for the detection of phishing emails. This paper analyzes the many advanced NLP techniques used to detect phishing emails at different phases of the assault, focusing on machine learning methodologies. These methodologies undergo a comparative evaluation and analysis. This provides an understanding of the issue, its immediate solutions, and anticipated future research trajectories.

L. Tang and Q. H. Mahmoud., A survey of machine learning-based solutions for phishing website detection. The advent of the Internet has heightened public awareness of network security. A safe network environment is fundamental for the swift and robust development of the Internet. Phishing is a critical category of cybercrime involving the deceitful manipulation of users into clicking on fraudulent links, so compromising their information, which is then used to impersonate them and illicitly access their accounts for financial gain. Network security is a cyclical challenge of offense and protection. The techniques of phishing and the technologies for phishing detection are always evolving. Conventional techniques for detecting phishing links depend on blacklists and whitelists, which are ineffective in recognizing novel phishing URLs. Consequently, we must determine ways to forecast whether a developing link is a phishing website and enhance the precision of the prediction. As machine learning technology has matured, prediction has emerged as a crucial capability. This study presents a comprehensive assessment of advanced methodologies for detecting phishing websites. The discussion begins with the life cycle of phishing, presents prevalent anti-phishing techniques, primarily emphasizes the identification of phishing links, and provides a comprehensive analysis of machine learning-based solutions, including data gathering, feature extraction, modeling, and performance assessment. This study presents a comprehensive analysis of several strategies for detecting phishing websites.

A. Odeh, I. Keshta, and E. Abdelfattah. Machine Learning Techniques for detection of website phishing: A review for promises and challenges. Phishing websites constitute a cyber-attack aimed at internet users to illicitly acquire their sensitive information, including login passwords and financial data. Malefactors deceive visitors by displaying a counterfeit website as authentic or credible to get their sensitive information. Various solutions to phishing website assaults have been suggested, including heuristics, blacklists or whitelists, and Machine Learning (ML) approaches. This study delineates the cutting-edge methodologies for detecting phishing websites using machine learning approaches. This study proposes answers to the website's phishing issue via machine learning approaches. The most methodologies analyzed concentrate on conventional machine learning techniques. The literature examines the robust machine learning approaches of Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and Ada Boosting. This survey report reveals deep learning algorithms that outperform traditional machine learning methods in identifying phishing websites. This article identifies challenges to machine learning approaches, including overfitting, poor precision, and ineffectiveness due to insufficient training data. This study indicates that Internet users must be aware of phishing to prevent cyber-attacks. This study further highlights the suggestion for an automated system to combat phishing websites.

S. Salloum, T. Gaber, S. Vadera, and K. Shaalan. A systematic literature review on phishing email detection using natural language processing techniques, Annually, phishing incurs damages up to billions of dollars and poses a significant danger to the Internet economy. Phishing attacks are mostly executed via email. Numerous review studies have been conducted to enhance understanding of the current research trends in phishing email detection. Nevertheless, it is crucial to evaluate this matter from several viewpoints. None of the studies have thoroughly examined the use of Natural Language Processing (NLP) methods for phishing detection, save for one that highlighted the use of NLP techniques for classification and training, while considering a few alternatives. This work seeks to thoroughly analyze and synthesize studies on the use of NLP for the detection of phishing emails. A total of 100 research publications published between 2006 and 2022 were discovered and analyzed based on particular predetermined criteria. We examine the principal research domains in phishing email detection using NLP, the machine learning algorithms used in phishing email identification, the textual properties present in phishing emails, the datasets and resources utilized in phishing research, and the evaluation metrics. The results indicate that the primary focus of phishing detection research is feature extraction and selection, followed by classification methods and optimization techniques for detecting phishing emails. Support vector machines (SVMs) are extensively used among various classification techniques for the detection of phishing emails. The predominant NLP approaches identified are TF-IDF and word embeddings. Moreover, the Nazario phishing corpus is the most often used dataset for evaluating phishing email detection techniques. Furthermore, Python is the most prevalently used language for phishing email detection. The results of this research are anticipated to benefit the scientific community, particularly in the domain of NLP applications for cybersecurity challenges. This survey is distinctive in that it connects works to their publicly accessible tools and resources. The examination of the submitted studies indicated that little research has been conducted on Arabic language phishing emails using NLP approaches. Consequently, several unresolved challenges pertain to the detection of Arabic phishing emails.

**PROPOSED SYSTEM**

In order to strengthen the capabilities of phishing detection, new ways are continuously being presented. The use of hybrid models that combine approaches from machine learning and feature engineering indicates that they have potential. By way of illustration, a more thorough knowledge of phishing characteristics may be achieved by merging lexical and host-based information with machine learning techniques. It is the goal of feature-rich models to capture a wider range of harmful activities. These models include lexical, content, and host-based information. Furthermore, developments in natural language processing methods and machine learning models give prospects for enhanced detection by interpreting complex patterns in phishing communications. This may be accomplished at several levels. These suggested algorithms make use of a variety of tactics in order to improve accuracy and resistance in the face of advanced phishing strategies.

**METHODOLOGIES**

**MODULES EXPLANATION:**

**Data Set:**

A structured set of data is called a dataset. It is often arranged in rows and columns, with each row denoting a distinct observation or instance and each column denoting a particular characteristic or feature of that instance. Spreadsheets, databases, text files, and unique formats for particular purposes are just a few of the many formats that datasets may take.

**Data Cleaning:**

Finding and fixing mistakes or inconsistencies in a dataset to increase its quality and dependability for analysis is known as data cleaning. It entails duties include dealing with outliers, resolving missing data, eliminating duplication, fixing errors, and standardizing formats.

**NLP Feature Extraction:**

Natural Language Processing (NLP) feature extraction involves transforming text data into numerical or categorical features that can be used for machine learning tasks.

**ML Model:**

Machine learning models are computer algorithms that discern patterns and correlations within data to generate predictions or make judgments. They include diverse approaches like as regression, classification, clustering, and deep learning, and are trained on labeled or unlabeled datasets to generalize patterns and address particular tasks, facilitating automated decision-making across several domains.

**Train Model:**

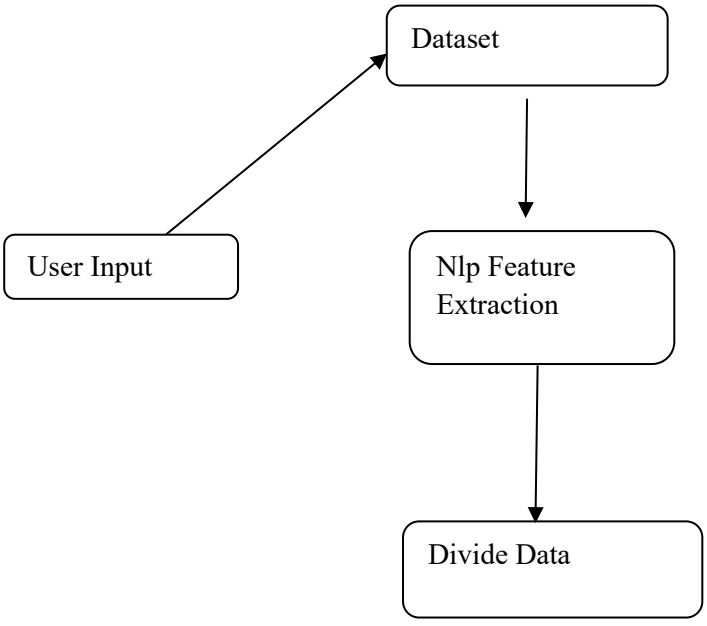
Training a model entails providing it with a dataset to discern patterns and correlations, while fine-tuning its parameters using optimization methods like gradient descent to reduce prediction errors. The model's performance is then assessed using a distinct validation dataset to determine its efficacy in generating correct predictions or classifications.

**Test and Deployement:**

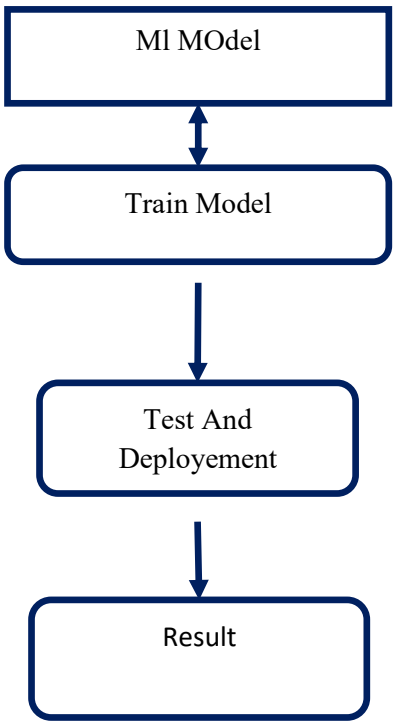
Evaluating a model entails examining its efficacy on novel data to confirm its generalizability and adherence to specified accuracy standards. Deployment entails the integration of the trained model into production systems, facilitating real-time predictions, and monitoring its performance for continuous optimization and maintenance.

**DATA FLOW DIAGRAM**

**Level 0**



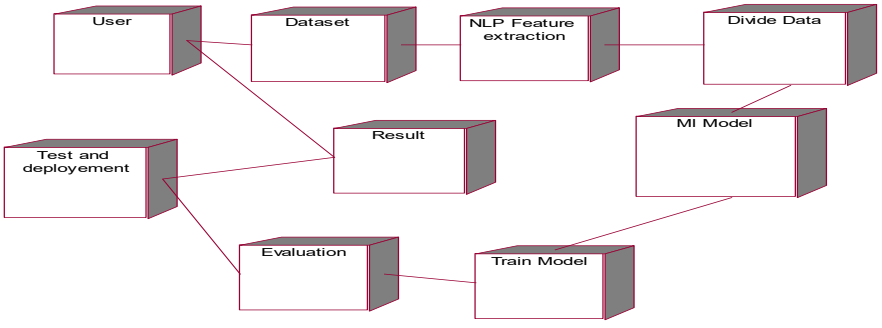
**Level 1**




**EXPLANATION:**
A graphical depiction of the "flow" of data through an information system, which models the many parts of the system's activity, is referred to as a data flow diagram (DFD). They are often a first phase that is used to develop an overview of the system, which can then be built upon at a later time. DFDs may also be used for the purpose of visualizing data processing, which is referred to as structured design.[28]

A data flow diagram (DFD) illustrates the types of data that will be input into and output from the system, as well as the originating and destination points of the data, as well as the storage locations for the data. It does not provide any information on the timing of processes, nor does it provide any information regarding whether processes will follow a sequential or parallel operation.

**DEPLOYMENT DIAGRAM**




**EXPLANATION:**
A Deployment Diagram delineates the actual hardware upon which the software system will operate. It also dictates the deployment of the program on the underlying hardware. It links software components of a system with the devices designated to execute them.

**1. SYSTEM ARCHITECTURE:**

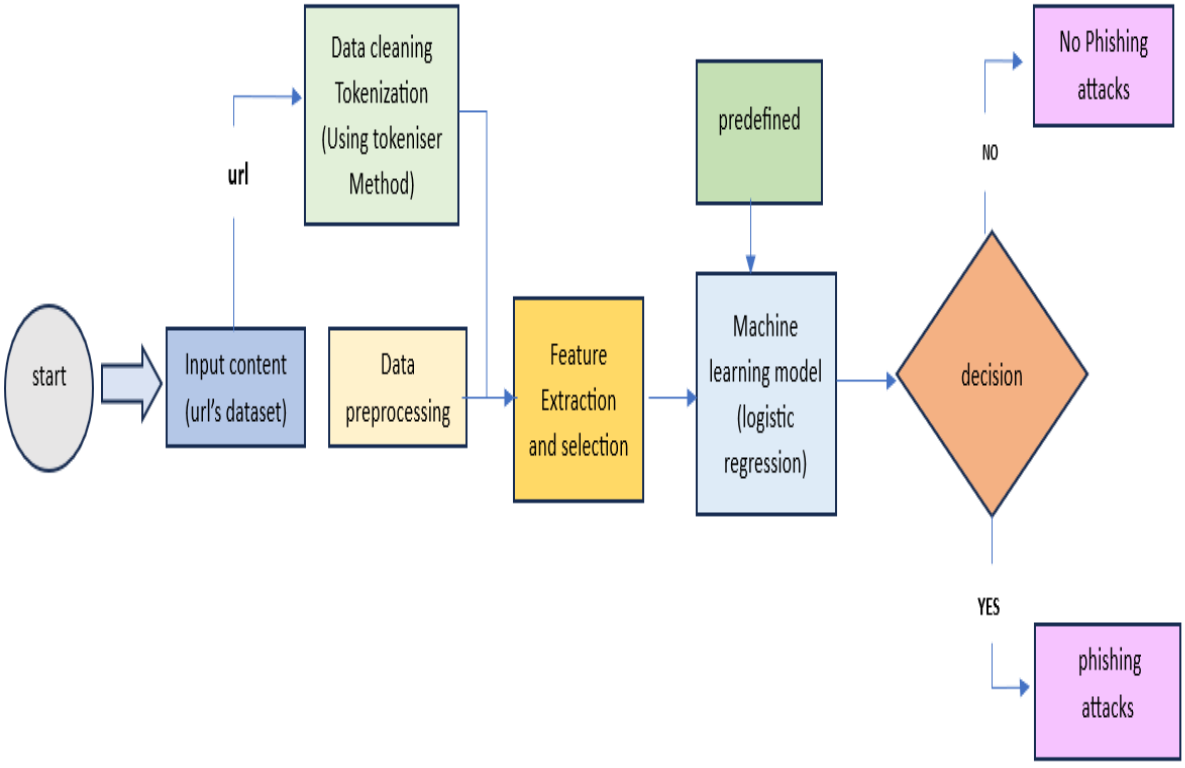



Fig 1: System Architecture

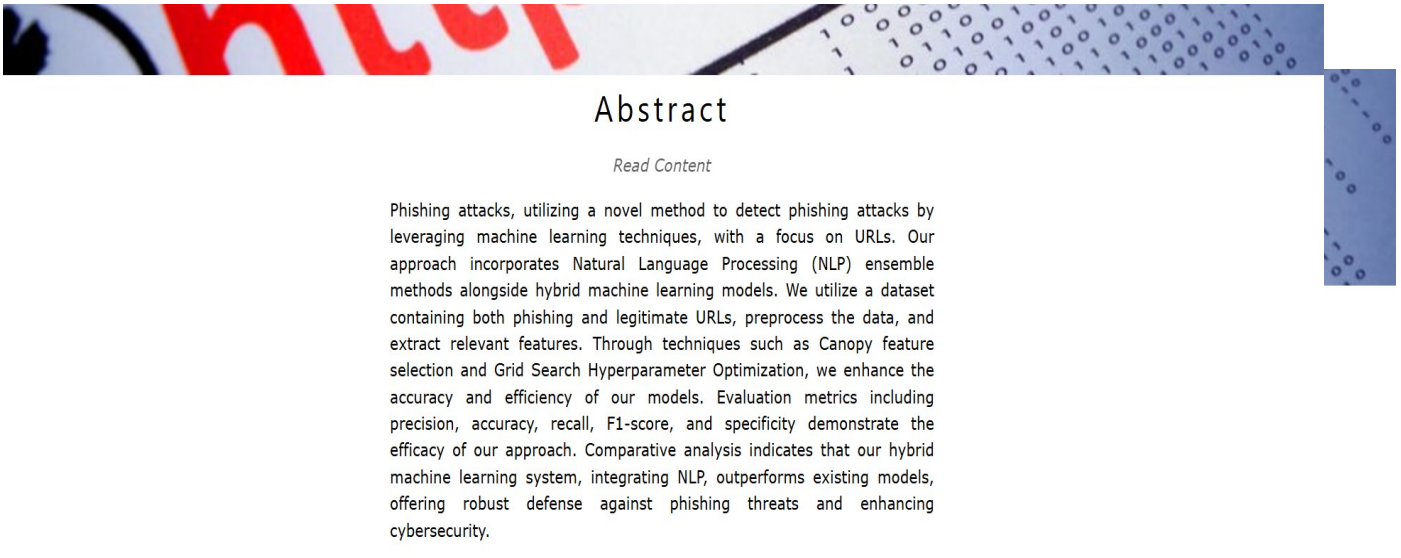
## 2. SNAPSHOTS

**General:**
This project is implements like application using python and the Server process is maintained using the SOCKET & SERVERSOCKET and the Design part is played by Cascading Style Sheet.

**SNAPSHOTS**
**HOME PAGE**


PHISHING DETECTION SYSTEM THROUGH HYBRID MACHINE LEARNING BASED ON URL

**ABSTRACT PAGE**




**LOGIN PAGE**

**OUTPUT PAGE**


**SOFTWARE TESTING**

The objective of testing is to identify mistakes. Testing is the procedure of identifying all potential flaws or deficiencies in a work output. It offers a method to evaluate the functioning of components, subassemblies, assemblies, and/or a completed product. The process of testing software is to verify that the software system fulfills its requirements and user expectations while avoiding undesirable failures. There are many categories of tests. Each test category fulfills a distinct testing need.

**FUTURE ENHANCEMENTS:**

Feature extraction involves the model reducing the quantity of input characteristics and using the most relevant training features. This stage distinguishes between deep learning and machine learning. Machine learning requires human intervention to extract characteristics from the incoming data. Nonetheless, deep learning emphasizes learning from the input and its corresponding label to extract the most relevant information. The predominant algorithm for feature extraction in recent years is CNN. The authors recommend using CNN for feature extraction.

## CONCLUSION

In recent years, deep learning has proven indispensable for addressing cybersecurity challenges, including phishing assaults. It is owing to its capacity to automatically extract features from the input data rather than manually. This study examines cutting-edge deep learning models for the detection of phishing assaults. The significance is in evaluating each deep learning model at every stage, from input data to model output. The significance of data preparation is equivalent to that of the deep learning model. input preparation significantly influences model performance in

any job, particularly when the model is used to identify real-time input. The model must categorize all incoming data, regardless of its absence in the model's dataset. Consequently, this study focuses on data preparation, emphasizing its merits and drawbacks. Subsequently, we evaluate the design of each deep learning model, emphasizing its advantages and disadvantages.

## REFERENCES


- Y. Zhang, Y. Xiao, K. Ghaboosi, J. Zhang, and H. Deng, ''A survey of cyber crimes,'' Secur. Commun. Netw., vol. 5, no. 4, pp. 422–437, 2012.
- APWG Developers. (2021). Phishing Activity Trends Report.
- M. Lei, Y. Xiao, S. V. Vrbsky, and C.-C. Li, ''Virtual password using random linear functions for on-line services, ATM machines, and pervasive computing,'' Comput. Commun., vol. 31, no. 18, pp. 4367–4375, Dec. 2008.
- P. Burda, L. Allodi, and N. Zannone, ''Don't forget the human: A crowdsourced approach to automate response and containment against spear phishing attacks,'' in Proc. IEEE Eur. Symp. Secur. Privacy Workshops (EuroS PW), Sep. 2020, pp. 471–476.
- P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, ''PhishNet: Predictive blacklisting to detect phishing attacks,'' in Proc. IEEE INFOCOM, Mar. 2010, pp. 1–5.
- W. Zhang, Y.-X. Ding, Y. Tang, and B. Zhao, ''Malicious web page detection based on on-line learning algorithm,'' in Proc. Int. Conf. Mach. Learn. Cybern., vol. 4, Jul. 2011, pp. 1914–1919.
- A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. González, ''Classifying phishing URLs using recurrent neural networks,'' in Proc. APWG Symp. Electron. Crime Res. (eCrime), 2017, pp. 1–8.
- B. Cui, S. He, X. Yao, and P. Shi, ''Malicious URL detection with feature extraction based on machine learning,'' Int. J. High Perform. Comput. Netw., vol. 12, no. 2, pp. 166–178, 2018.
- Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang, ''Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism,'' IEEE Access, vol. 7, pp. 56329–56340, 2019.
- J. Feng, L. Zou, O. Ye, and J. Han, ''Web2Vec: Phishing webpage detection method based on multidimensional features driven by deep learning,'' IEEE Access, vol. 8, pp. 221214–221224, 2020.
- H. Cheng, J. Liu, T. Xu, B. Ren, J. Mao, and W. Zhang, ''Machine learning based low-rate DDoS attack detection for SDN enabled IoT networks,'' Int. J. Sens. Netw., vol. 34, no. 1, pp. 56–69, 2020.
- S. Christin, É. Hervet, and N. Lecomte, ''Applications for deep learning in ecology,'' Methods Ecol. Evol., vol. 10, no. 10, pp. 1632–1644, Oct. 2019.
- A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, ''PhishAri: Automatic realtime phishing detection on Twitter,'' in Proc. eCrime Res. Summit, Oct. 2012, pp. 1–12.
- H. Ma, Y. Zuo, and T. Li, ''Vessel navigation behavior analysis and multiple-trajectory prediction model based on AIS data,'' J. Adv. Transp., vol. 2022, pp. 1–10, Jan. 2022.
- J. Fang, B. Li, and M. Gao, ''Collaborative filtering recommendation algorithm based on deep neural network fusion,'' Int. J. Sens. Netw., vol. 34, no. 2, pp. 71–80, 2020.
- E. S. Gualberto, R. T. De Sousa, T. P. De Brito Vieira, J. P. C. L. Da Costa, and C. G. Duque, ''The answer is in the text: Multi-stage methods for phishing detection based on feature engineering,'' IEEE Access, vol. 8, pp. 223529–223547, 2020.
- W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, ''Short-term residential load forecasting based on LSTM recurrent neural network,'' IEEE Trans. Smart Grid, vol. 10, no. 1, pp. 841–851, Jan. 2019.
- E. Zhu, Y. Chen, C. Ye, X. Li, and F. Liu, ''OFS-NN: An effective phishing websites detection model based on optimal feature selection and neural network,'' IEEE Access, vol. 7, pp. 73271–73284, 2019.
- S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, ''Phishing email detection using natural language processing techniques: A literature survey,'' Proc. Comput. Sci., vol. 189, pp. 19–28, Jan. 2021.
- M. Korkmaz, O. K. Sahingoz, and B. Diri, ''Feature selections for the classification of webpages to detect phishing attacks: A survey,'' in Proc. Int. Congr. Hum.-Comput. Interact., Optim. Robotic Appl. (HORA), Jun. 2020, pp. 1–9.
- Dr.Abdul Bari ,Dr. Imtiyaz khan , Dr. Rafath Samrin , Dr. Akhil Khare , " VPC & Public Cloud Optimal Perfomance in Cloud Environment ",Educational Administration: Theory and Practic,ISSN No : 2148-2403 Vol 30- Issue -6 June 2024
- C. Singh and Meenu, ''Phishing website detection based on machine learning: A survey,'' in Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), Coimbatore, India, 2020, pp. 398–404. [Online]. Available: https://ieeexplore.ieee.org/document/9074400/authors#authors, doi: 10.1109/ICACCS48705.2020.9074400.
- L. Tang and Q. H. Mahmoud, ''A survey of machine learning-based solutions for phishing website detection,'' Mach. Learn. Knowl. Extraction, vol. 3, no. 3, pp. 672–694, Aug. 2021.
- A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, ''A comprehensive survey of AI-enabled phishing attacks detection techniques,'' Telecommun. Syst., vol. 76, pp. 139–154, Oct. 2020.
- M. Vijayalakshmi, S. M. Shalinie, M. H. Yang, and M. R. Meenakshi, ''Web phishing detection techniques: A survey on the state-of-the-art, taxonomy and future directions,'' IET.
- N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma, and H. Fujita, ''Deep learning for phishing detection: Taxonomy, current challenges and future directions,'' IEEE Access, vol. 10, pp. 36429–36463, 2022.
- A. Odeh, I. Keshta, and E. Abdelfattah, ''Machine LearningTechniquesfor detection of website phishing: A review for promises and challenges,'' in Proc. IEEE 11th Annu. Comput. Commun. Workshop Conf. (CCWC), Jan. 2021, pp. 0813–0818.
- S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, ''A systematic literature review on phishing email detection using natural language processing techniques,'' IEEE Access, vol. 10, pp. 65703–65727, 2022.
- T. Mahjabin, Y. Xiao, T. Li, and C. L. P. Chen, ''Load distributed and benign-bot mitigation methods for IoT DNS flood attacks,'' IEEE Internet Things J., vol. 7, no. 2, pp. 986–1000, Feb. 2020.