Phishing Website Detection: A Dataset-Centric Approach for Enhanced Security

6 authors, including:

Sultan Ahmad
Prince Sattam Bin Abdulaziz University
165 PUBLICATIONS   **2,352** CITATIONS

Md Alimul Haque
Veer Kunwar Singh University
117 PUBLICATIONS   **965** CITATIONS

Hikmat A. M. Abdeljaber
Applied Science Private University
39 PUBLICATIONS   **340** CITATIONS

Jabeen Nazeer
Prince Sattam Bin Abdulaziz University
29 PUBLICATIONS   **243** CITATIONS

ORIGINAL

Phishing Website Detection: A Dataset-Centric Approach for Enhanced Security

Detección de sitios web de phishing: Un enfoque centrado en conjuntos de datos para mejorar la seguridad

Sultan Ahmad[1,2] ✉, Alimul Haque[3] ✉, Hikmat A. M. Abdeljaber[4,5] ✉, M. U. Bokhari[6] ✉, Jabeen Nazeer[1], B. K. Mishra[7] ✉

[1]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, P.O.Box. 151, Alkharj 11942, Saudi Arabia.

[2]School of Computer Science and Engineering, Lovely Professional University, Phagwara, 144411, Punjab, India.

[3]Department of Computer Science, Veer Kunwar Singh University, Ara- 802301, India.

[4]Department of Computer Science, Faculty of Information Technology, Applied Science Private University, Amman, Jordan.

[5]MEU Research Unit, Middle East University, Amman 11831, Jordan.

[6]Department of Computer Science, Aligarh Muslim University, Aligarh - 202002, India.

[7]P.G. Department of Physics, Veer Kunwar Singh University, Ara- 802301, India.

ABSTRACT

Introduction: phishing involves cybercriminals creating fake websites that appear to be real sites with the aim of obtaining personal information. With the increasing sophistication of phishing websites, machine learning today provides a useful approach to scan and counter such attacks.
Objective: in this study, we seek to apply machine learning algorithms on the dataset - Phishing_Legitimate_full.csv – which consists of phishing websites and genuine websites that have been labeled.
Method: this paper aims to identify the most effective feature selection method for predicting phishing websites.
Result: the findings highlight the potential of machine learning in enhancing cybersecurity by automating threat detection and intelligence. Phishing attacks rely on social engineering strategies to present deceptive links as trustworthy sources, deceiving individuals into sharing confidential data.
Conclusion: this study explores the utilization of curated datasets and machine learning algorithms to develop adaptive and efficient phishing detection mechanisms, providing a robust defense against such malicious activities.

Keywords: Machine Learning; Phishing Attacks; Cybersecurity.

RESUMEN

Introducción: el phishing consiste en que los ciberdelincuentes crean sitios web falsos que parecen ser sitios reales con el objetivo de obtener información personal. Con la creciente sofisticación de los sitios web de phishing, el aprendizaje automático proporciona hoy en día un enfoque útil para escanear y contrarrestar este tipo de ataques.
Objetivo: en este estudio, buscamos aplicar algoritmos de aprendizaje automático en el conjunto de datos - Phishing_Legitimate_full.csv - que consiste en sitios web de phishing y sitios web genuinos que han sido etiquetados.

Método: el objetivo de este artículo es identificar el método de selección de características más eficaz para predecir sitios web de phishing.
Resultado: los resultados ponen de manifiesto el potencial del aprendizaje automático para mejorar la ciberseguridad mediante la automatización de la detección de amenazas y la inteligencia. Los ataques de phishing se basan en estrategias de ingeniería social para presentar enlaces engañosos como fuentes fiables, engañando a las personas para que compartan datos confidenciales.
Conclusión: este estudio explora la utilización de conjuntos de datos curados y algoritmos de aprendizaje automático para desarrollar mecanismos de detección de phishing adaptables y eficientes, proporcionando una defensa robusta contra tales actividades maliciosas.

Palabras clave: Aprendizaje Automático; Ataques de Phishing; Ciberseguridad.

INTRODUCTION

As phishing attacks continue to rise in frequency and sophistication, the need for automated detection systems has become increasingly urgent. Phishing websites deceive users into disclosing sensitive personal and financial information, often leading to significant financial losses. Traditional detection techniques, which relied on heuristic approaches and static blacklists, have struggled to keep up with the dynamic tactics employed by cybercriminals.[1] As a result, more advanced, adaptive detection methods are required to counter these ever-evolving threats.[2] Phishing attacks often involve emails containing deceptive URLs (Uniform Resource Locators) designed to mislead recipients. Unsuspecting individuals may unknowingly click on these fraudulent links, ultimately compromising sensitive information. As highlighted by Anupam and Kar (2021),[3] attackers employ various strategies to ensnare their targets. These include techniques such as phishing via email, deploying malware, spear phishing, whaling, smishing, and vishing. Each method exploits specific weaknesses or characteristics of individuals or organizations, making them vulnerable to such malicious schemes.[4]

In recent years, the emergence of machine learning (ML) has opened up new possibilities for improving phishing detection systems. By analyzing patterns and characteristics of websites, ML models can differentiate between legitimate and phishing websites with greater precision and speed.[5] These models are trained on extensive datasets of phishing and legitimate sites, allowing them to learn from the complex features of each category. The aim of modern research is to harness the power of ML to create systems that can adapt to new phishing techniques, providing more reliable real-time protection.[6] The key advantage of ML-based phishing detection systems lies in their ability to learn and evolve with the threat landscape. Unlike traditional methods, which often rely on static features or signatures, ML models can dynamically adapt to new phishing strategies by analyzing various website attributes. Features such as URL structure, HTML content, domain age, and hosting information are used to identify phishing sites. Researchers have explored various ML algorithms for this task, including Decision Trees, Support Vector Machines (SVM), Random Forests, and Neural Networks.

One significant breakthrough in phishing detection using ML has been the focus on ensemble learning, which combines the predictions of multiple algorithms to increase overall accuracy. For instance, a recent study by Mohammad et al. (2022)[7] demonstrated that ensemble-based models such as Gradient Boosting and Random Forests performed significantly better than individual classifiers in detecting phishing websites. These models achieved high accuracy rates by leveraging multiple weak learners and combining their predictions to produce a stronger overall result. Other research has emphasized the role of deep learning techniques such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) in identifying phishing websites through visual and textual analysis of website content. The structure of phishing websites has been extensively studied to identify patterns and trends that are common across phishing campaigns. Phishing sites often exhibit telltale signs, such as misspellings, shortened URLs, and the use of fake logos and brand names. By understanding these common characteristics, researchers can design features that help ML models distinguish between phishing and legitimate sites more effectively. Despite the success of ML models, there are still challenges in achieving perfect phishing detection. One issue is the imbalance in the dataset, where phishing websites are often outnumbered by legitimate ones. This imbalance can lead to models that are biased towards the majority class (legitimate sites) and fail to detect a significant number of phishing attempts. Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and cost-sensitive learning are being explored to address this problem and improve the detection of phishing sites.[8,9]

Phishing detection methods are generally classified into two broad categories. The first category is known as user education or awareness, which focuses on teaching users to recognize and distinguish between phishing and legitimate emails. The second category involves software-based detection, which utilizes techniques such as blacklists, heuristic analysis, visual similarity checks, and increasingly, machine learning (ML) approaches to identify phishing attempts. Machine Learning, a vital area within computer science and artificial intelligence

(AI), aims to replicate the way humans learn by uncovering meaningful patterns in data through specialized algorithms.[10,11] ML has found widespread application across numerous fields due to its ability to process large datasets and identify hidden insights. For instance, it has been effectively applied in domains such as medical diagnosis, where it aids in detecting diseases and conditions with higher accuracy. In cybersecurity, ML is extensively used for malware prediction, improving the ability to identify harmful software before it causes significant damage. Similarly, ML models are employed in weather forecasting to predict climatic conditions with greater precision. Fraud detection, which is critical in financial sectors, has been significantly enhanced by ML algorithms, enabling more efficient identification of unusual patterns or transactions indicative of fraud. ML has been applied in areas such as scene classification, which is used in image processing and computer vision to categorize different objects or scenarios in photos and videos. In the context of phishing detection, ML serves as a powerful tool by analyzing diverse features such as email structure, URL patterns, and website content to differentiate between phishing and legitimate activities.[12,13] The adoption of ML in this field allows for a more dynamic and accurate approach to identifying phishing attacks, making it an invaluable component of modern cybersecurity efforts.[14,15]

The structure of the paper is outlined as follows: The second section provides a comprehensive review of existing research and data collection techniques relevant to the study. Section 3 outlines the methodology employed in the research. An in-depth discussion is included on how the data for the experiments was gathered and analyzed. In Section 4, the paper presents the results of experiments aimed at detecting and mitigating phishing attacks. Finally, Section 5 summarizes the key findings, explores future prospects, and provides recommendations to enhance cybersecurity measures against phishing threats.

Literature review

Research studies focused on enhancing cybersecurity through machine learning for phishing detection as shown in table 1. The rapid evolution of phishing techniques has necessitated advanced and automated mechanisms for detection, with machine learning emerging as a promising solution. The studies summarized in the table provide an in-depth exploration of various machine learning methodologies applied to phishing detection, each showcasing unique approaches and significant results.

Several researchers, such as Mohammad Alauthman et al., and Rao and Pais,[16] emphasized the use of classical algorithms like Decision Tree, Naïve Bayes, and Gradient Boosting for detecting phishing websites. Their work demonstrated the efficacy of these models, with accuracies surpassing 95 %, showcasing their reliability in traditional phishing scenarios. Similarly, Abdulkarim R. Muda et al. and Chiew et al.[17] highlighted the effectiveness of ensemble models like Random Forest, achieving accuracy rates above 96 %, solidifying their prominence in phishing detection tasks. More recent approaches incorporate deep learning techniques, as demonstrated by Zhang et al. and Sahingoz et al.[18] utilizing Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models. These methods excelled in mobile and natural language processing-based phishing detection, reflecting their adaptability to modern attack vectors. Furthermore, hybrid models like the one proposed by Hong et al. achieved remarkable accuracy of 98,5 %, underlining the potential of integrating multiple machine learning techniques for robust defense. Other studies focused on domain-specific applications, such as email-based phishing (Verma and Das)[19] and dynamic feature extraction, offering targeted solutions to address specific vulnerabilities. Collectively, these works demonstrate the adaptability and scalability of machine learning in phishing detection, paving the way for future research to refine these models for enhanced cybersecurity resilience.

Table 1. Cybersecurity through Machine Learning for phishing detection					
Researchers	**Research Field**	**Machine Learning Techniques**	**Data Sets Used**	**Number of Features**	**Attained Results**
Mohammad Alauthman et al.[16]	Phishing Website Detection	Decision Tree, Naïve Bayes	UCI Repository Dataset	30	95,5 % accuracy with Decision Tree
NZ Jhanjhi, IA Shah et al.[20]	Comparative ML for Phishing	SVM, Random Forest, k-NN	PhishTank and Alexa	20	Random Forest achieved 98,2 % accuracy
Mohammad Reza Ebrahimi et al.[21]	Phishing URL Detection	Logistic Regression, SVM	Private dataset	40	SVM achieved 97 % accuracy
Chiew et al.[22]	URL-based Phishing Detection	Random Forest, k-NN	PhishTank	24	Random Forest achieved 96,7 % accuracy
Verma and Das[19]	Email Phishing Classification	Naïve Bayes, Neural Network	Kaggle Phishing Email Dataset	48	Neural Network achieved 98 % accuracy
Jain and Gupta[23]	Anti-phishing System	Ensemble Learning	PhishTank, Legitimate URLs	30	Ensemble method achieved 96,9 % accuracy
Zhang et al.[24]	Mobile Phishing Detection	Deep Learning (CNN)	Mobile-specific phishing dataset	18	CNN achieved 94,3 % accuracy

Rao and Pais[25]	Phishing URL Identification	Gradient Boosting, Random Forest	UCI Repository Dataset	20	Gradient Boosting achieved 97,1 % accuracy
Mohammad et al.[26]	Blacklist-based Detection	Random Forest, Neural Network	Public phishing and legitimate URL datasets	15	Neural Network achieved 98,3 % accuracy
Pham et al.[27]	Dynamic Feature Extraction	SVM, Logistic Regression	Dynamic phishing URL dataset	25	SVM achieved 95,8 % accuracy
Anupam and Kar[3]	Multi-vector Phishing Defense	k-NN, Decision Tree	PhishTank	30	k-NN achieved 94,5 % accuracy
Marchal et al.[28]	Heuristic-based Detection	Naïve Bayes	Legitimate and Phishing Emails Dataset	50	Naïve Bayes achieved 93 % accuracy
Sahingoz et al.[29]	NLP in Phishing Detection	Deep Learning (LSTM)	Twitter and PhishTank	10	LSTM achieved 96 % accuracy
Basnet et al.[30]	URL Classification	Logistic Regression, Decision Tree	Phishing Dataset from UCI	30	Decision Tree achieved 95,3 % accuracy
Hong et al.[31]	Phishing Defense Framework	Hybrid Ensemble	Multi-source phishing dataset	45	Hybrid ensemble achieved 98,5 % accuracy

This table highlights the diversity of research methods, machine learning models, datasets, and achieved results, showcasing the advancement of phishing detection systems through machine learning. Although many studies have examined the use of machine learning algorithms for phishing detection, a significant gap remains in the integration of these techniques with expert-curated datasets. Most existing research prioritizes machine learning models trained solely on raw data, neglecting the valuable insights that curated lists could offer. This disconnect highlights an underexplored opportunity to enhance phishing detection systems by combining the predictive power of machine learning with the domain-specific expertise embedded in these curated resources. Bridging this gap could lead to more robust and accurate approaches to combating phishing threats.

METHOD
The methodology used in this research consists of several steps:

Dataset Overview
The dataset contains two classes of websites: phishing and legitimate.[32] Each instance in the dataset includes several features (such as URL length, presence of suspicious characters, etc.) that characterize phishing behavior. Summarize the dataset in a tabular form based on its phishing detection purpose in table 2. The dataset is first examined for any missing or outlier data. The dataset contains 10 000 entries and 50 columns. These columns represent various attributes that can be used to identify phishing websites. Some of the notable columns include:
- NumDots: Number of dots in the URL.
- SubdomainLevel: The level of the subdomain in the URL.
- UrlLength: The total length of the URL.
- NumDash: Number of dashes ('-') in the URL.
- NumUnderscore: Number of underscores ('_') in the URL.
- NoHttps: A binary feature indicating whether HTTPS is absent in the URL.
- IpAddress: Indicates whether an IP address is used in the URL instead of a domain name.
- PctExtHyperlinks: Percentage of external hyperlinks in the website.
- PctExtResourceUrls: Percentage of external resource URLs in the website.
- CLASS_LABEL: The target variable indicating whether the website is phishing (1) or legitimate (0).

Data Description
The dataset compiled serves the purpose of developing and accessing various classification techniques aimed at detecting phishing websites. This is accomplished by analyzing key features such as the properties of the uniform resource locator (URL), URL resolution metrics, and information from external services.

Dataset Insights from Visualizations
I will now generate specific visualizations to explore patterns in the dataset, particularly focusing on the distribution of key features and their relationship to the target variable (CLASS_LABEL). Let's start by visualizing the following:
- The distribution of URL lengths for phishing and legitimate websites.
- The correlation heatmap of all features to identify the most impactful ones.

•   The count of phishing and legitimate websites based on various attributes like subdomain level, presence of IP addresses, and HTTPS usage.

Distribution of Phishing Vs Legitimate URLs
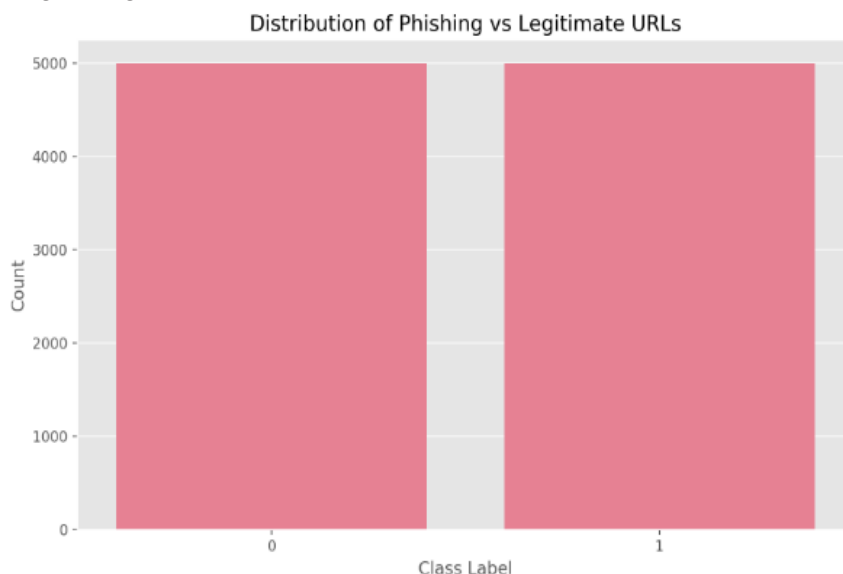
Figure 1. Distribution of Phishing Vs Legitimate URLs[(32)]

The URL length is simply the number of characters in a URL. It can be an important feature in distinguishing between phishing and legitimate URLs, as phishing URLs often have certain characteristics, such as being unusually long or containing many subdomains. By visualizing the distribution of URL lengths, we can observe:

•   The range of lengths present in the dataset.
•   The frequency of URLs of different lengths.
•   Any patterns or trends that may indicate a relationship between URL length and the likelihood of being classified as phishing or legitimate.

Overall, analyzing URL length distribution helps in understanding how this feature can contribute to the classification of URLs and can be a valuable part of a phishing detection model.

URL Length Distribution by Class
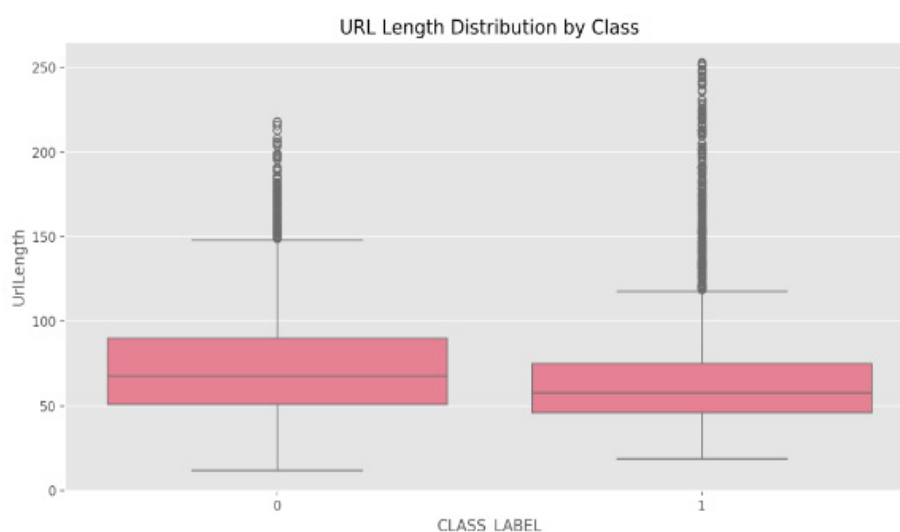
Figure 2. URL Length Distribution by Class[(32)]

This likely refers to a visual representation of the data, such as a histogram, box plot, or bar chart, that illustrates the distribution of URL lengths for each class. The figure would help in visually comparing how URL

lengths differ across the various classes.

Correlation Heatmap of Key Features

A correlation heatmap is a graphical representation of the correlation coefficients between multiple variables in a dataset. Correlation coefficients measure the strength and direction of the relationship between pairs of features (variables). Values range from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no correlation.

A heatmap uses color coding to represent these correlation values, making it easy to visualize which features are positively or negatively correlated with each other. Explaining a correlation heatmap would involve identifying which features have strong correlations (either positive or negative), discussing any surprising or expected relationships, and considering how these correlations might inform further analysis or model building. Both tasks involve analyzing and interpreting data visualizations to gain insights into the structure and relationships within a dataset.

Figure 3. Correlation Heatmap of Key Features[32]

A correlation heatmap is a graphical representation of the correlation coefficients between multiple variables in a dataset. Correlation coefficients measure the strength and direction of the relationship between pairs of features (variables). Values range from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no correlation.

A heatmap uses color coding to represent these correlation values, making it easy to visualize which features are positively or negatively correlated with each other. Explaining a correlation heatmap would involve identifying which features have strong correlations (either positive or negative), discussing any surprising or expected relationships, and considering how these correlations might inform further analysis or model building. Both tasks involve analyzing and interpreting data visualizations to gain insights into the structure and relationships within a dataset.
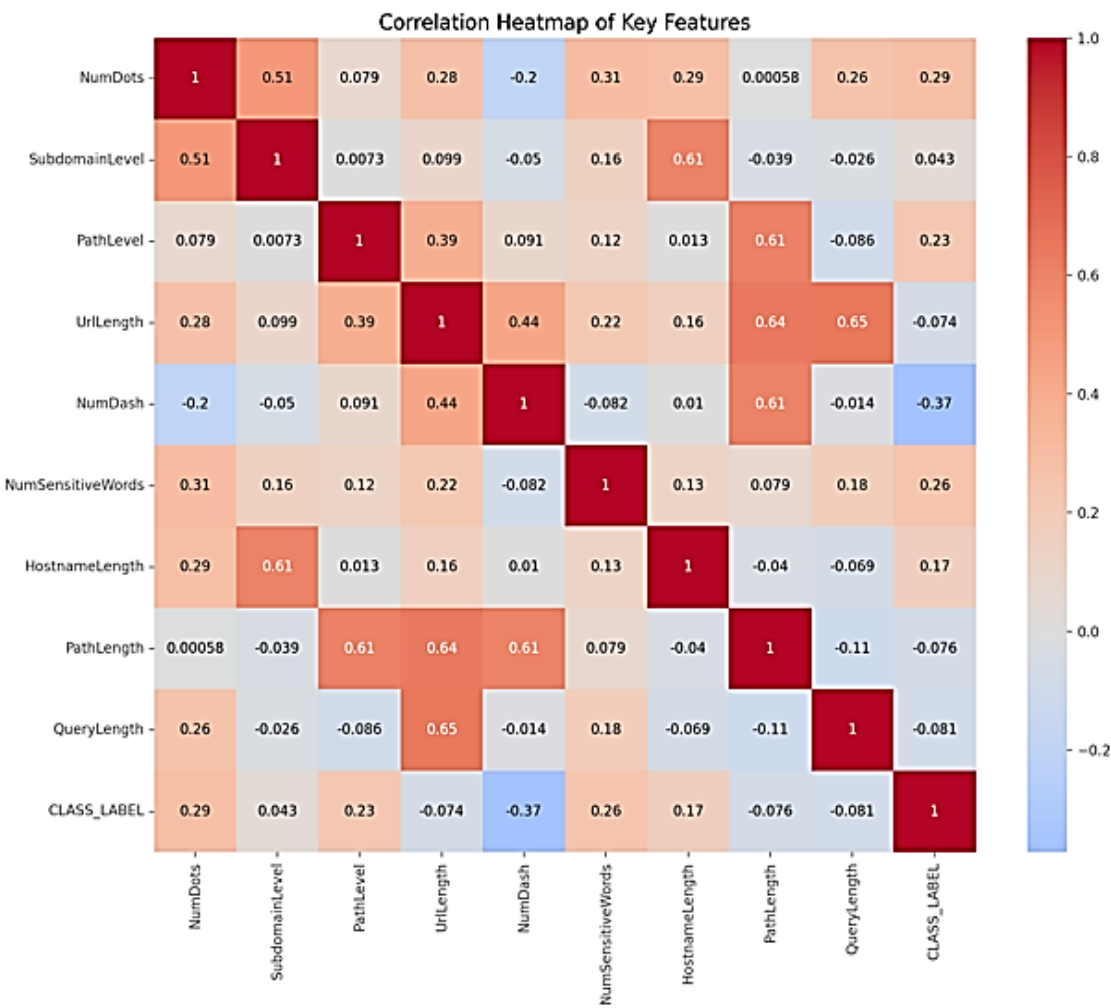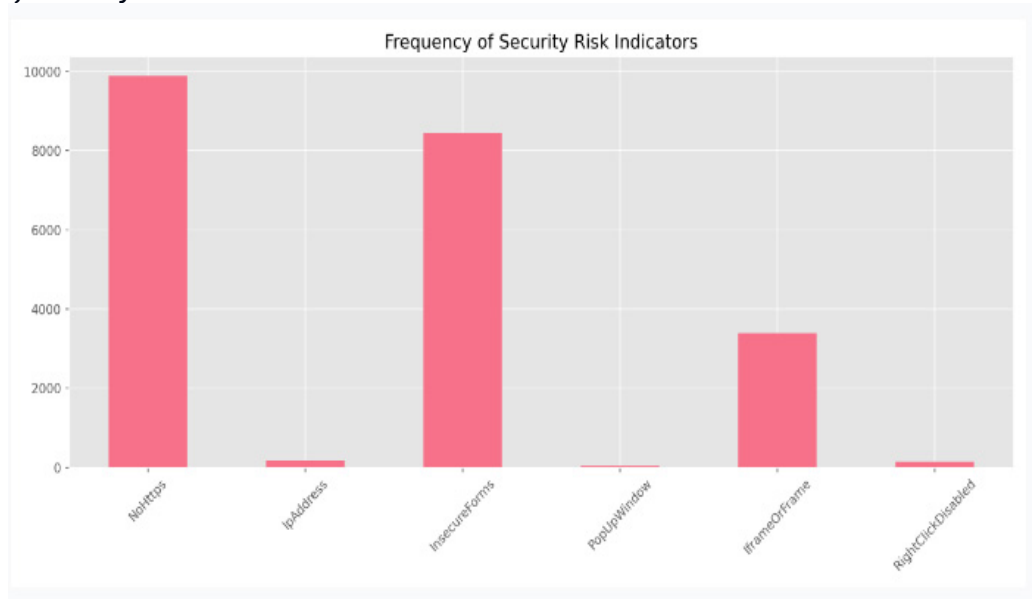
Frequency of Security Risk Indicators

Figure 4. Frequency of Security Risk Indicators[32]

The "Frequency of Security Risk Indicators" refers to the count of specific features in the dataset that are associated with potential security risks in URLs. These indicators are characteristics that may suggest whether a URL is more likely to be phishing or legitimate.
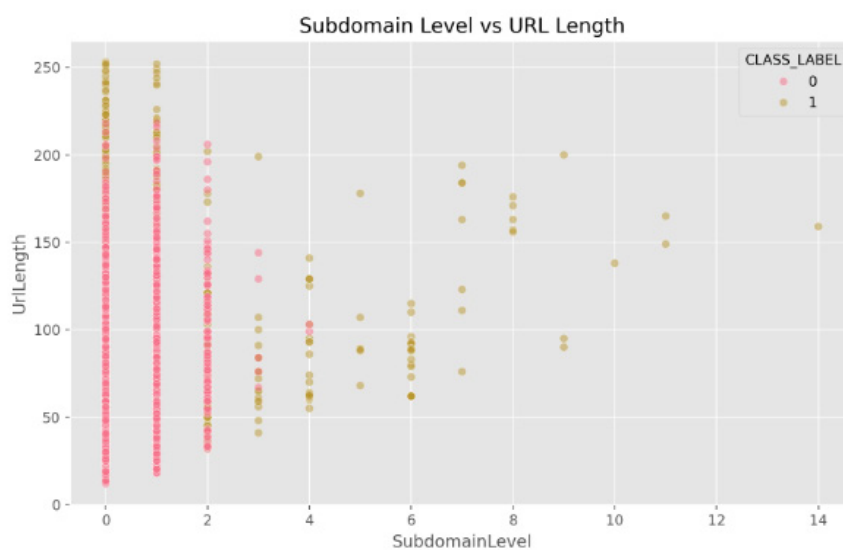
Subdomain Level Vs URL Length

Figure 5. Subdomain Level Vs URL Length[32]

The scatter plot that visualizes "Subdomain Level vs URL Length" aims to show how the number of subdomains in a URL correlates with its overall length. By plotting these two variables against each other, we can observe patterns or trends that may indicate whether certain combinations of subdomain levels and URL lengths are more likely to be associated with phishing or legitimate URLs.

RESULT AND DISCUSSION

The dataset employs a perfectly balanced sampling design with equal representation of legitimate and phishing URLs Feature Engineering and Feature Types. All features were engineered as numerical variables, enabling direct mathematical analysis.

The figure depicts boxplots of key URL metrics categorized by class labels (0 for legitimate websites and 1 for phishing websites). The metrics include UrlLength, NumDots, and PathLength. For UrlLength and PathLength, phishing websites (class 1) tend to have longer URLs and paths compared to legitimate sites (class 0). NumDots

shows a relatively small variation but indicates that phishing websites tend to have more dots in the URL. The distribution suggests that phishing websites generally have higher values for these URL-related features, which could be used to distinguish them from legitimate websites.
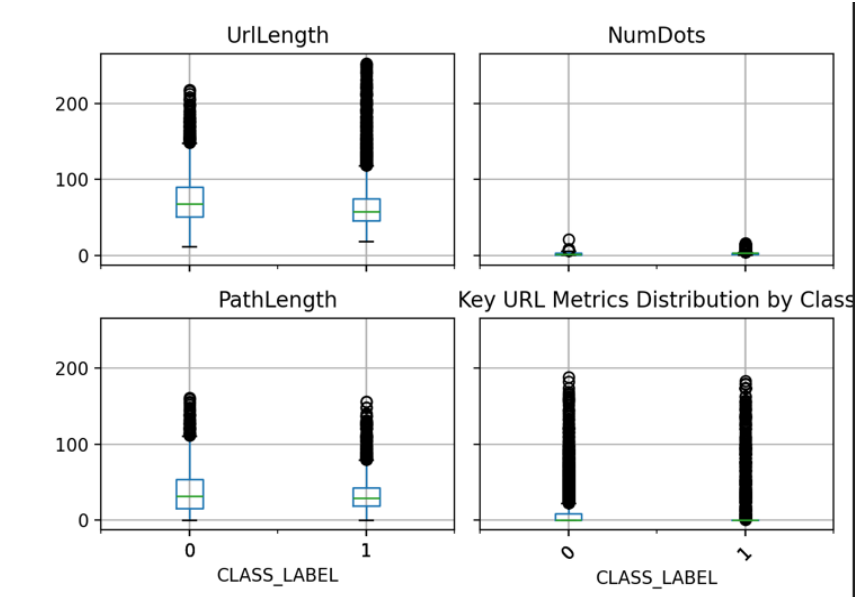
Figure 6. UrlLength and PathLength, phishing websites (class 1) tend to have longer

URLs and paths compared to legitimate sites (class 0)[32]

Security Feature Implementation
It refers to the various attributes or characteristics within the dataset that indicate the presence or absence of security measures in a URL. These features can help assess the potential risk associated with a URL, particularly in the context of phishing detection. Here are some examples of security features that might be included in the dataset:

NoHttps	0.9888
IpAddress	0.0172
InsecureForms	0.844
PopUpWindow	0.0049
IframeOrFrame	0.3396
RightClickDisabled	0.014

Figure 7. Potential risks associated with a URL[32]

By analyzing these security features, one can gain insights into the potential risks associated with a URL and improve the effectiveness of phishing detection models.

CONCLUSION
Through the examination of metrics such as URL length, number of dots, and path length, clear differences in distribution between phishing and legitimate sites were observed. The dataset enabled the application of various machine learning algorithms, and a detailed feature selection analysis helped identify optimal approaches for phishing detection. The comparative results based on performance metrics like Accuracy, Precision, and Recall provided a clear understanding of which classifiers and feature selection methods are most effective in predicting phishing sites. Future research can focus on integrating additional features, such as content-based metrics or deeper URL analysis, and experimenting with ensemble learning techniques to further improve detection accuracy. Furthermore, there is potential in incorporating real-time data to refine these models and enhance the robustness of phishing detection systems for practical deployment. This research contributes meaningfully to the ongoing efforts to mitigate phishing attacks, offering insights into efficient methods of identifying and countering this growing cyber threat.

REFERENCES

1. Wu L, Du X, Wu J. Effective defense schemes for phishing attacks on mobile computing platforms. IEEE Transactions on Vehicular Technology. 2015;65(8):6678-91.

2. Ahmad S, Jha S, Alam A, Alharbi M, Nazeer J. Analysis of Intrusion Detection Approaches for Network Traffic Anomalies with Comparative Analysis on Botnets (2008-2020). Security and Communication Networks. 2022;2022.

3. Anupam S, Kar AK. Phishing website detection using support vector machines and nature-inspired optimization algorithms. Telecommunication Systems. 2021;76(1):17-32.

4. Xiang G, Hong JI. A hybrid phish detection approach by identity discovery and keywords retrieval. En: Proceedings of the 18th international conference on World wide web. 2009. p. 571-80.

5. Ahmad S, Afzal MM. A Study and Survey of Security and Privacy issues in Cloud Computing. International Journal of Engineering Research & Technology (IJERT), ISSN. :181-2278.

6. Whig V, Othman B, Gehlot A, Haque MA, Qamar S, Singh J. An Empirical Analysis of Artificial Intelligence (AI) as a Growth Engine for the Healthcare Sector. En: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE; 2022. p. 2454-7.

7. Haakenstad A, Irvine CMS, Knight M, Bintz C, Aravkin AY, Zheng P, et al. Measuring the availability of human resources for health and its relationship to universal health coverage for 204 countries and territories from 1990 to 2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet. 2022;399(10341):2129-54.

8. Hossain MdA, Haque MA, Ahmad S, Abdeljaber HAM, Eljialy AEM, Alanazi A, et al. AI-enabled approach for enhancing obfuscated malware detection: a hybrid ensemble learning with combined feature selection techniques. International Journal of System Assurance Engineering and Management. 2024;

9. Haque MA, Ahmad S, Abboud AJ, Hossain MA, Kumar K, Haque S, et al. 6G wireless Communication Networks: Challenges and Potential Solution. International Journal of Business Data Communications and Networking (IJBDCN). 2024;19(1):1-27.

10. Haque MA, Haque S, Kumar K, Singh NK. A Comprehensive Study of Cyber Security Attacks, Classification, and Countermeasures in the Internet of Things. En: Digital Transformation and Challenges to Data Security and Privacy. IGI Global; 2021. p. 63-90.

11. Haque MA, Ahmad S, Sonal D, Abdeljaber HAM, Mishra BK, Eljialy AEM, et al. Achieving Organizational Effectiveness through Machine Learning Based Approaches for Malware Analysis and Detection. Data and Metadata. 2023;2:139.

12. Haque MA, Ahmad S, John A, Mishra K, Mishra BK, Kumar K, et al. Cybersecurity in Universities: An Evaluation Model. SN Computer Science. 2023;4(5):569.

13. Wei W, Nazura Bt. AM, Bin Abd Rahman MR. Research on the Issues and Paths of Citizen Privacy Protection in China in the Era of Big Data. Salud, Ciencia y Tecnología [Internet]. 8 de agosto de 2024 [citado 12 de enero de 2025];4. Disponible en: https://sct.ageditor.ar/index.php/sct/article/view/948

14. Haque A, Raza S, Ahmad S, Hossain A, Abdeljaber HAM, Eljialy AEM, et al. Implication of Different Data Split Ratio on the Performance of Model in Price Prediction of Used Vehicles Using Regression Analysis. Data and Metadata. 2024;3:425.

15. Rosario Quiroz FJ, Calla Vásquez KM, Ochoa Tataje FA, Morí Holguín JY, Villanueva-Batallanos M. Systemic review of studies of cyberbullying in Hispanic American adolescents. Salud, Ciencia y Tecnología [Internet]. 14 de febrero de 2024 [citado 12 de enero de 2025];4:800. Disponible en: https://sct.ageditor.ar/index.php/sct/article/view/623

16. Alauthman M, Aslam N, Al-Kasassbeh M, Khan S, Al-Qerem A, Choo KKR. An efficient reinforcement

learning-based Botnet detection approach. Journal of Network and Computer Applications. 2020;150:102479.

17. Chiew KL, Yong KSC, Tan CL. A survey of phishing attacks: Their types, vectors and technical approaches. Expert Systems with Applications. 2018;106:1-20.

18. Sahingoz OK. Networking models in flying ad-hoc networks (FANETs): Concepts and challenges. Journal of Intelligent & Robotic Systems. 2014;74:513-27.

19. Das A, Baki S, El Aassal A, Verma R, Dunbar A. SoK: a comprehensive reexamination of phishing research from the security perspective. IEEE Communications Surveys & Tutorials. 2019;22(1):671-708.

20. Jhanjhi NZ, Shah IA. Navigating Cyber Threats and Cybersecurity in the Logistics Industry. IGI Global; 2024.

21. Keyvanpour MR, Javideh M, Ebrahimi MR. Detecting and investigating crime by means of data mining: a general crime matching framework. Procedia Computer Science. 2011;3:872-80.

22. Chiew KL, Chang EH, Tiong WK. Utilisation of website logo for phishing detection. Computers & Security. 2015;54:16-26.

23. Jain AK, Gupta BB. A survey of phishing attack techniques, defence mechanisms and open research challenges. Enterprise Information Systems. 2022;16(4):527-65.

24. Zhang J, Pan Y, Wang Z, Liu B. URL based gateway side phishing detection method. En: 2016 IEEE Trustcom/BigDataSE/ISPA. IEEE; 2016. p. 268-75.

25. Rao RS, Vaishnavi T, Pais AR. CatchPhish: detection of phishing websites by inspecting URLs. Journal of Ambient Intelligence and Humanized Computing. 2020;11:813-25.

26. Mohamed G, Visumathi J, Mahdal M, Anand J, Elangovan M. An effective and secure mechanism for phishing attacks using a machine learning approach. Processes. 2022;10(7):1356.

27. Nguyen TP, Pham CC, Ha SVU, Jeon JW. Change detection by training a triplet network for motion feature extraction. IEEE Transactions on Circuits and Systems for Video Technology. 2018;29(2):433-46.

28. Karunakaran B, Misra D, Marshall K, Mathrawala D, Kethireddy S. Closing the loop—Finding lung cancer patients using NLP. En: 2017 IEEE international conference on big data (big data). IEEE; 2017. p. 2452-61.

29. Sahingoz OK, Buber E, Demir O, Diri B. Machine learning based phishing detection from URLs. Expert Systems with Applications. 2019;117:345-57.

30. Basnet RB, Sung AH. Learning to Detect Phishing Webpages. J Internet Serv Inf Secur. 2014;4(3):21-39.

31. Hong J, Kim H, Oh S, Im Y, Jeong H, Kim H, et al. Combating phishing and script-based attacks: a novel machine learning framework for improved client-side security. The Journal of Supercomputing. 2025;81(1):1-24.

32. Phishing Detection Dataset [Internet]. Disponible en: https://www.kaggle.com/datasets/sharmi3754/phishing-detection-dataset

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHOR CONTRIBUTIONS

Conceptualization: Sultan Ahmad, Alimul Haque, Hikmat A. M. Abdeljaber, Jabeen Nazeer.

Investigation: Alimul Haque, Sultan Ahmad, M. U. Bokhari, B.K. Mishra.

Methodology: Sultan Ahmad, Alimul Haque, M. U. Bokhari, B. K. Mishra, Hikmat A. M. Abdeljaber, Jabeen Nazeer.

Writing - original draft: Sultan Ahmad, Alimul Haque, M.U. Bokhari, Hikmat A. M. Abdeljaber, Jabeen Nazeer, B. K. Mishra.

Writing - review and editing: Sultan Ahmad, Alimul Haque, M.U. Bokhari, Hikmat A. M. Abdeljaber, Jabeen Nazeer, B. K. Mishra.