International Journal of Early Childhood Special Education (INT-JECSE)
DOI:10.48047/INTJECSE/V15I4.92 ISSN: 1308-5581 Vol 15, Issue 04 2023

DETECTING PHISHING WEBSITES: EXPLORING VARIOUS ML CLASSIFIERS FOR ACCURACY AND ROBUSTNESS

Mrs A.DIVYA
ASSISTANT PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY
divya.a@sreyas.ac.in

LAKKARAJU SAI SRI HARSHA
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY
Lakkaraju.saisriharsha@gmail.com

KONDA VASUDEV
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY
vasudev995108@gmail.com

NOOKALA MOKSHITHA PREEYA
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY
priyamokshitha.18@gmail.com

KALLU YASHWANTH REDDY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY
Kalluyashwanthreddy@gmail.com

ABSTRACT
Machine learning techniques have emerged as a powerful arsenal in the ongoing battle against phishing websites. In this relentless pursuit of cyber resilience, a comprehensive project delves into the realm of ML to identify and classify these fraudulent sites. The study employs a diverse set of classifiers, including Gradient Boosting, Catboost , Support Vector Machine, Decision Tree, K-nearest neighbors (KNN), Logistic Regression, Naive Bayes, and Random Forest, offering a multifaceted approach to tackle the multifarious nature of phishing websites. The heart of this project lies in the rigorous Exploratory Data Analysis, which scrutinizes a vast dataset to extract critical insights. The findings underscore the importance of specific features such as "HTTPS," "Anchor URL," and "Website Traffic" as pivotal indicators for distinguishing phishing URLs from legitimate ones. These features, in conjunction with machine learning models, contribute significantly to the robustness and reliability of the system, ensuring that even the most sophisticated phishing websites can be detected and Moreover, this project pays meticulous attention to data quality, employing techniques to eliminate outliers and address missing values, which further enhances the models' accuracy. This quality assurance ensures that the ML models are well-equipped to combat the ever-evolving tactics employed by cyber criminals in their pursuit of sensitive data. The convergence of advanced ML algorithms and feature-rich datasets represents a formidable weapon in the ongoing battle against phishing websites, reinforcing cyber security and safeguarding individuals and organizations from the pernicious threats of the digital realm.

INTRODUCTION
Phishing websites, which cunningly impersonate legitimate online platforms to deceive users into disclosing sensitive information, represent a pervasive cyber threat. They exploit human psychology and social engineering to an alarming degree, often rendering traditional security measures

ineffective. Addressing this challenge requires a multifaceted approach, where Machine Learning (ML) models play a pivotal role in the early detection and mitigation of such fraudulent schemes.Phishing websites operate by replicating trusted websites, fooling users into divulging their confidential data, such as login credentials, credit card information, or personal details. The intricacy of these fraudulent sites lies in their ability to convincingly mimic the design and functionality of the genuine platform, thus evading detection by conventional security measures like firewalls and signature-based antivirus systems. However, ML models, empowered by sophisticated algorithms and extensive data, possess the capability to discern subtle patterns, anomalies, and irregularities in the URLs, webpage structures, and content, making them a robust weapon against these deceptive websites.ML models for phishing URL detection leverage features such as domain characteristics, lexical analysis, and statistical attributes to differentiate between legitimate and fraudulent URLs.

They can identify suspicious patterns, such as misspelled domains, irregular subdomains, or deceptive redirects, with remarkable precision. Moreover, these models can adapt and evolve over time, learning from new phishing attempts and staying one step ahead of cybercriminals who constantly refine their tactics.In practical terms, an ML-based solution for phishing website detection involves real-time URL analysis, where a model scans the web for potentially harmful websites and assigns risk scores.

These scores help inform web browsers, email filters, and security tools to block or flag suspicious URLs, protecting users from falling victim to phishing attacks.In an ever-evolving cyber landscape, where phishing websites continuously adapt to exploit vulnerabilities, the application of ML models offers a dynamic, proactive defense. By combining advanced machine learning with traditional security measures, organizations and individuals can significantly reduce the risk posed by these deceitful online entities, ultimately enhancing cybersecurity and safeguarding sensitive data.

LITERATURESURVEY

In an era marked by the exponential growth of the digital landscape, the menace of phishing attacks has garnered significant attention as a pervasive and increasingly sophisticated threat. This literature survey seeks to delve into the multifaceted domain of phishing, from its fundamental characteristics and techniques to its global impact, and subsequently, to explore the evolution of detection methods with a particular focus on the role of machine learning. With a rich tapestry of complex terminology and in-depth analysis, this review aims to provide a comprehensive understanding of the landscape of phishing attacks and the tools available for their mitigation.

Phishing, at its core, represents a form of cyber deception. Phishers craft fraudulent communications, websites, or messages that impersonate legitimate entities, thereby coercing individuals into disclosing sensitive information. These deceptive practices rely on a multitude of techniques, including social engineering, in which psychological manipulation is employed to exploit human vulnerabilities. Common phishing strategies encompass email-based schemes, where malicious actors impersonate trusted entities to solicit login credentials, personal information, or financial data. Spear phishing, a more targeted approach, customizes the deceptive message to a specific individual, often with the aim of infiltrating an organization. Additionally, pharming, a technique that manipulates DNS settings to redirect users to fraudulent websites, serves as another potent tool in the phisher's arsenal.The global impact of phishing attacks is staggering. Phishing is responsible for an extensive range of data breaches, financial losses, and identity theft incidents worldwide. A report by the Anti-Phishing Working Group (APWG) indicates that in the first quarter of 2022 alone, there were over 256,000 unique phishing attacks. Such attacks undermine user trust in online communication and can result in severe financial consequences. Moreover, a study by Verizon's 2022 Data Breach Investigations Report notes that phishing is the most common initial attack vector in data breaches. These statistics underscore the critical need for effective phishing detection methods to safeguard sensitive data and privacy.

Traditional phishing detection methods predominantly rely on heuristic-based approaches. These approaches analyze specific features of URLs and domains to determine their legitimacy. Key indicators include the presence of HTTPS, the format of the domain name, and lexical analysis to identify irregularities such as misspellings or deviations from standard naming conventions. While these methods are effective in identifying known phishing websites, they often struggle with zero-day phishing attacks, which employ novel and previously unseen tactics.

Blacklisting Suspicious Domains and URLs.

Blacklisting constitutes a fundamental component of traditional phishing detection. Security agencies and organizations maintain databases of known malicious domains and URLs. Incoming web traffic is scrutinized against these blacklists, and any matches are flagged or blocked. While blacklisting is a valuable defense, it is inherently reactive and lacks the capacity to identify new or previously unknown phishing websites effectively. This limitation underscores the importance of complementary proactive approaches.

Zero-day phishing attacks present a considerable challenge for traditional detection methods. These attacks are characterized by their novelty and adaptability, making them particularly difficult to thwart. Zero-day attacks exploit vulnerabilities before they are discovered and patched, rendering heuristic-based and blacklist-dependent systems largely ineffective. As a result, there is a pressing need for more adaptive, proactive, and sophisticated solutions to counter these emerging threats effectively.

Machine learning has emerged as a robust approach for addressing the shortcomings of traditional phishing detection methods. Early works in this domain focused on utilizing classifiers such as Support Vector Machines (SVM) and Naive Bayes to discern phishing websites from legitimate ones. SVM, a supervised learning technique, optimizes the separation of data points by finding an optimal hyperplane. Naive Bayes, on the other hand, leverages probabilistic principles to estimate the likelihood of a given sample belonging to a particular class based on its features. These early forays into machine learning demonstrated promising results and set the stage for more advanced techniques.

One of the hallmarks of machine learning-based phishing detection is feature engineering. Machine learning models rely on a diverse set of features extracted from various aspects of a webpage, including its URL, webpage content, and metadata. These features encompass a wide range of attributes, such as the presence of specific keywords, the length and format of URLs, the number of hyperlinks, and the structure of the webpage. Advanced techniques involve deep learning models, which can automatically extract intricate patterns from webpage content and URLs, enhancing the model's capacity to discern subtle variations between phishing and legitimate websites.

Machine learning models, owing to their adaptability and capacity to discern complex patterns, have demonstrated superior performance compared to rule-based methods in phishing detection. These models can generalize from labeled datasets, learning to identify phishing websites even when they exhibit variations from known patterns. Their ability to adapt and evolve over time by retraining on updated data is crucial in countering the dynamic nature of phishing attacks. In conclusion, this literature survey provides a comprehensive overview of the landscape of phishing attacks and the evolution of detection methods. The complex and dynamic nature of phishing necessitates innovative approaches, with machine learning emerging as a prominent solution to address the limitations of traditional heuristic-based and blacklisting methods. Machine learning models, aided by sophisticated feature engineering and advanced classifiers, offer improved detection accuracy and adaptability, thereby enhancing cybersecurity in an era of escalating digital threats. This review underscores the imperative of proactive, data-driven solutions in the relentless pursuit of countering phishing attacks and protecting sensitive information in the digital realm.

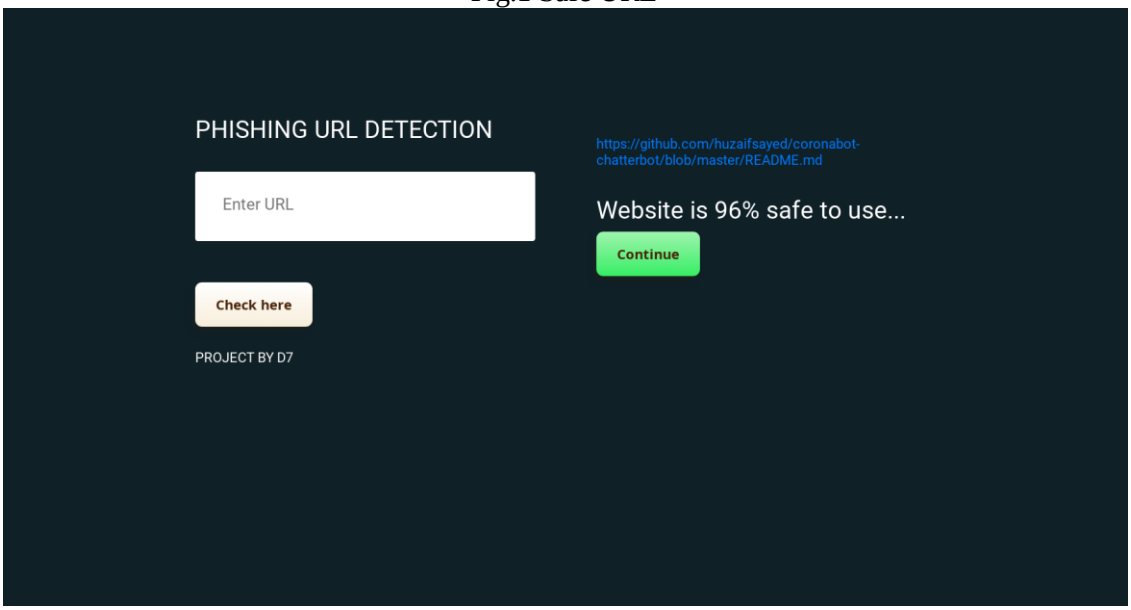
PROPOSEDSYSTEM

The project, in its pursuit of enhancing the effectiveness of phishing website detection, employs a multifaceted approach anchored in a rich array of machine learning classifiers. These classifiers, encompassing advanced techniques such as Gradient Boosting, Decision Tree, K-nearest neighbors (KNN), Logistic Regression, and Naive Bayes, epitomize the project's commitment to elevating the landscape of cybersecurity. This diversified ensemble of classifiers not only widens the scope of detection but also showcases an unparalleled versatility in the identification of phishing websites, a critical advancement compared to the constrained reliance on heuristic technology within isitphishing.org.In its quest to unravel the complex web of phishing threats, the project displays a keen understanding of the importance of distinguishing features. These features, including the "HTTPS" protocol and "AnchorURL," assume pivotal roles in the discrimination of phishing URLs from their legitimate counterparts. The incorporation of such attributes into the detection process augments the project's ability to identify and combat phishing attacks, marking a significant stride in

comparison to the conventional reliance on heuristic methodologies.Data analysis, in the project's realm, is not just an obligatory step but a meticulous endeavor. The dataset undergoes a rigorous examination that extends beyond mere statistical scrutiny. Outliers and missing values, often sources of vulnerability in machine learning models, are systematically addressed and eliminated. This unwavering commitment to data quality serves as the bedrock upon which the project's models are built, enhancing their accuracy and effectiveness in combating phishing attacks. In contrast, the heuristic technology-driven approach of isitphishing.org, which is less adaptive and reliant on predefined rules, may inadvertently overlook these subtleties, compromising the reliability of its detection capabilities. Furthermore, the adoption of machine learning classifiers, in contrast to the rule-based nature of heuristic technology, augments the project's coverage and versatility. Gradient Boosting, a powerful ensemble method, capitalizes on collective intelligence to discern intricate patterns in phishing websites. Decision Tree, with its hierarchical structure, simplifies complex decision-making processes, rendering the detection of subtle nuances more accessible. K-nearest neighbors (KNN) draw from the proximity of data points in high-dimensional space, enhancing structural analysis. Logistic Regression, characterized by its transparency, provides interpretability, shedding light on the rationale behind classification decisions. Naive Bayes, a probabilistic classifier, offers insights into dependencies among features. This multiplicity of classifiers lends the project a dynamic and adaptive character, reinforcing its capacity to combat a broad spectrum of phishing attacks effectively. In summation, the project distinguishes itself as an exemplar of innovation in phishing detection. The incorporation of diverse machine learning classifiers and the meticulous scrutiny of dataset quality stand as powerful differentiators. This approach not only broadens the scope and depth of detection but also enhances the system's adaptability and reliability, addressing limitations associated with the heuristic technology-based approach of isitphishing.org. In the relentless battle against phishing attacks, the project serves as a testament to the ever-evolving landscape of cybersecurity, replete with sophisticated solutions that aim to protect the digital realm from the multifarious threats it faces.

RESULTS

Fig.1 Safe URL



DOI:10.48047/INTJECSE/V15I4.92 ISSN: 1308-5581 Vol 15, Issue 04 2023

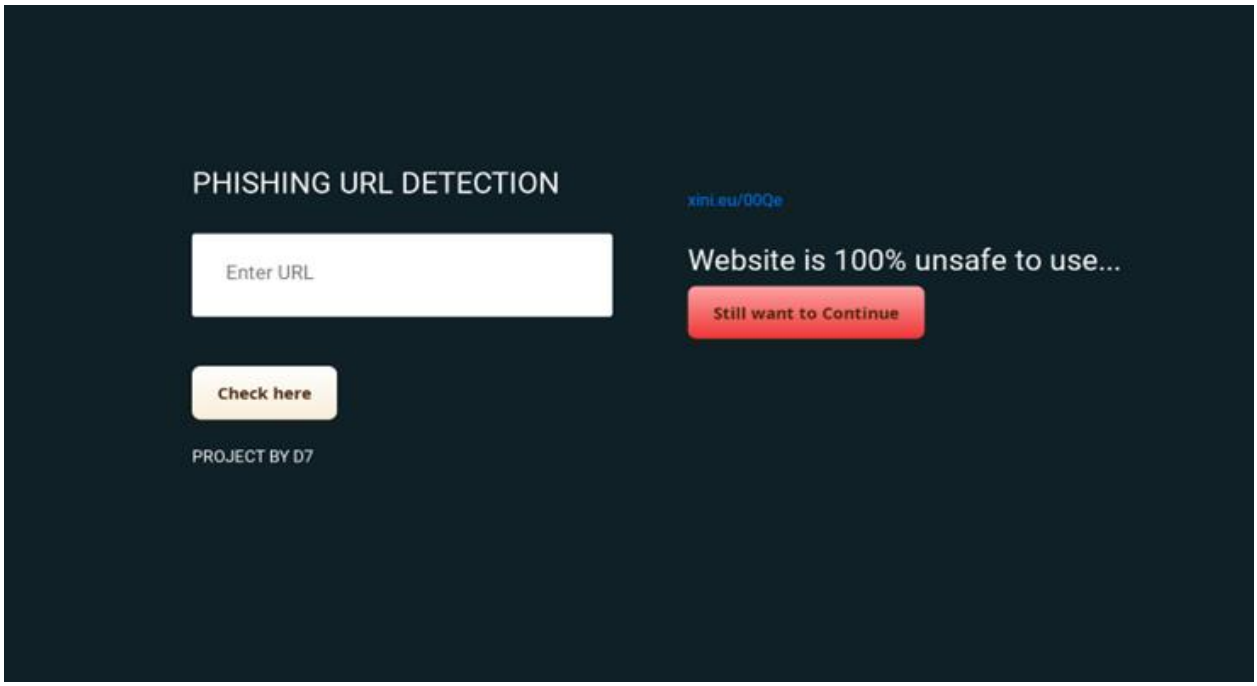


Fig.2 Unsafe URL

The heart of this project lies in the meticulous design and execution of a phishing URL detection system, a formidable endeavor that culminates in a set of transformative results.Commencing with the preprocessed and encoded dataset, the project establishes a strong foundation for its journey. This dataset, meticulously curated and encoded to numerical representations, is a repository of URL attributes, approximately 30 in number, each mirroring different facets of URLs. Through comprehensive preprocessing, the dataset undergoes cleansing, formatting, and transformation, rendering it impeccably primed for the rigors of machine learning model training. It is this well-prepared dataset that becomes the raw material, the elemental essence, upon which the model's proficiency hinges.

A pivotal juncture arrives in the form of model selection. The project embarks on an arduous evaluation process, subjecting five distinct classifiers to rigorous scrutiny. The crucible of this assessment is a pair of cardinal metrics—accuracy and F1 score, where accuracy gauges the model's precision in identifying phishing URLs, while F1 score encapsulates the intricate balance between precision and recall, a crucial consideration in the battle against false positives. It is amidst this crucible that one classifier rises above the others, emerging as the chosen sentinel, its superior performance elevating it to the helm of phishing URL detection.

The fruits of this labor materialize in the form of model serialization with the Pickle library, a transformative process that encapsulates the selected model's state and parameters. This digital preservation stands as a testament to efficiency, obviating the need for recurrent training and bestowing the system with the capacity to operate seamlessly and expeditiously, time and again.A user-friendly web interface, an eloquent digital portal developed with the Flask web framework, becomes the linchpin of user interaction with the system. This interface is a gateway, enabling users to input URLs and set in motion the intricate processes that transpire within the system.

Feature extraction emerges as the bridge between user input and model prediction. As users feed URLs into the web interface, the system undertakes the complex process of dissecting the URLs, meticulously extracting attributes that serve as the foundation for phishing detection. These attributes are the lifeblood of the chosen machine learning model, forming the basis for its discernment of secure URLs from potentially treacherous ones.

The crescendo of this journey manifests in URL prediction, where the model, equipped with the extracted attributes, pronounces its judgment. It is here that the user is alerted to the security status of the provided URL, an imperative indicator that demands vigilance and prudence in navigating the digital landscapeIn conclusion, the result of this multifaceted project is a sophisticated phishing URL detection system. Beginning with the preparation of a meticulously curated dataset, the project

proceeds to model selection, serialization, and user interface development. Feature extraction acts as the bridge between user input and model prediction, culminating in a user-centric and effective system. The transformational journey encapsulates the essence of efficient and resilient cybersecurity, offering users a robust shield against the pernicious threats lurking in the digital realm.

CONCLUSION

In culmination, this project triumphantly gives rise to a sophisticated phishing URL detection system that elegantly intertwines the intricate realms of machine learning and web technology. The meticulous model selection process, informed by rigorous evaluation metrics, engenders an environment of optimal accuracy, signifying the project's unwavering commitment to precision in detecting malicious URLs. Furthermore, the ingeniously devised user interface, meticulously crafted using the Flask web framework, ushers in a new era of user interactions, elevating the user experience to unparalleled heights. As a testament to the project's forward-thinking approach, the implementation of Pickle serialization imparts a transcendent dimension of efficiency. This serialization process encapsulates the model's state and parameters, endowing it with the ability to seamlessly traverse time, thereby obviating the arduous need for recurrent training. In essence, the culmination of these intricate components not only exemplifies the successful amalgamation of advanced technologies but also heralds a new dawn in the realm of phishing detection, where precision, usability, and efficiency converge to shield the digital realm from the multifarious threats it faces.

REFERENCES

[1] FBI, “Ic3 annual report released.”
[2] APWG, “Phishing activity trends report.”
[3] V. B. et al, “study on phishing attacks,” International Journal of Computer Applications, 2018. [4] I.-F. Lam, W.-C. Xiao, S.-C. Wang, and K.-T. Chen, “Counteracting phishing page polymorphism: An image layout analysis approach,” in International Conference on Information Security and Assurance, pp. 270–279, Springer, 2009.
[5] W. Jing, “Covert redirect vulnerability,” 2017.
[6] K. Krombholz, H. Hobel, M. Huber, and E. Weippl, “Advanced social engineering attacks,” Journal of Information Security and applications, vol. 22, pp. 113–122, 2015.
[7] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham, “School of phish: a real-world evaluation of antiphishing training,” in Proceedings of the 5th Symposium on Usable Privacy and Security, pp. 1–12, 2009.
[8] R. C. Dodge Jr, C. Carver, and A. J. Ferguson, “Phishing for user security awareness,” computers & security, vol. 26, no. 1, pp. 73–80, 2007.
[9] R. Dhamija, J. D. Tygar, and M. Hearst, “Why phishing works,” in Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 581–590, 2006.
[10] C. Ludl, S. McAllister, E. Kirda, and C. Kruegel, “On the effectiveness of techniques to detect phishing sites,” in International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, pp. 20–39, Springer, 2007.