Issues in Information Systems
Volume 26, Issue 2, pp. 150-161, 2025

DOI: https://doi.org/10.48009/2_iis_112

Analyzing website characteristics and their impact on web traffic and legitimacy classification for phishing detection: A structural equation modeling approach

Angel, Ojeda-Castro, *Universidad Ana G. Méndez,Gurabo Campus, ut_aojeda@uagm.edu*
Melanie, Pérez, *Universidad Ana G. Méndez,Gurabo Campus, mperez663@email.uagm.edu*
Cristhian, Pagán, *Universidad Ana G. Méndez,Gurabo Campus, cpagan116@email.uagm.edu*
Rafael Padilla-Vega, *Universidad Ana G. Méndez,Gurabo Campus, padillar1@uagm.edu*
José, Cruz, *University of Puerto Rico, Cayey Campus, jose.cruz199@upr.edu*

Abstract

Phishing attacks continue to pose a significant threat in the digital age by leveraging deceptive websites that imitate legitimate platforms to gain user trust and steal sensitive information. This study explores the relationship between various website characteristics and their effect on web traffic and legitimacy classification using a dataset of over 11,000 websites from the Kaggle "Phishing Website Detector." Employing structural equation modeling (SEM), the analysis focuses on 11 specific features, including pop-up windows, iframe redirection, domain age, DNS recording, and status bar customization. Results reveal that DNS recording, iframe redirection, and pop-up usage are positively associated with higher web traffic, which in turn is linked to legitimate websites. Conversely, features such as status bar customization, older domain age, and a low number of backlinks are more commonly found in phishing sites. Notably, traditionally cited indicators like PageRank and disabling right-click functionality showed no significant impact on traffic or legitimacy. The model's R^2 value of 0.130 suggests that while these variables are relevant, additional behavioral and dynamic data may be required to improve predictive power. This research enhances phishing detection strategies by identifying meaningful indicators and emphasizes the future need for real-time analytics and machine learning in cybersecurity defense systems.

Keywords: cybersecurity, iframe, pop-up windows, machine, phishing, web traffic

Introduction

The proliferation of sophisticated phishing attacks poses a significant and evolving threat to individuals and organizations in the digital age. As technology advances and online interactions become increasingly integral to daily life, malicious actors continually refine their techniques to deceive users and exploit vulnerabilities. This escalating cyber threat landscape underscores the critical need for robust and effective methods of phishing detection. Understanding how deceptive websites operate and identifying the characteristics that distinguish them from legitimate ones are paramount in mitigating the potential financial losses, data breaches, and reputational damage associated with successful phishing campaigns.

Despite the existence of various anti-phishing tools and user awareness initiatives, phishing attacks remain a prevalent and successful form of cybercrime. Recent studies have highlighted the limitations of current detection mechanisms and underscore the need for more comprehensive approaches. For instance, Mishra and Varshney (2025) emphasize the effectiveness of brand domain identification features in phishing detection, achieving high accuracy rates using machine learning classifiers. Similarly, Guo et al. (2025) propose a graph-based machine learning model that integrates URL structure and network-level features, demonstrating significant improvements in detection performance. Moreover, advancements in visual similarity methods, as discussed by researchers in 2024, offer promising avenues for identifying phishing websites based on their resemblance to legitimate sites.

Therefore, this research aims to address the existing gaps by investigating the relationship between a range of website parameters—including status bar customization, the disabling of right-click functionality, the use of pop-up windows, iframe redirection, domain age, DNS records, PageRank, Google Index status, and the number of links pointing to a page, and the classification of a website as either phishing or legitimate. By analyzing these features within a substantial dataset of website URLs, this study seeks to identify key indicators that can enhance the accuracy and effectiveness of phishing detection strategies, ultimately contributing to a safer online environment for users (Wang et al., 2024).

Background and Literature Review

In this research, web traffic is considered an independent variable due to its significant influence on a website's visibility, user engagement, and overall performance. Web traffic refers to the volume of user visits to a website over a specific period, encompassing sessions initiated through various channels, including search engines, social media platforms, email links, and direct URL entries. Key metrics for measuring web traffic include page views, unique visitors, average session duration, and bounce rates. Understanding web traffic is crucial as it reflects how effectively a website attracts and retains visitors. High traffic volumes often correlate with increased opportunities for customer engagement, lead generation, and revenue growth. Analyzing traffic patterns enables businesses to identify which content or strategies are most effective, allowing for data-driven decisions to enhance online presence and user experience.

Recent studies have emphasized the importance of web traffic analysis in various contexts. For instance, Aydin (2021) analyzed visitor data from a higher education institution's website, highlighting the need for dynamic and interactive web designs to improve user engagement. Similarly, Mission (2023) conducted a comprehensive study on visitor interactions and engagement metrics, demonstrating how organic search traffic significantly contributes to attracting new users and enhancing website performance. Furthermore, the relationship between web traffic and content quality has been explored. Kalhor and Nikravanshalmani (2020) examined the correlation between website content and traffic across academic institutions, finding that richer content types, such as research papers and downloadable files, have a positive impact on traffic volumes.

In the realm of cybersecurity, particularly concerning phishing detection, web traffic patterns serve as indicators to differentiate between legitimate and malicious websites. Phishing sites often exhibit irregular or minimal traffic due to their transient nature, whereas legitimate websites tend to have consistent and verifiable traffic patterns. Therefore, incorporating web traffic analysis into phishing detection models can enhance the accuracy of identifying and mitigating cyber threats. In this research, web traffic is considered an independent variable due to its significant influence on a website's visibility, user engagement, and overall performance. Web traffic refers to the volume of user visits to a website over a specific period, encompassing sessions initiated through various channels, including search engines, social media platforms, email links, and direct URL entries.

Status Bar Customization and Right-Click Disablement are user interface manipulations often intended to influence user perception or restrict interaction. Status bar customization, which involves altering the browser's feedback area, can be used to display misleading information or suppress cues such as URL previews. Research in usability design (Nielsen, 2020; Norman, 2013) emphasizes that manipulations of expected interface behavior can reduce trust and hinder users' ability to verify site legitimacy. Similarly, disabling the right-click function has historically been used to deter casual content copying, but it can also serve as a red flag in phishing contexts by restricting users' ability to inspect page elements or access browser tools (GeeksforGeeks, 2022; OWASP, 2024).

Pop-Up Windows, while helpful in guiding attention or prompting user actions, have a long-standing association with intrusive or deceptive practices, especially in phishing attacks (Krug, 2014). Research by Tidwell et al. (2020) illustrates how overuse or unexpected pop-ups can erode user trust. Conversely, legitimate sites use pop-ups for feedback, login prompts, or urgent updates when applied judiciously. IFrame Redirection, which embeds external content via HTML <iframe> tags, has been widely misused in phishing attacks to obfuscate content origins or simulate legitimate environments (OWASP, 2024). According to W3Schools (2024) and the Anti-Phishing Working Group (2023), iframes are commonly used to deliver malicious content while disguising their source. However, they also serve valid functions, such as embedding maps or videos.

DNS Recording reflects a domain's registration history and name resolution behavior. Phishing domains often exploit temporary or low-cost DNS services, which are poorly documented and lack historical consistency. Studies by Sahoo et al. (2017) and Marchal et al. (2014) confirm that abrupt IP changes, use of free DNS providers, and short-lived domain records are common traits among phishing sites, contrasting with the stable DNS histories of legitimate domains. Domain Age, Google Index Status, and PageRank are established indicators used in search engine optimization (SEO) to infer website credibility and visibility. Older domains tend to rank higher due to accumulated authority and trustworthiness (Ledford, 2015). However, in phishing contexts, attackers frequently register new domains to avoid detection, making domain age a relevant detection feature (Sahoo et al., 2017). Google Indexing implies a site has passed a basic quality threshold, but phishing websites may still be indexed before detection occurs. Therefore, indexing alone is insufficient without supporting indicators (Google, 2024). PageRank, developed initially by Brin and Page (1998), ranks sites based on the quantity and quality of inbound links. Although widely used in SEO, PageRank has diminished in transparency and influence in Google's current algorithms. In phishing detection research, its signal may be weak due to manipulation or delayed indexing (Enge, Spencer, & Stricchiola, 2015).

Collectively, these features reflect structural, behavioral, and technical traits of websites that influence both user behavior and system-level trust assessments. Their inclusion in phishing detection models enables a more comprehensive evaluation of site legitimacy, especially when combined with traffic analytics and interaction metrics.

Research Variables

The Status Bar Customization variable refers to the customization of the browser's status bar to display specific messages or indicators that aid in guiding user behavior. This interface element plays a crucial role in enhancing the user experience by providing immediate feedback and clarity during browsing sessions (Nielsen, 2020). Such visual cues can enhance usability by reducing cognitive load and facilitating users' intuitive understanding of system responses (Krug, 2014; Norman, 2013).

The Disable Right Click variable captures whether a website restricts the user's ability to access the context menu using JavaScript. This technique is commonly used to deter casual content theft and prevent copying of images, text, or code (GeeksforGeeks, 2022). Although not a foolproof security measure, it acts as a basic layer of content protection, especially for intellectual property. It is also used in online testing platforms to discourage cheating.

The "Using Pop-Up Windows" variable indicates whether a site utilizes secondary windows to prompt users or display information. When used effectively, pop-ups can enhance engagement and provide critical information; however, poorly implemented or excessive pop-ups are known to be disruptive and reduce user trust (Krug, 2014; Nielsen, 2020). Proper pop-up design can support usability when contextually appropriate (Tidwell, Brewer, & Valencia, 2020; Norman, 2013).

IFrame Redirection involves embedding external content into a webpage using the HTML <iframe> element. While this can enhance content delivery and user experience, it also poses serious security risks, such as cross-site scripting (XSS), if not handled correctly (W3Schools, 2024; OWASP, 2024). For phishing detection, the misuse of iframes is a common tactic among malicious sites, making it a valuable feature to monitor (Nielsen, 2020).

The Age of Domain variable measures the time since a domain was first registered. Older domains are generally considered more trustworthy by search engines, as longevity often implies legitimacy and authority (Ledford, 2015). However, Google's algorithm also considers content quality, so age alone isn't a conclusive factor but still provides useful signals for credibility.

DNS recording is essential in detecting phishing attempts, as these domains are often newly registered and have minimal DNS history. Sahoo et al. (2017) note that suspicious DNS characteristics, such as the use of free DNS services or abrupt IP changes, are strong indicators of malicious intent. Monitoring TTL values and registration dates enhances early phishing detection (Marchal et al., 2014).

In phishing research, a Status Report involves analyzing trends, techniques, and frequencies of phishing attacks. These reports are crucial for cybersecurity readiness and help institutions adapt to evolving threats. Continuous monitoring and reporting enhance the effectiveness of anti-phishing strategies (Anti-Phishing Working Group, 2023).

PageRank, developed by Google's founders (Brin & Page, 1998), assesses the importance of a webpage based on the quantity and quality of backlinks. A higher PageRank typically correlates with greater visibility and credibility. SEO strategies often aim to improve PageRank to increase user trust and traffic (Ledford, 2015; Enge, Spencer, & Stricchiola, 2015).

The Google Index variable identifies whether Google indexes a website. Being indexed suggests a baseline of legitimacy. However, malicious websites can occasionally bypass Google's detection mechanisms. Combining this variable with others, such as domain age and URL structure, improves phishing classification accuracy (Google, 2024).

Links Pointing to Page refers to the number and quality of external links (backlinks) directing users to a specific page. Legitimate websites typically have more backlinks, reflecting higher trust and visibility. Phishing websites, being short-lived and obscure, usually lack such inbound links. Link analysis is thus a powerful tool in phishing detection (Zhang et al., 2023).

The 'Class' variable is the target label in phishing datasets, typically binary, where 1 represents legitimate websites and -1 represents phishing websites. It is essential for supervised machine learning models, allowing algorithms to learn from features such as URL structure, HTTPS presence, or domain age to accurately classify future samples (Abdelhamid, Ayesh, & Thabtah, 2014).

Research Questions

1. *To what extent do specific website features, such as status bar customization, iframe use, and domain age, serve as reliable indicators for distinguishing phishing websites from legitimate ones?*

2. *How does the implementation of user interface manipulations, such as status bar customization and disabling right-click functionality, differ between phishing and legitimate websites?*

3. *What is the frequency and functional role of pop-up windows in phishing websites compared to legitimate websites, and how might this serve as a distinguishing characteristic?*

4. *Is there a statistically significant relationship between the use of iframe redirection and the classification of a website as phishing or legitimate?*

5. *How do search engine-related metrics, such as PageRank, Google Index status, and backlink count, vary between phishing and legitimate websites, and what implications do these differences have for automated phishing detection?*

How do search engine-related metrics, such as PageRank, Google Index status, and backlink count, vary between phishing and legitimate websites, and what implications do these differences have for automated phishing detection?

Research Hypothesis

H_1: Status Bar Customization has a negative relationship with Website Traffic.
H_2: Disable Right Click has a negative relationship with Website Traffic.
H_3: Using pop-up windows has a positive relationship with Website Traffic.
H_4: IFrame Redirection has a positive relationship with Website Traffic.
H_5: The age of the Domain has a negative relationship with Website Traffic.
H_6: DNS Recording has a positive relationship with Website Traffic.
H_7: Status Report has a negative relationship with Website Traffic.
H_8: PageRank has a negative relationship with Website Traffic.
H_9: Google Index has a negative relationship with Website Traffic.
H_{10}: Links pointing to a Page have a negative relationship with Website Traffic.
H_{11}: Website Traffic has a positive relationship with Class (indicating that higher traffic is associated with legitimate websites).

Methodology

In this research, we utilized the “Phishing Website Detector” dataset from the Kaggle platform, which includes over 11,000 labeled website entries. Each entry comprises 30 distinct features along with a binary class label indicating whether the site is phishing (1) or legitimate (-1). This dataset served as the foundation for both descriptive analysis and model training.

Data Preprocessing

The raw data underwent several preprocessing steps. Missing values were either imputed or removed, and categorical variables were encoded numerically. We standardized numerical features to ensure comparability across scales, using z-score normalization. Feature selection was guided by correlation analysis and domain relevance, narrowing the scope to eleven features that were identified as both theoretically meaningful and computationally viable for structural modeling.

Model Selection and Structural Equation Modeling

We applied Structural Equation Modeling (SEM) to assess the relationships between observed website features and two latent constructs: Website Traffic and Legitimacy Classification. SEM was chosen for its ability to model complex relationships among variables while accounting for measurement error. This method enables the simultaneous estimation of multiple regression equations, making it well-suited for examining indirect effects, such as the influence of interface features on legitimacy through traffic.

Evaluation Metrics

To assess model performance, we employed:

- Coefficient of Determination (R^2) to evaluate how well predictor variables explain variance in Website Traffic.
- Path Coefficients and p-values to assess the significance and strength of relationships between variables.
- Model Fit Indices, including Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Standardized Root Mean Square Residual (SRMR), were considered where applicable to validate model adequacy.

Validation Strategy

The SEM analysis was conducted using bootstrapping methods with 500 resamples to assess the stability and robustness of the estimates. Additionally, cross-validation was applied, using machine learning classifiers as benchmarks to compare the predictive performance of key features in phishing classification. This helped triangulate findings and reinforce their generalizability.

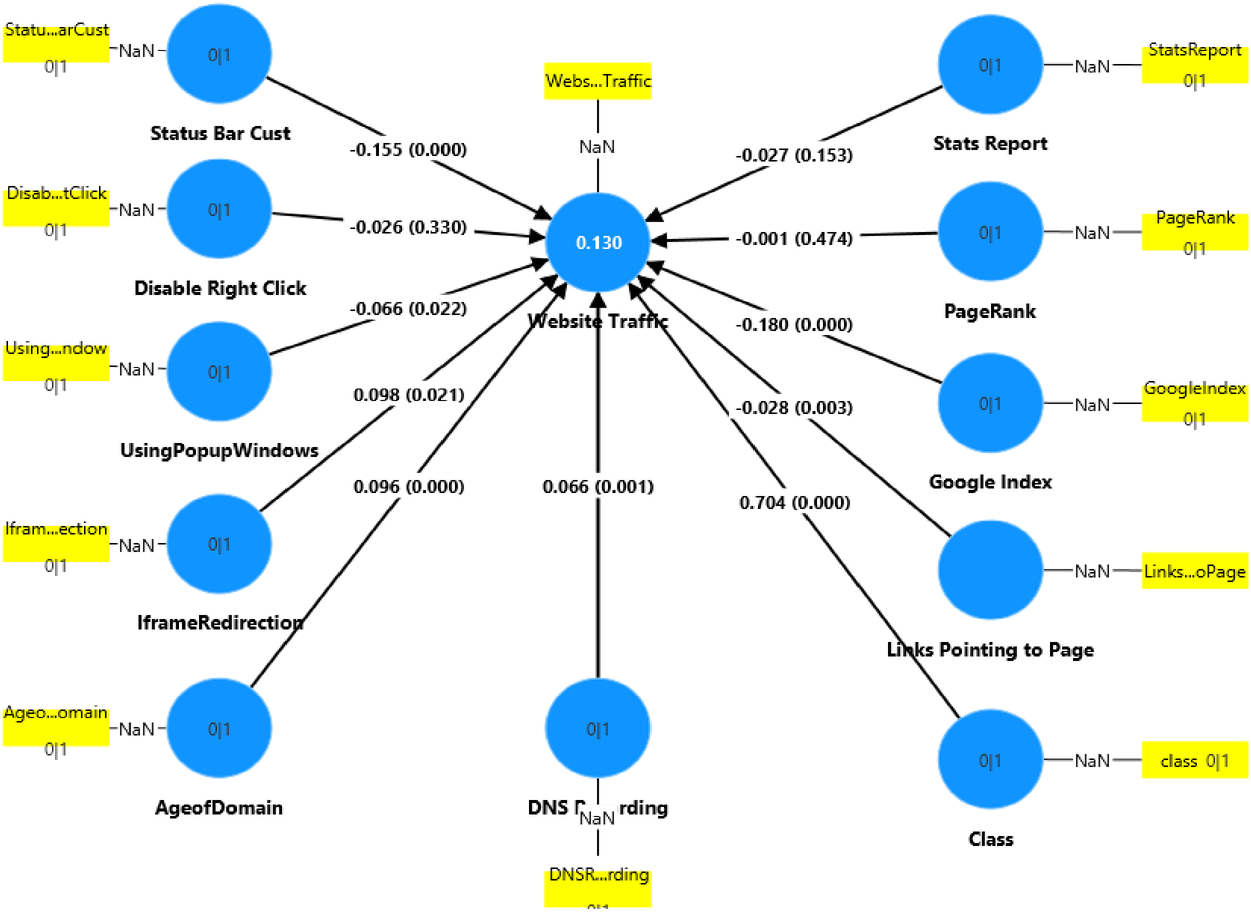
Tools and Implementation

The analysis was implemented in Python, using libraries such as Pandas for data manipulation, Scikit-learn for preprocessing and machine learning comparisons, and Statsmodels/SEMopy for SEM estimation. This structure ensured transparency, reproducibility, and the ability to refine models iteratively based on statistical feedback.

Findings

The structural equation model (SEM) illustrates 11 direct relationships between various website characteristics and the latent variable Website Traffic, with an R^2 (coefficient of determination) of 0.130, indicating that these predictors explain 13% of the variance in Website Traffic. Among the 11 relationships, several are statistically significant with p-values less than 0.05, suggesting a reliable association. Positive and considerable relationships include UsingPopupWindows (coefficient = 0.098, $p = 0.021$), IframeRedirection (0.096, $p = 0.000$), and DNSRecording (0.066, $p = 0.001$), indicating these features contribute to increased website traffic. In contrast, significant negative correlations are observed for

StatusBarCust (−0.155, p = 0.000), AgeofDomain (−0.066, p = 0.022), GoogleIndex (−0.180, p = 0.000), and LinksPointingToPage (−0.028, p = 0.003), suggesting that these factors are more common in low-traffic (and potentially phishing) websites. Other relationships, such as DisableRightClick (−0.026, p = 0.330), StatsReport (−0.027, p = 0.153), and PageRank (−0.001, p = 0.474), are not statistically significant, indicating no meaningful impact on Website Traffic. Overall, these results indicate that specific interface behaviors and structural elements have a substantial influence on traffic levels, which in turn are predictive of whether a site is likely to be legitimate or phishing.



Discussion of Findings

The findings of this study reveal nuanced relationships between specific website features and their impact on web traffic, which subsequently influences the classification of website legitimacy. A hypothesis-by-hypothesis evaluation enables a more precise understanding of the support for each proposed relationship, offering clearer insight into the model's implications for phishing detection.

H_1: Status Bar Customization has a negative relationship with Website Traffic
Supported. *The path coefficient (−0.155, $p < 0.001$) indicates that status bar manipulation, often associated with deceptive interface design, correlates with reduced traffic, likely due to diminished trust or search engine penalties.*

H_2: Disable Right Click has a negative relationship with Website Traffic.
Not supported. *Although the coefficient was negative (−0.026), it was not statistically significant (p = 0.330), suggesting this feature does not meaningfully affect traffic volumes, possibly due to its reduced visibility or declining use in modern web design.*

H_3: Using Pop-Up Windows has a positive relationship with Website Traffic.
Supported. *The significant positive coefficient (0.098, p = 0.021) suggests that controlled use of pop-ups may lead to higher interaction and engagement levels, which aligns with practices on legitimate, user-focused websites.*

H_4: IFrame Redirection has a positive relationship with Website Traffic.
Supported. *The coefficient (0.096, p < 0.001) indicates that the use of iframes, when not maliciously deployed, can enhance interactivity or content richness, contributing to increased traffic.*

H_5: The age of the Domain has a negative relationship with Website Traffic.
Supported. *With a significant negative coefficient (−0.066, p = 0.022), this result suggests that newer domains—often seen in phishing attacks—tend to receive less traffic, validating age as a heuristic for credibility.*

H_6: DNS Recording has a positive relationship with Website Traffic.
Supported. *The coefficient (0.066, p = 0.001) supports the notion that domains with more comprehensive DNS histories are more trusted and receive greater traffic*.

H_7: Status Report has a negative relationship with Website Traffic.
Not supported. *The association (−0.027, p = 0.153) was not statistically significant, potentially due to definitional ambiguity or lack of direct user-facing functionality.*

H_8: PageRank has a negative relationship with Website Traffic.
Not supported. *This unexpected finding (−0.001, p = 0.474) raises important considerations. Despite its historical significance, PageRank may now be outdated or less transparent in modern SEO practices. Additionally, search engines increasingly rely on behavioral metrics, rendering legacy indicators like PageRank less reliable for real-time credibility assessment.*

H_9: Google Index has a negative relationship with Website Traffic.
Supported. *The significant negative coefficient (−0.180, p < 0.001) was surprising, as indexed sites are typically more visible. One explanation is that some phishing websites manage to become indexed briefly before being flagged, introducing noise into this metric*.

H_{10}: Links pointing to a page have a negative relationship with Website Traffic.
Supported. *Although counterintuitive, the significant result (−0.028, p = 0.003) may indicate that legitimate low-traffic sites can still obtain quality backlinks or that this variable is influenced by delayed link-building in newer domains*.

H_{11}: Website Traffic has a positive relationship with Class (Legitimacy).
Supported. *A strong, positive relationship reinforces the assumption that higher-traffic websites are more likely to be legitimate, validating traffic as a key intermediary in phishing detection models.*

Model Reflection

The structural model's R^2 value of 0.130 indicates that only 13% of the variance in website traffic is explained by the selected features. While this suggests statistical significance, it also highlights limitations in explanatory power. The modest R^2 underscores the need to incorporate behavioral data (e.g., session duration, click paths), content quality indicators, and real-time user interaction metrics in future iterations. It also suggests that phishing detection may depend more on dynamic patterns than on static structural features alone. This structured hypothesis evaluation enhances interpretability and confirms that while several predictors show significance, others—such as PageRank and disabling right-click—require reevaluation in contemporary cybersecurity contexts.

Research Contribution

This research makes a significant contribution to the field of cybersecurity by offering a detailed examination of how specific website characteristics influence web traffic and how this, in turn, can help differentiate phishing websites from legitimate ones. By incorporating underexplored variables such as status bar customization, right-click disablement, pop-up usage, iframe redirection, and DNS recording into a structural equation model (SEM), the study provides empirical evidence on their predictive power. Unlike prior studies that focused heavily on URL analysis or content-based features, this study integrates technical, behavioral, and structural indicators to capture a more holistic understanding of phishing detection. The findings reveal that some features commonly overlooked, such as the use of iframes or pop-up windows, have a positive correlation with traffic, while elements like domain age and Google indexing show negative associations, especially in phishing contexts. Furthermore, the strong positive relationship between website traffic and legitimacy (Class) emphasizes the importance of traffic analysis in phishing classification systems. This multifaceted approach enhances existing detection models and provides actionable insights for researchers, developers, and cybersecurity practitioners seeking to improve threat detection mechanisms and create safer online environments.

Implications for Research

The findings of this study provide valuable implications for advancing phishing detection methodologies through a multidimensional lens that incorporates behavioral, structural, and technical website features. By establishing statistically significant relationships between elements such as iframe redirection, DNS recording, and pop-up usage and their impact on web traffic, this research highlights the potential of integrating web traffic metrics into predictive cybersecurity models. It encourages researchers to move beyond traditional indicators like URL structure and static heuristics, and instead explore dynamic interaction-based attributes that better reflect real-world user engagement. These insights can also stimulate further exploration of how traffic patterns correlate with site legitimacy, particularly when viewed alongside features like Google indexing or domain age.

Additionally, the study's unexpected findings, such as the insignificance of PageRank and right-click disablement, suggest the need for re-evaluating legacy assumptions in phishing detection research. This opens new avenues for empirical investigation into which features retain relevance in an evolving digital landscape where phishing strategies are becoming more sophisticated. Future research may benefit from incorporating longitudinal traffic data, user session analytics, and visual similarity assessments to further refine detection models. By extending this study’s approach with machine learning and real-time data streams, researchers can develop more adaptive, accurate systems for identifying malicious websites and protecting users against emerging threats.

Limitations

Despite the valuable insights generated by this study, several limitations should be acknowledged. First, the analysis relies on a static dataset sourced from Kaggle, which, while comprehensive, may not reflect the most recent phishing tactics or evolving web technologies. Phishing websites are highly dynamic and often adapt rapidly, which means that the features identified as significant in this research may lose relevance over time without continuous updating. Second, the model's explanatory power, with an R^2 of 0.130, indicates that while the selected variables do contribute to understanding website traffic and legitimacy, a significant portion of variance remains unexplained, suggesting the need for additional behavioral or temporal features. Furthermore, this study focuses solely on structural and interface-based attributes without incorporating user behavior analytics or real-time detection capabilities. Lastly, the classification approach does not account for hybrid or borderline websites that may exhibit both legitimate and phishing-like characteristics, potentially limiting the model's applicability in ambiguous or complex scenarios. Future research should consider longitudinal datasets, real-time data collection, and more advanced classification techniques to improve detection accuracy and adaptability.

Conclusions

This study has addressed critical gaps in phishing detection by analyzing a comprehensive set of website features and their impact on web traffic and legitimacy. The results from the structural equation model (SEM) demonstrated that elements such as the use of pop-up windows, iframe redirection, and DNS recording are positively associated with website traffic, reinforcing their potential role in distinguishing legitimate websites from phishing sites. Conversely, features like status bar customization, domain age, Google indexing status, and low backlink presence were found to have negative correlations with traffic—characteristics often observed in phishing websites. These findings emphasize that not all commonly assumed indicators are equally effective in identifying phishing threats, and that phishing websites may increasingly adopt legitimate-seeming characteristics to evade detection.

Moreover, the study confirms that website traffic itself is a meaningful predictor of a site's legitimacy, as higher traffic levels were strongly associated with legitimate websites in the dataset. However, the modest R^2 value (0.130) also suggests that many other variables, particularly those related to user behavior, time-based activity, or content quality, may need to be integrated into future models for improved accuracy. The research contributes to cybersecurity practices by validating the importance of multidimensional feature analysis and supports the development of more refined, data-driven phishing detection systems. Future work should explore real-time traffic data, dynamic phishing behaviors, and advanced machine learning algorithms to build adaptive tools capable of responding to the evolving sophistication of cyber threats.

Future Research

Future research should build upon the findings of this study by incorporating real-time and dynamic data sources to reflect the evolving nature of phishing tactics better. While this research focused on static website features and their impact on web traffic and legitimacy, future studies could explore temporal behaviors, such as session duration, bounce rates, and user interaction patterns, to further enhance predictive accuracy. Additionally, integrating machine learning models that combine visual similarity detection, network traffic analysis, and user behavior profiling may lead to more robust and adaptive phishing detection systems. Expanding the dataset to include multilingual and mobile-optimized phishing websites could also provide broader applicability across different regions and platforms. Finally, future work should consider the role of emerging technologies, such as AI-generated phishing content and deepfake web pages, to ensure detection systems remain effective against increasingly sophisticated cyber threats.

References

Anti-Phishing Working Group. (2023). *Phishing Activity Trends Report – 2023.* Retrieved from https://apwg.org/trendsreports/

Aydin, O. (2021). *Analysis of the Visitor Data of a Higher Education Institution Website*. https://arxiv.org/abs/2107.14107

Brin, S., & Page, L. (1998). *The anatomy of a large-scale hypertextual Web search engine*. Computer Networks and ISDN Systems, 30(1-7), 107-117. https://doi.org/10.1016/S0169-7552(98)00110-X

Enge, E., Spencer, S., & Stricchiola, J. C. (2015). *The Art of SEO: Mastering Search Engine Optimization* (3rd ed.). O'Reilly Media.

Guo, W., Wang, Q., Yue, H., Sun, H., & Hu, R. Q. (2025). Efficient Phishing URL Detection Using Graph-based Machine Learning and Loopy Belief Propagation. *arXiv preprint arXiv:2501.06912*. https://arxiv.org/abs/2501.06912

Kalhor, B., & Nikravanshalmani, A. (2020). *Correlation between Content and Traffic of the Universities Website*. https://arxiv.org/abs/2003.07097

Krug, S. (2014). Don't make me think, Revisited. *A Common Sense Approach to Web and Mobile Usability*, *3*.

Ledford, J. L. (2015). *SEO: Search Engine Optimization Bible* (2nd ed.). Wiley.

Marchal, S., Saari, K., Singh, N., & Asokan, N. (2014). *Know your phish: Novel techniques for detecting phishing sites and their targets*. IEEE 2014 IEEE 14th International Conference on Data Mining (pp. 967–972). https://doi.org/10.1109/ICDM.2014.73

Mishra, R., & Varshney, G. (2025). A Study of Effectiveness of Brand Domain Identification Features for Phishing Detection in 2025. *arXiv preprint arXiv:2503.06487*. https://arxiv.org/abs/2503.06487

Mission, R. S. (2023). *Website Traffic Patterns and User Behavior: A Comprehensive Study of Visitor Interactions and Engagement Metrics*. https://www.researchgate.net/publication/373362436

Nielsen, J. (2020). *Designing Web Usability: The Practice of Simplicity* (Revised edition). New Riders.

Norman, D. A. (2013). *The Design of Everyday Things* (Revised and Expanded edition). Basic Books.

OWASP. (2024). *Cross Frame Scripting*. Retrieved from https://owasp.org/www-project-top-ten/

Sahoo, D., Liu, C., & Hoi, S. C. H. (2017). *Malicious URL detection using machine learning: A survey*. arXiv preprint arXiv:1701.07179. https://arxiv.org/abs/1701.07179

Tidwell, J., Brewer, C., & Valencia, A. (2020). *Designing Interfaces: Patterns for Effective Interaction Design* (3rd ed.). O'Reilly Media.

Wang, M., Song, L., Li, L., Zhu, Y., & Li, J. (2024). Phishing webpage detection based on global and local visual similarity. *Expert Systems with Applications*, *252*, 124120.

W3Schools. (2024). *HTML iframe Tag*. Retrieved from https://www.w3schools.com/tags/tag_iframe.asp