IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 8, March 2025

Intelligent Machine Learning Algorithms for Reliable Phishing URL Identification: Review

Sonali Dawange[1] and Dr. Prashant Yawalkar[2]
Student, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik[1]
Associate Professor, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik[2]
sonaliwaje91@gmail.com

Abstract: *Phishing attacks continue to pose a significant threat to cybersecurity, with attackers using deceptive techniques to lure unsuspecting users into divulging sensitive information such as login credentials, financial details, and personal data. As the volume and sophistication of phishing attacks increase, there is a growing need for effective detection mechanisms to thwart these malicious activities. Machine learning (ML) has emerged as a promising approach for detecting phishing URLs due to its ability to analyze large datasets and identify patterns indicative of malicious intent. This study presents a comprehensive literature review focusing on the methodologies employed in detecting phishing URLs using ML models. The review encompasses various ML techniques such as supervised learning, unsupervised learning, and deep learning, highlighting their strengths and limitations in the context of phishing URL detection. Additionally, the study explores the challenges faced in this domain, feature extraction techniques, and the dynamic nature of phishing attacks. Furthermore, the study examines the types of features commonly used in ML-based phishing URL detection, such as lexical features (e.g., URL length, domain age), content-based features (e.g., presence of keywords), and structural features (e.g., URL hierarchy). The analysis considers the relevance of these features in differentiating between legitimate and malicious URLs and discusses strategies for feature selection and extraction. This research provides valuable insights into the state-of-the-art methodologies, technologies, features, and datasets in ML-based phishing URL detection.*

Keywords: phishing URL detection, machine learning models, detection technologies, feature selection, datasets, cybersecurity defenses, malicious intent

I. INTRODUCTION

The era of digital technology has led to an increase in cyber threats, with phishing attacks standing out as a pervasive menace in the online landscape. According to the 2021 Data Breach Investigations Report by Verizon, phishing remains one of the top cybersecurity threats, accounting for approximately 36% of all data breaches. These attacks often leverage deceptive URLs to trick users into disclosing sensitive information, underscoring the critical need for robust detection mechanisms. Machine learning (ML) has emerged in the fight against phishing, leveraging algorithms to analyze patterns and detect malicious URLs with increasing accuracy. As per the Global Phishing Survey 2021, ML-based phishing detection systems have shown a significant improvement in detection rates, surpassing traditional rule-based methods.

This study delves into the realm of phishing URL detection using ML models, building upon the growing body of research and advancements in cybersecurity technology. With the rapid evolution of phishing tactics, including spear phishing and social engineering, organizations face escalating risks of data breaches and financial losses. The Cybersecurity Ventures Cybercrime Report 2021 predicts that global cybercrime costs will reach $6 trillion annually by 2021, underscoring the urgency of effective cybersecurity measures. By harnessing ML algorithms for phishing URL detection, organizations can bolster their defenses and mitigate the impact of cyber threats.

ISSN 2581-9429 IJARSCT

DOI: 10.48175/IJARSCT-24519

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 8, March 2025

A comprehensive understanding of the methodologies, technologies, and challenges in ML-based phishing URL detection is paramount for developing proactive defense strategies. The 2021 Phishing Activity Trends Report by the Anti-Phishing Working Group (APWG) highlights the increasing sophistication of phishing attacks, emphasizing the importance of continuous innovation in cybersecurity solutions. Through a meticulous literature review and analysis of key statistics and trends, this study aims to contribute valuable insights to the cybersecurity community. By leveraging ML's capabilities to detect malicious URLs and thwart phishing attempts, organizations can enhance their resilience against cyber threats and safeguard sensitive data in an increasingly digital world.

II. RELATED WORK

In thereseach, Lee and Kim [1] delve into the escalating issue of malicious activities on Twitter, particularly focusing on the dissemination of spam, phishing attempts, and malware via shared URLs. They critically evaluate existing detection methods, highlighting their limitations in effectively recognizing and mitigating these risks. Traditional approaches often falter due to their reliance on account features or network relationships within Twitter, which are susceptible to manipulation or resource-intensive operations. Similarly, detection schemes for suspicious URLs face evasion tactics such as time-based evasion and crawler evasion. To address these challenges, Lee and Kim introduce WARNINGBIRD, a novel system tailored for detecting suspicious URLs within Twitter's environment. WARNINGBIRD capitalizes on analyzing the correlations among URL redirect chains from multiple tweets, identifying frequently shared URLs within these chains to discern suspicious patterns and assess potential threat levels. The system integrates techniques like lexical analysis and dynamic behavior assessment, offering a promising solution to combat online threats effectively.

Utilizing a substantial dataset from the Twitter public timeline, Lee and Kim develop a statistical classifier within WARNINGBIRD to detect and categorize suspicious URLs efficiently. Their evaluation demonstrates WARNINGBIRD's efficacy in accurately pinpointing malicious URLs and its suitability for real-time monitoring of Twitter activity. Despite its strengths, WARNINGBIRD faces limitations related to the sophistication of evasion tactics and challenges in handling complex redirection mechanisms. For future advancements, integrating machine learning algorithms could enhance detection accuracy and adaptability.

The study[2] focuses on defensive strategies against phishing attacks, a pervasive cybersecurity threat. In response to the emergence of DeepPhish, a neural network-based system for generating phishing URLs, the authors emphasize the critical need for robust detection mechanisms. Their solution, PhishHaven, is an ensemble ML-based detection system designed to identify both AI-generated and human-crafted phishing URLs. This research marks a significant advancement as the first attempt to address the detection of phishing attacks orchestrated by both human and AI actors. Leveraging lexical analysis for feature extraction, PhishHaven integrates innovative techniques such as URL HTML Encoding to enhance its detection capabilities, especially concerning tiny URLs, a persisting issue in the field. Moreover, the authors introduce a URL Hit approach and an unbiased voting mechanism within PhishHaven to ensure precise classification and minimize misclassification occurrences. The implementation of multi-threading for parallel execution facilitates real-time detection, showcasing PhishHaven's practical applicability in combating phishing threats. The authors also provide theoretical insights into the effectiveness of their solution, showcasing its ability to consistently detect tiny URLs and future AI-generated phishing URLs with 100% accuracy based on selected lexical features. Through extensive experimentation using a benchmark dataset of more than 100,000 phishing and normal URLs, the efficacy of PhishHaven is empirically demonstrated, achieving an impressive accuracy rate of 98.00%. This performance outperforms existing lexical-based systems tailored for human-crafted phishing URL detection, underscoring the efficacy and superiority of PhishHaven in addressing evolving cybersecurity challenges. The study not only presents a novel approach to combating phishing attacks but also contributes valuable insights and methodologies to the broader landscape of cybersecurity research and defense mechanisms, paving the way for future advancements in this domain.

ISSN 2581-9429 IJARSCT

DOI: 10.48175/IJARSCT-24519

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 8, March 2025

In this study, the author [3] explores the intricate landscape of user vulnerability to phishing attacks by conducting a systematic review and meta-analysis of prior research findings.The primary objective is to offer a comprehensive understanding of the factors influencing susceptibility to phishing, with a particular focus on age and gender differences. The results reveal a nuanced scenario, with contradictory outcomes across reviewed studies. Despite widespread assumptions about older users being more vulnerable to phishing attacks, over half of the examined studies show no statistically significant correlation between age and susceptibility. Additionally, there are variations in the influence of gender, with some studies indicating higher susceptibility among females and others finding no substantial difference. The meta-analysis conducted by the author uncovers several significant insights. Firstly, it establishes a notable association between participants' age and susceptibility to phishing attacks, challenging previously held assumptions. Furthermore, the findings suggest higher susceptibility among females compared to males, indicating a potential focus area for targeted interventions or awareness initiatives. Lastly, the meta-analysis highlights the efficacy of user training in improving detection capabilities, offering a proactive strategy for mitigating phishing risks. This comprehensive examination significantly advances our understanding of phishing susceptibility and provides valuable insights for guiding future research endeavors and enhancing cybersecurity practices.

III. ANATOMY OF PHISHING ATTACKS

Phishing attacks are a prevalent form of cyber threat that target individuals, businesses, and organizations with the aim of stealing sensitive information or causing financial harm. Understanding the anatomy of phishing attacks is crucial for identifying red flags, implementing effective cybersecurity measures, and mitigating the risks associated with such malicious activities.

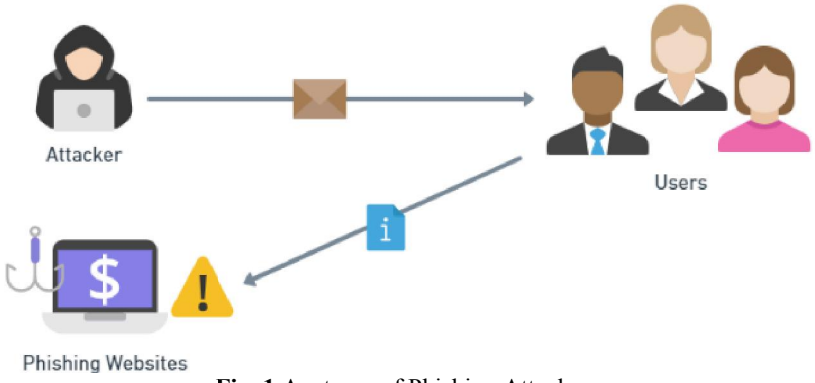

Fig. 1. Anatomy of Phishing Attack

Social Engineering Tactics: Phishing attacks often rely on social engineering tactics to manipulate human psychology and trick individuals into taking actions that benefit the attacker. This may involve creating a sense of urgency (e.g., claiming an account will be suspended unless immediate action is taken), impersonating trusted entities (e.g., banks, government agencies, or reputable companies), or exploiting emotions (e.g., using fear or curiosity to prompt a response).

Deceptive Communication: Phishing attacks typically involve deceptive communication methods, such as emails, text messages, or phone calls, that appear legitimate at first glance. Attackers often use sophisticated techniques to spoof email addresses or mimic official websites, making it challenging for recipients to discern between genuine and fraudulent communications.

Malicious URLs and Attachments: Phishing emails often contain links to malicious websites or attachments infected with malware. These URLs may appear genuine by using domain names similar to legitimate ones or employing URL shortening services to obfuscate the destination. Clicking on such links or downloading malicious attachments can compromise device security and lead to data theft or system infiltration.

DOI: 10.48175/IJARSCT-24519

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 8, March 2025

Credential Theft: One of the primary objectives of phishing attacks is to steal login credentials, such as usernames and passwords, for online accounts or sensitive systems. Attackers use various techniques, such as phishing pages that mimic login portals or fake login prompts within emails, to deceive users into divulging their credentials unwittingly.

Information Gathering: Phishing attackers often conduct extensive information gathering to personalize their attacks and increase their chances of success. This may involve researching targets on social media platforms, gathering publicly available information, or using previously compromised data to craft convincing and tailored phishing messages.

Spoofed Identities: Phishing attacks frequently involve spoofing trusted identities, such as known brands, colleagues, or contacts, to establish credibility and lower victims' defenses. By masquerading as a familiar entity, attackers increase the likelihood of recipients engaging with the malicious content or providing sensitive information.

Phishing Campaign Variants: Phishing attacks come in various forms, including spear phishing, and vishing (phishing via phone calls or voice messages). Each variant employs tailored strategies and tactics to achieve its objectives, emphasizing the adaptability and persistence of phishing threat actors.

Post-Compromise Actions: In successful phishing attacks, threat actors may exploit compromised credentials or systems to carry out further malicious activities, such as unauthorized access, data exfiltration, or deploying ransomware. This underscores the importance of prompt detection, response, and mitigation measures following a phishing incident.

understanding the anatomy of phishing attacks involves recognizing the social engineering tactics, deceptive communication methods, malicious payloads, and post-compromise actions employed by threat actors. By staying vigilant, implementing cybersecurity best practices, and educating users about phishing risks, organizations can enhance their resilience against these pervasive and evolving cyber threats.

URL and its components

A URL, which stands for Uniform Resource Locator, is a standardized addressing format used to specify the location of a resource on the internet. It serves as a way to access web pages, files, images, videos, and other resources hosted on servers across the World Wide Web. A URL consists of several components, each providing specific information about the resource and its location.

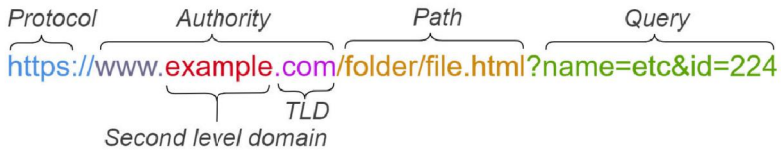

Fig. 2.URL with its Components

Let's break down the components of a URL:

Scheme: The scheme, also known as the protocol, indicates the communication protocol used to access the resource. Common schemes include "http://" for Hypertext Transfer Protocol (HTTP), "https://" for HTTP Secure (HTTPS), "ftp://" for File Transfer Protocol (FTP), "mailto://" for email addresses, and "tel://" for telephone numbers. The scheme is followed by a colon and two forward slashes ("://").

Domain: The domain, also referred to as the hostname, identifies the specific server or network location where the resource is hosted. It can be a human-readable domain name (e.g., "example.com") or an IP address (e.g., "192.168.1.1"). Domains are hierarchical, with subdomains (e.g., "blog.example.com") indicating different sections or services within a domain.

Port: The port number, if specified, indicates the communication endpoint on the server where the resource is hosted. For example, "http://example.com:80" specifies port 80 for HTTP communication, while "https://example.com:443" specifies port 443 for HTTPS communication. Default ports are often omitted, as HTTP uses port 80 and HTTPS uses port 443 by default.

ISSN 2581-9429 IJARSCT

DOI: 10.48175/IJARSCT-24519

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 8, March 2025

Path: The path component specifies the location of the specific resource within the server's file system or directory structure. It follows the domain and optional port number, separated by a forward slash ("/"). For example, in the URL "https://example.com/products/shoes", "/products/shoes" is the path indicating the "shoes" resource within the "products" directory on the server.

Query Parameters: Query parameters, also known as query strings, are additional information appended to the URL to pass data to the server or modify the request. They are preceded by a question mark ("?") and consist of key-value pairs separated by ampersands ("&"). For instance, in the URL "https://example.com/search?q=keywords&page=1", "q=keywords" and "page=1" are query parameters specifying search keywords and page number, respectively.

Fragment Identifier: The fragment identifier, often denoted by a hash symbol ("#"), points to a specific section or anchor within the resource, such as a specific section of a webpage. It is used for navigation purposes within the resource. For example, in the URL "https://example.com/page#section2", "#section2" indicates the "section2" anchor within the "page" resource.

IV. PROPOSED SYSTEM

A simplified machine learning (ML) methodology for phishing URL detection involves several key steps. First, gather a dataset consisting of labeled URLs, categorizing them as either phishing or legitimate. Next, preprocess the data by extracting features such as URL length, domain age, presence of suspicious keywords, and domain reputation scores. Then, split the dataset into training and testing sets and select an appropriate ML algorithm, such as logistic regression, decision trees, or random forests. Train the model using the training data, optimizing hyperparameters through techniques like cross-validation. Evaluate the model's performance using the testing set, assessing metrics such as accuracy, precision, recall, and F1 score. Finally, deploy the trained ML model to classify new URLs as phishing or legitimate based on their features, thereby aiding in phishing detection.

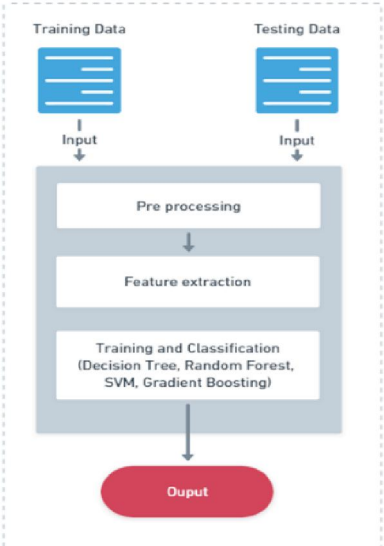

Fig. 3. Proposed System Architecture

The proposed system goes through several crucial stages to effectively detect phishing URLs. Initially, a varied dataset containing both legitimate and phishing URLs is compiled from various sources to ensure it reflects real-world scenarios. After gathering data, preprocessing steps are taken to refine the dataset, which involves tasks like removing duplicates, managing missing values, and standardizing URL formats. Feature engineering methods are then used to extract meaningful attributes from the URLs, such as domain age, URL length, and lexical features, which are essential for training the model.

ISSN 2581-9429 IJARSCT
DOI: 10.48175/IJARSCT-24519

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 8, March 2025

In our proposed system, we extract various URL features using different techniques, including structural components, lexical elements, and semantic attributes, to obtain comprehensive information from URLs. By examining structural components like URL length, redirects presence, and subdomains, we gain insights into URL organization and behavior. Additionally, by analyzing lexical elements such as symbols, prefixes or suffixes, and email-related details, we gather information about URL syntax and composition. Furthermore, our analysis extends to semantic attributes like HTTPS usage, domain registration length, domain age, website traffic, and Google index status, providing valuable insights into URL credibility, trustworthiness, and popularity.

4.1 Feature Extraction Techniques

Feature extraction techniques for URLs encompass three main categories: Structural Components, Lexical Elements, and Semantic Attributes. Structural Components involve analyzing the URL's hierarchical structure, including the domain name, path, query parameters, and fragment identifier, to extract features such as URL length, domain age, presence of subdomains, and depth of directory structure. Lexical Elements focus on the textual content of the URL, extracting features like the presence of suspicious keywords, misspellings, hyphens, digits, or special characters indicative of phishing attempts. Semantic Attributes delve deeper into the context and semantics of the URL, considering factors such as domain reputation, SSL encryption status, WHOIS information, and the presence of redirections or obfuscation techniques. By combining features from these three categories, feature extraction techniques can create a comprehensive set of attributes that machine learning models can leverage to effectively detect phishing URLs and differentiate them from legitimate ones.

Table 1. Feature Extraction Techniques

Technique Used	Explanation
Structural Components	These functions focus on analyzing the structural elements of URLs, such as their length (longUrl), presence of redirects (redirecting), presence of subdomains (SubDomains), existence of certain tags like <iframe> (IframeRedirection), and the presence of links in script tags (LinksInScriptTags). These features provide insights into the organization and behavior of URLs.
Lexical Elements	Functions in this category examine specific lexical elements within the URLs, such as the presence of certain symbols (symbol), prefixes or suffixes (prefixSuffix), and the existence of email-related information (InfoEmail). These elements offer clues about the syntax and composition of the URLs.
Semantic Attributes	These functions delve deeper into the semantics of URLs, considering factors such as the use of HTTPS (Hppts), domain registration length (DomainRegLen), age of domain (AgeofDomain), DNS recording (DNSRecording), website traffic (WebsiteTraffic), PageRank (PageRank), Google index status (GoogleIndex), and links pointing to the page (LinksPointingToPage). These attributes provide insights into the credibility, trustworthiness, and popularity of the URLs.

4.1 Machine Learning Models

Phishing URL detection is commonly approached as a binary classification problem, where the goal is to determine whether a given web page is legitimate or malicious (phishing). Several supervised learning algorithms are utilized to address this challenge effectively. One of the traditional classifiers often employed is the Naive Bayes Classifier, which is based on Bayes' theorem and assumes independence among features. Support Vector Machine (SVM) is another popular choice, known for its ability to handle high-dimensional data and nonlinear decision boundaries effectively. Random Forest, a robust ensemble learning method, is favored for its capability to handle noisy data and feature interactions by constructing multiple decision trees. Gradient Boosting Classifier, a boosting algorithm that combines

DOI: 10.48175/IJARSCT-24519

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 8, March 2025

weak learners to create a strong classifier, is also widely used due to its high accuracy and robustness against overfitting.

These machine learning models leverage various features extracted from the URL, including structural components (such as URL length, domain age, and path depth), lexical elements (such as presence of suspicious keywords or unusual characters), and semantic attributes (such as domain reputation and SSL encryption status). By training these models on labeled datasets containing both legitimate and phishing URLs, they learn to distinguish between benign and malicious web pages based on the patterns and characteristics present in the data. The models are then evaluated using metrics such as accuracy, precision, recall, and F1 score to assess their performance in correctly classifying URLs.

techniques such as feature selection, hyperparameter tuning, and ensemble methods (e.g., combining multiple classifiers for improved accuracy) are often employed to enhance the effectiveness of these ML models for phishing URL detection. Continuous research and development in this area aim to improve the detection capabilities and robustness of these algorithms against evolving phishing techniques and tactics employed by cybercriminals.

Table 2. Accuracy of various model

No.	ML Model	Accuracy	F1 Score	Recall	Precision
1	Gradient Boosting Classifier	97.4%	97.7%	99.4%	98.6%
2	Random Forest	96.7%	99.3%	99.3%	99.0%
3	Support Vector Machine	96.4%	96.8%	98.0%	96.5%
4	Naive Bayes Classifier	60.5%	45.4%	29.2%	99.7%

Table 2 - Comparison of Phishing Detection Systems

Paper Title	Methodology/Technology	Observations/Limitations	Analysis
S. Lee and J. Kim, "WarningBird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream"	WARNINGBIRD system: Analyzing URL redirect chains, lexical analysis, dynamic behavior assessment	Limitations: Sophistication of evasion tactics, handling complex redirection mechanisms	Utilizing a statistical classifier within WARNINGBIRD showcased efficacy in pinpointing malicious URLs, albeit facing challenges with evasion tactics. Integration of machine learning algorithms is suggested for future advancements to enhance accuracy and adaptability.
M. Sameen, K. Han and S. O. Hwang, "PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System"	PhishHaven system: Ensemble machine learning-based, lexical analysis, URL HTML Encoding, URL Hit approach, unbiased voting	Observations: Achieved 98.00% accuracy, superior to existing systems; Emphasized detection of AI-generated and human-crafted phishing URLs	PhishHaven demonstrated superior performance in detecting phishing URLs, especially AI-generated ones, using innovative techniques and achieving high accuracy rates. The study suggests the practical applicability of PhishHaven in combating evolving phishing threats, contributing valuable

DOI: 10.48175/IJARSCT-24519

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 8, March 2025

			insights and methodologies to cybersecurity research.
S. Baki and R. M. Verma, "Sixteen Years of Phishing User Studies: What Have We Learned?"	Systematic review and meta-analysis of prior research findings	Observations: Contradictory outcomes regarding age and gender susceptibility; Suggested efficacy of user training	The meta-analysis in the study challenges assumptions about age and gender vulnerability to phishing, suggesting potential focus areas for interventions. It highlights the importance of user training in improving detection capabilities, contributing significantly to advancing understanding of phishing susceptibility and guiding future research in enhancing cybersecurity practices.
Ehsan Nowrooziand Abhishek, "Novel Lexical and Web-scrapped Features for Fraudulent Advertisement URL Detection using Machine Learning"	In this study, a novel approach is proposed for detecting fraudulent advertisement URLs using machine learning (ML) techniques. A unique set of lexical and Web-scrapped features is extracted and combined into six different categories.	vulnerabilities in decision tree-based models to limited knowledge attack scenarios are observed. Exploratory attacks during the test phase and Zeroth Order Optimization adversarial attacks on detection models are implemented to analyze their vulnerability.	The proposed approach significantly enhances fraudulent URL detection capabilities, demonstrating potential for improved security against cyber-attacks. While achieving high accuracy, vulnerabilities in decision tree-based models underscore the need for robustness against adversarial attacks.
R. R. Rout, G. Lingam and D. V. L. N. Somayajulu, "Detection of Malicious Social Bots Using Learning Automata with URL Features in Twitter Network."	Learning Automata-Based Algorithm (LA-MSBD)	Dependency on URL-based features reduces detection time; Social graph-based features are time-consuming; Difficulty for bots in manipulating URL redirection chains	Integration of trust computation model enhances accuracy; LA-MSBD algorithm shows improvement in precision, recall, F-measure, and accuracy compared to existing methods

ISSN 2581-9429 IJARSCT
DOI: 10.48175/IJARSCT-24519

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 8, March 2025

V. CONCLUSION

The intelligent utilization of machine learning algorithms plays a crucial role in identifying phishing URLs reliably. Various supervised learning techniques like the Naive Bayes Classifier, Support Vector Machine, Random Forest, and Gradient Boosting Classifier have significantly advanced the accurate discrimination between legitimate and malicious URLs. These algorithms harness a wide range of URL features, encompassing structural components, lexical elements, and semantic attributes, to construct robust models capable of detecting phishing attempts with exceptional precision and recall.The continuous enhancement of these machine learning models, combined with progress in feature selection, hyperparameter tuning, and ensemble methods, has led to heightened reliability and efficacy in phishing URL identification. However, it's imperative to acknowledge the dynamic nature of cyber threats, necessitating ongoing research and development endeavors to stay ahead of sophisticated phishing techniques employed by malicious actors. By harnessing the power of intelligent machine learning algorithms, organizations can fortify their cybersecurity defenses and mitigate the risks posed by phishing attacks, thereby safeguarding sensitive data and upholding user trust in online platforms.

In the proposed system, a Gradient Boosting Classifier model is currently being meticulously developed and evaluated for phishing URL detection, showing robust performance across various evaluation metrics such as accuracy, F1 score, recall, and precision. It achieved an accuracy of 98.9% on the training data and 97.4% on the test data, demonstrating strong classification capabilities by accurately distinguishing between phishing and legitimate URLs. The F1 score, considering both precision and recall, was notably high at 99.0% on the training data and 97.7% on the test data, indicating a balanced performance in correctly identifying phishing URLs while minimizing false positives. The model also exhibited exceptional recall scores of 99.4% on the training data and 98.9% on the test data, highlighting its effectiveness in capturing a significant portion of actual phishing URLs. However, the precision scores, though high at 98.6% on the training data and 96.6% on the test data, suggest some instances of misclassification, particularly false positives, which could be further improved to enhance the model's reliability in accurately identifying legitimate URLs.

REFERENCES

[1]. S. Lee and J. Kim, "WarningBird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream," in IEEE Transactions on Dependable and Secure Computing, vol. 10, no. 3, pp. 183-195, May-June 2013, doi: 10.1109/TDSC.2013.3.

[2]. M. Sameen, K. Han and S. O. Hwang, "PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System," in IEEE Access, vol. 8, pp. 83425-83443, 2020, doi: 10.1109/ACCESS.2020.2991403.

[3]. S. Baki and R. M. Verma, "Sixteen Years of Phishing User Studies: What Have We Learned?," in IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 2, pp. 1200-1212, 1 March-April 2023, doi: 10.1109/TDSC.2022.3151103.

[4]. E. Nowroozi, Abhishek, M. Mohammadi and M. Conti, "An Adversarial AttackAnalysis on Malicious Advertisement URL Detection Framework," in IEEE Transactions on Network and Service Management, vol. 20, no. 2, pp. 1332-1344, June 2023, doi: 10.1109/TNSM.2022.3225217.

[5]. R. R. Rout, G. Lingam and D. V. L. N. Somayajulu, "Detection of Malicious Social Bots Using Learning Automata with URL Features in Twitter Network," in IEEE Transactions on Computational Social Systems, vol. 7, no. 4, pp. 1004-1018, Aug. 2020, doi: 10.1109/TCSS.2020.2992223.

[6]. Y. Sönmez, T. Tuncer, H. Gökal and E. Avcı, "Phishing web sites features classification based on extreme learning machine," 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, Turkey, 2018, pp. 1-5, doi: 10.1109/ISDFS.2018.8355342.

[7]. M. H. Alkawaz, S. J. Steven and A. I. Hajamydeen, "Detecting Phishing Website Using Machine Learning," 2020 16th IEEE International Colloquium on Signal Processing &Its Applications (CSPA), Langkawi, Malaysia, 2020, pp. 111-114, doi: 10.1109/CSPA48992.2020.9068728.

ISSN 2581-9429 IJARSCT

DOI: 10.48175/IJARSCT-24519

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 8, March 2025

[8]. L. Wu, X. Du and J. Wu, "Effective Defense Schemes for Phishing Attacks on Mobile Computing Platforms," in IEEE Transactions on Vehicular Technology, vol. 65, no. 8, pp. 6678-6691, Aug. 2016, doi: 10.1109/TVT.2015.2472993

[9]. H. BOUIJIJ and A. BERQIA, "Machine Learning Algorithms Evaluation for Phishing URLs Classification," 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Alkhobar, Saudi Arabia, 2021, pp. 01-05, doi: 10.1109/ISAECT53699.2021.9668489.

[10]. A Dawabsheh, M. Jazzar, A. Eleyan, T. Bejaoui and S. Popoola, "An Enhanced Phishing Detection Tool Using Deep Learning From URL," 2022 International Conference on Smart Applications, Communications and Networking (SmartNets), Palapye, Botswana, 2022, pp. 1-6, doi: 10.1109/SmartNets55823.2022.9993984.

[11]. M. Singla, K. S. Gill, R. Chauhan, V. Singh and D. Banerjee, "Phishing URL Classification Using K-Nearest Neighbour and Logistic Regression Machine Learning Approaches," 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/SMARTGENCON60755.2023.10442844.

[12]. S. Mondal, D. Maheshwari, N. Pai and A. Biwalkar, "A Review on Detecting Phishing URLs using Clustering Algorithms," 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, 2019, pp. 1-6, doi: 10.1109/ICAC347590.2019.9036837.

ISSN 2581-9429 IJARSCT

DOI: 10.48175/IJARSCT-24519