Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

E-Commerce Phishing Websites in Vietnam: An Exploratory Study

Joshua Dwight

joshua.dwight@rmit.edu.vn

RMIT University https://orcid.org/0000-0002-6247-2703

Research Article

Keywords: phishing websites, e-commerce, Vietnam, exploratory data analysis

Posted Date: July 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10504449/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

E-commerce Phishing Websites in Vietnam: An Exploratory Study

Joshua Dwight

Royal Melbourne Institute of Technology (RMIT) – Vietnam, School of Science, Engineering, and Technology, Hanoi, Vietnam

Joshua.dwight@rmit.edu.vn

ORCID#: 0000-0002-6247-2703

Abstract

Vietnam has experienced rapid digital transformation, with e-commerce growing at a rate of 25% in 2023, yet this expansion is increasingly threatened by social engineering attacks. This exploratory study investigates the nature and characteristics of e-commerce phishing websites in Vietnam, specifically comparing data from 2022, 2023, and 2024. Utilizing a quantitative research design, the study analyzed 607 e-commerce phishing websites collected from the National Cyber Security Center (NCSC) of Vietnam's weekly reports. The methodology employed exploratory data analysis (EDA), descriptive statistics, and analysis of variance (ANOVA) with Post Hoc Tukey tests to evaluate key features including URL character length, protocols, Top Level Domains (TLDs), Levenshtein distance (similarity score), and VirusTotal community scores.

The results of this exploratory study indicate that major platforms such as Shopee, Tiki, and Lazada are the most frequently impersonated brands. Statistical analysis revealed significant differences across the three years regarding average URL character lengths, similarity scores, and VirusTotal scores, suggesting that malicious actors continuously adapt their tactics. While the use of HTTPS and ".com" domains remains prevalent to establish fake credibility, fluctuations in URL length and protocols suggest ongoing experimentation by phishers to evade detection. These findings contribute to a better understanding of the evolving phishing landscape in Vietnam and highlight the need for enhanced security awareness and technical countermeasures in the region's digital economy.

Keywords: phishing websites, e-commerce, Vietnam, exploratory data analysis

Introduction

Vietnam has shown incredible socio-economic development over the past 35 years and has become one of the most open economies in the world (Nguyen, 2022). More recently, Vietnam and other Southeast Asian nations have been adopting Industry 4.0 to transform the countries into digital economies (Van et al., 2021). Per the 2024 Vietnam E-Commerce Business Index Report, e-commerce had a growth rate of 25% in 2023, with online retail sales accounting for $17.3 billion (VECOM, 2024). The Vietnam Ministry of Industry and Trade also projects that e-commerce export value will reach $5.8 billion by 2028 (Vietnam ECommerce and Digital Economy Agency, 2024). E-commerce is one of the key growth areas of a digital economy; however, there are issues such as phishing, fraud, and other unethical practices that impact the success associated with e-commerce (Van et al., 2021). Phishing attacks are a critical security threat that leads to financial, reputational, technical, and consumer damage (Sabir et al., 2022). The Anti-Phishing Work Group (APWG) 2024 reports note there were over 3.7 million reported phishing attacks globally with 7% of attacks targeting e-commerce. See Appendix A. Since Vietnam is experiencing and focusing on digital growth, the digital threats for the country are high. Yet, there is limited research on phishing websites targeting Vietnamese e-commerce firms and respective consumers.

This study aims to explore the nature of e-commerce phishing websites in Vietnam by analyzing empirical data from the Vietnam National Cyber Security Center (NSCS). By investigating key technical phishing indicators such as URL structure, protocol usage, and brand impersonation trends, the research seeks to answer:

RQ1: What are the characteristics of e-commerce phishing websites targeting Vietnam?

RQ2: How do e-commerce phishing website tactics compare between 2022, 2023, and 2024 in Vietnam?

This study will provide a background on phishing and phishing websites. This includes breaking down the components of a website, identifying and analyzing phishing websites, and describing various security measures. The research design is a quantitative study using exploratory data analysis (EDA). This study utilizes EDA to identify trends or patterns with descriptive statistics and data visualizations to generate hypotheses. Then statistically validates by testing hypotheses using one-way ANOVA and Kruskal-Willis tests.

E-commerce in Vietnam

E-commerce is the exchange of goods and services using the Internet and related technologies such as payment systems. E-commerce provides convenient and effective online shopping experiences with flexibility in access, payment, and delivery of products and services. These technologies can include websites, mobile apps, and social media. There are several different e-commerce business models, including business-to-business (B2B), business-to-consumer (B2C), consumer-to-business (C2B), consumer-to-consumer (C2C), and business-to-government (B2G) (Dwight et al., 2025). E-commerce generally falls under the business-to-consumer model.

E-commerce is one of the key growth areas of the digital economy (Van et al., 2021), and Vietnamese e-commerce had a growth rate of 25% and $17.3 billion in sales in 2023 (VECOM, 2024). Vietnamese e-commerce is expected to grow, and the top e-commerce platforms in Vietnam are Shopee, Lazada, Tiki, TikTok Shop, and Sendo. These organizations accounted for $9.5 billion in revenue in 2023 (Hoa, 2024). However, Vietnam's e-commerce growth has also been observed by malicious actors.

Phishing and Phishing Websites

Phishing is a well-known social engineering problem (Charmet et al., 2024). However, it is very difficult to deal with. Phishing is where malicious actors gather personal and financial data and information about people and organizations using deceptive techniques such as impersonation (Alkhalil et al., 2021; Desolda et al., 2021). Additionally, advances in generative artificial intelligence have drastically increased the velocity of creation of sophisticated attacks. Phishing continues to be one of the most effective cyberattacks (Desolda et al., 2021).

Phishing techniques generally encompass a communication method, deceptive content, and a target. Phishing attacks adopt many communication methods such as social media, emails, websites, messaging, forums, SMS, eFax, and voice to deceive people into disclosing personal or important information (Alkhalil et al., 2021; Ayeni et al., 2024; Desolda et al., 2021; Hamid & Abawajy, 2014). Malicious actors continue to creatively use and adapt different modes of communication and techniques to carry out scams and fraud (Dwight et al., 2025). A phishing website is where malicious actors construct counterfeit websites and connect them to a communication method to trick users into providing personal information or downloading malicious software while using their device (Ayeni et al., 2024).

Phishing tactics continue to change and evolve in sophistication to defeat anti-phishing tactics and security measures (Hamid & Abawajy, 2014). Phishing takes advantage of human vulnerabilities such as the lack of knowledge, lack of resources, lack of awareness, lack of norms, and complacency (Desolda et al., 2021) with the use of intelligent technical techniques to give the impression users are dealing with legitimate websites (Mohammed et al., 2015). Subramani et al. (2022) noted that phishing websites do not need to be an exact facsimile of a branded website, but just enough details and web pages to be sufficiently believable. These technical techniques can include domain

spoofing (Dawkins & Jacobs, 2023), domain obfuscation (Skula & Kvet, 2024), and domain squatting (Tian et al., 2018).

Domain spoofing is where a domain name appears familiar and legitimate (Dawkins & Jacobs, 2023), and often the fake domain name will include misspellings or additional characters. Domain obfuscations are tactics used to conceal or make it difficult to recognize the actual destination of the domain or uniform resource locator (URL). Malicious actors will do so by manipulating one or multiple components of a URL (Skula & Kvet, 2024). Domain squatting is the act of registering domain names that are similar to existing brands and trademarks (Tian et al., 2018). The next section provides a breakdown of the various components of a website URL that malicious actors can target for spoofing and obfuscation.

Anatomy of a URL

A uniform resource locator (URL) is a reference to a web resource that specifies its location and the protocol for retrieving the resource (Dodson et al., 2021). A URL is composed of a protocol, subdomain, domain name, and path. A domain name is a combination of a top-level domain and 2nd level domain. Hypertext transfer protocol (HTTP) is a standard communication method between clients and web servers. Hypertext transfer protocol secure (HTTPS) is the secure version of the standard communication method (Akram et al., 2020). The World Wide Web (WWW) is the global system that allows content sharing and access across the Internet. Each resource on the Internet is identified by a unique domain name.

Top-level domains (TLDs) are known as root domains (Rose et al., 2025). Most commonly, ".com" is widely used, although there are over 1400 TLDs globally (IANA, 2025). There are generally three types of TLDs. First are country code domains, which represent specific countries and territories. Sponsored generic domains represent specific interests such as ".edu", ".gov", and ".mil". Last are unsponsored generic domains, which refer to ".com", ".net", and ".org" domains (Rose et al., 2025). TLDs can have 2nd and 3rd level domains, which are known as children or enterprise-level domains (Rose et al., 2025). Enterprise-level domains typically represent an organizational brand for ease of access and recognition, such as ".shopee.com".

The URL path specifies the location of a resource on a server. These paths are used to navigate specific pages, files, or directories on a website. Manipulating these components and acquiring specific domain names can help malicious actors evade detection and increase credibility for brand impersonation.

Phishing Website Duration

The phishing website duration includes planning, collection, and fraud (Mohammed et al., 2015). The Moore and Clayton (2007) study indicated that the average lifetime of a phishing site was 2.57 days based on 18, 680 URLs from PhishTank. A later study by Mohammed et al. (2015) estimated the phishing duration is between 3 and 4 days. A study by Oest et al. (2020) examined and monitored 4.8 million victims over one year and suggested that the average campaign lasted 0.88 days. In contrast, Drury et al. (2022) specified the estimated average duration of a phishing campaign was 11.94 days based on certificates, crt.sh, WHOIS, and carbon14 information. The study reviewed 133,667 URLs from PhishTank, OpenPhish, and PhishStats. The study concluded that there was no definitive way to get complete information about all websites (Drury et al, 2022). More recently, Lee et al. (2025) analyzed 286,237 unique phishing URLs from the Anti-Phishing Work Group (APWG) and showed that the average lifespan of a phishing website of well-known brands is 2.25 days, with most sites taken down within 5.46 hours. The study also noted that unique brands could have phishing sites that lasted more than 30 days. A study by Agarwal and Vasek (2025) reviewed 14,112 domains and concluded an average lifetime of 8.6 days. Table 1 summarizes the phishing duration stages.

Table 1: Phishing duration stages

Authors	Planning	Collection	Fraud

Moore & Clayton (2007)		2.57 days	
Mohammed et al. (2015)	0.04 days	0.50 days	3.00 days
Oest et al. (2020)		0.88 days	
Drury et al. (2022)		11.94 days	
Lee et al. (2025)		2.25 days	
Agarwal & Vasek (2025)		8.60 days	

Phishing Website Security Measures

Many studies have outlined different security measures for phishing and phishing websites. There tend to be four categories of security measures, including technical content measures, technical non-content measures, security awareness, and threat information sharing (Alkhalil et al., 2021; Johnson et al., 2016). Phishing identification, analysis, and response can use non-content-based and content-based machine learning and deep learning approaches (Afroz & Greenstadt, 2011; Alkhalil et al., 2021). Dawkins & Jacobs (2023) created the NIST Phish Scale, which analyzes and categorizes phishing cues. This is generally directed towards phishing emails, but some of the technical cues can be applied specifically to phishing websites. These include cues such as spelling irregularities and domain spoofing.

Technical and non-technical content measures

Content-based approaches focus on examining site content such as grammar, links, input fields, images, colors, hypertext markup language (HTML), JavaScript, cascading style sheets (CSS), metadata, pages, page transitions, CAPTCHA, and other website-related content (Afroz & Greenstadt, 2011; Alkhalil et al., 2021; Subramani et al., 2022). Charmet et al. (2024) proposed machine learning and deep learning that focused on visual detection. Additionally, Subramani et al. (2022) utilized machine learning-based web crawlers and input validation to identify phishing websites. However, Lee et al. (2025) noted malicious actors can adapt a phishing website's visual appearance to avoid detection and prolong the phishing campaign.

Non-content measures

Non-content approaches concentrate on URL blacklisting and whitelisting (Afroz & Greenstadt, 2011; Alkhalil et al., 2021). Current detection and prevention methods can be mainly classified into black/whitelists, rule-based approaches, machine learning, and deep learning (Alkhalil et al., 2021; He et al., 2024). Linh and Hung (2025) created a feature-based dataset for machine learning evaluation due to a limitation of phishing datasets. The authors noted that ML and LLMs have strong potential for URL-based classification but suffer from limitations in reproducibility and fair comparison across models. Linh and Hung specify using features including URL length, domain length, digit ratio, entropy, HTTPS usage, URL, domain, and TLD for classification (2025).

Choo et al. (2023) conducted a study on malicious URLs that used hosting, lexical, and WHOIS features. WHOIS is a widely used internet protocol to query information about registered users of domain names, IP addresses, and autonomous systems. The protocol can provide information related to contact details, registration dates, and nameservers (WHOIS, 2025). Online scanners can help identify malicious websites. One of the most popular online scanners is VirusTotal (van Liebergen et al., 2023). VirusTotal is a tool that aggregates results from 95 phishing detection scanners and can provide information to the cybersecurity community (Choo et al., 2023). Users can submit malicious online and cybersecurity professionals can utilize the VirusTotal API for machine learning purposes (van Liebergen et al., 2023).

Security awareness measures

Organizations and staff need to understand potential cyber threats to successfully manage and protect the organization (Tran et al., 2025). To this end, organizations need to implement an effective governance framework

and training to ensure staff understand their duties and the procedures to protect sensitive and confidential information (Tran et al., 2025). Security awareness campaigns and phishing simulation campaigns are security measures that try to influence the behavior of employees and users involved with an organization (Braun et al., 2025). Braun et al. conducted a study on 36 organizations of various sizes to test phishing security awareness and simulation campaigns (2025). Carella et al. (2014) conducted phishing simulation campaigns and found that people with no security awareness training had the highest click-through rates, and people with document or class security awareness training performed better. However, they also noted that the longer the time passed since the security training, the more people fell for the phishing attempts (Carella et al, 2014). Nguyen and Le (2024) surveyed 400 Vietnamese bank workers and found that younger employees showed a more positive attitude and behavior towards security awareness than older employees. Additionally, employees with higher levels of education showed a positive association with information security knowledge and behavior (Nguyen & Le, 2024). Security awareness measures have benefits and drawbacks but ultimately should be part of a comprehensive security posture.

Threat Information Sharing measures

Several organizations and initiatives collect and share information related to cyber threats. Per NIST SP 800-150, threat information is any information related to malicious actors and actions, including indicators, alerts, reports, and tactics, techniques, and procedures (Johnson et al., 2016). Sharing threat information along with internally gathered information can improve an organization's security posture, situational awareness, defensive agility, and knowledge maturity (Johnson et al., 2016). However, there are general challenges of threat information sharing such as trust, interoperability, protecting sensitive and classified data, and enabling information consumption (Johnson et al., 2016). There is an opportunity to increase sharing of data and information on phishing websites (Oest et al., 2020). However, Oest et al. (2020) also suggest that malicious actors do not appear to be fearful of current detection measures since phishing websites are so rampant.

Limitations of security measures

There are obvious limitations to preventing and responding to phishing websites, as phishing attacks remain one of the highest reported types of cyber-attack. Machine learning for both technical and content-related classification and identification can help but is still laborious. Even though much academic research showcases the use of machine learning with many positive machine learning algorithms, there are still several limitations. Sabir et al. (2022) conducted a study on machine learning (ML) based approaches and identified practical limitations in ML studies such as phishing datasets, timely detection, and scalability. Lee et al. (2025) identified that even the practical Google Safe Browser did not detect 81.59% of the phishing URLs. So both academic research and industry applications of machine learning techniques have limitations in the detection and identification of phishing websites. Sharing threat information can help prevent and respond to many tactics and techniques of malicious actors. However, due to the sensitive data and the nature of cybersecurity, issues of trust, interoperability, and the protection of data are strong barriers to this measure. Security awareness training is a cost-efficient measure that can reduce engagement and click-through of phishing campaigns; however, as time passes, people forget the training. Conversely, Generative AI is making it more difficult to deal with phishing.

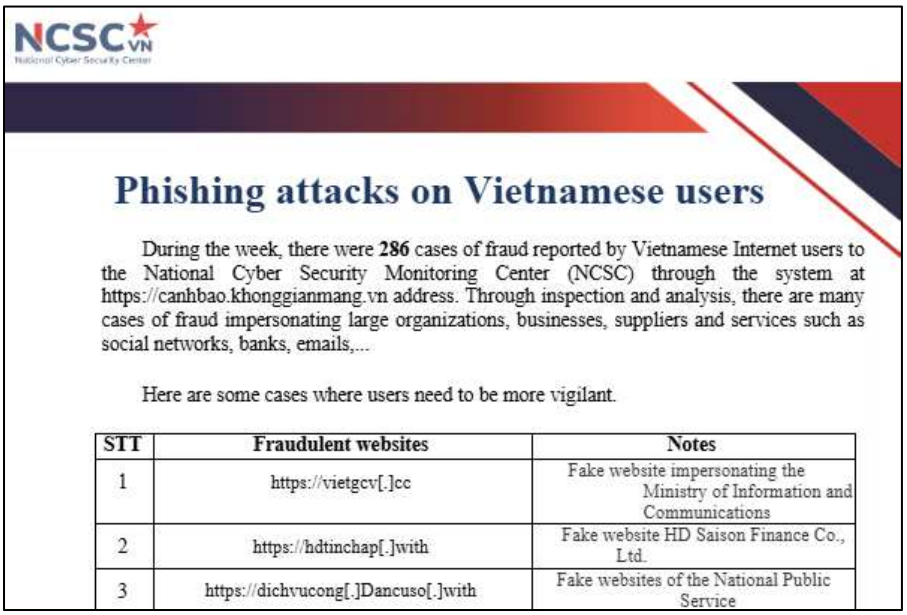
Methodology

Research Design

This study uses quantitative research design and methodology. The research design utilized for this study is an exploratory data analysis of phishing websites related to e-commerce in Vietnam. The methodology includes data collection and cleaning, exploratory data analysis (EDA) through descriptive statistics, box plots, and bar charts. Second, hypotheses are generated based on the EDA. Third, hypotheses will be tested using analysis of variance (ANOVA) and Kruskal-Wallis methods and follow up with Post Hoc Tukey test as needed.

Part 1: Data Collection and Cleaning

The National Cyber Security Center of Vietnam releases a weekly report on cybersecurity. The data for this study were downloaded, translated, and reviewed from weekly public cybersecurity reports dated from February 2022 to December 2024. No public reports were available before February 2022, and public reports became unavailable in 2025 after Vietnamese government restructuring. Only data related to phishing websites were collected from each weekly report. Figure 1 showcases an example screenshot of the translated report.



NCSC VN
National Cyber Security Center

Phishing attacks on Vietnamese users

During the week, there were **286** cases of fraud reported by Vietnamese Internet users to the National Cyber Security Monitoring Center (NCSC) through the system at https://canhbao.khonggianmang.vn address. Through inspection and analysis, there are many cases of fraud impersonating large organizations, businesses, suppliers and services such as social networks, banks, emails,...

Here are some cases where users need to be more vigilant.

STT	Fraudulent websites	Notes
1	https://vietgcv[.]cc	Fake website impersonating the Ministry of Information and Communications
2	https://hdtinchap[.]with	Fake website HD Saison Finance Co., Ltd.
3	https://dichvucong[.]Dancuso[.]with	Fake websites of the National Public Service

Figure 1. NCSC Report Screenshot

The weekly reports included other areas such as vulnerabilities, botnet threats, and other cybersecurity-related public sharing. After data are collected from the government report, the URL character length, protocols, top-level domain, similarity score (Levenshtein distance), duration, and brand impersonation are determined. Lastly, the phishing URLs are submitted to the VirusTotal website to collect additional information such as VirusTotal community score and when the URL was first reported. Figure 2 showcases a screenshot of the VirusTotal Analysis of a phishing URL.

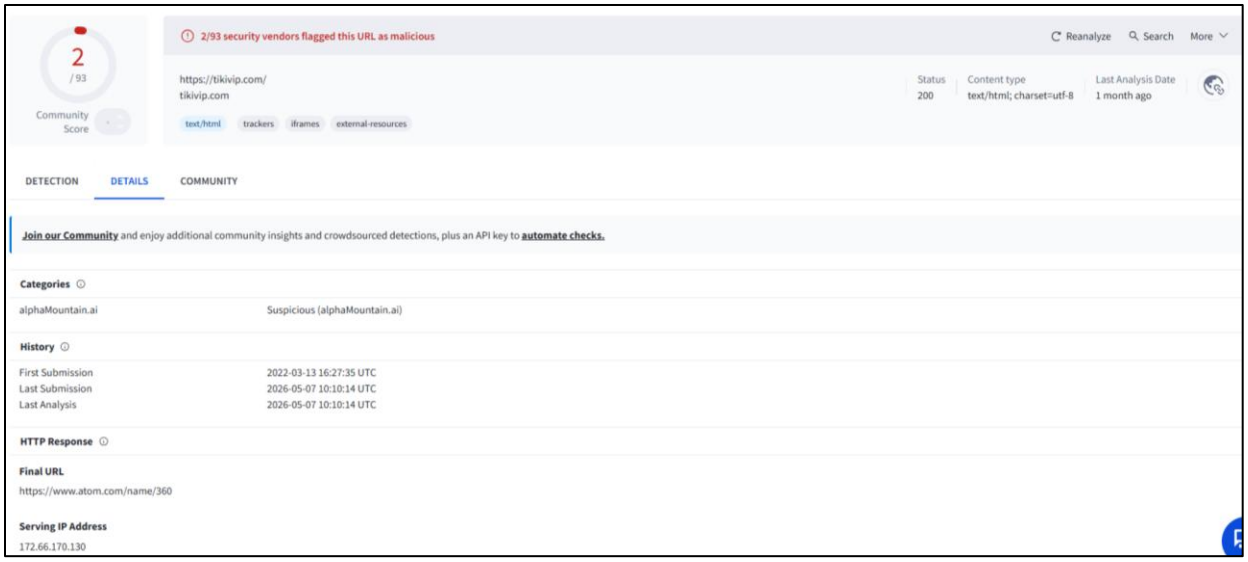


Figure 2: VirusTotal Screenshot

Data Cleaning

Data was reviewed and cleaned. Some websites were not identified by NCSC in the report but updated for classification purposes by the researcher. For example, the website https://alibabavn[.]shop was just identified as a "fake website" by NCSC, but not as a fake website of the Chinese Brand Alibaba. Second, the website data were updated to ensure safety and were not accidentally opened. This included putting brackets around the period between the top-level domain and 2nd-level domains as necessary. Brackets were also put around the real websites for the Levenshtein distance (similarity score) analysis. Lastly, it appeared that some top-level domain names were changed to ".clip" by NCSC for safety and were not actual TLDs and were updated by the researcher to ".club" as part of the data cleaning process. Duplicates were not removed as the reports showcased some of the same websites were re-reported.

Population and Sample

The total population of fraudulent websites in the Vietnamese government reports were 1467 phishing websites from 2022 to 2024. The sample size for this study was 607 phishing websites that were classified as fake e-commerce sites and chosen for this study.

Part 2: Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) aims to examine data for understanding and visualization to generate hypotheses (Komorowski et al, 2016). The EDA process starts with the problem, collects data, analyzes data, decides the imposition of a model, then draws conclusions (NIST/SEMATECH, 2019). EDA techniques can include descriptive statistics, histograms, heatmaps, scatterplots, and boxplots (Komorowski et al., 2016; NIST/SEMATECH, 2019; Sandfeld, 2024). Usually, the analysis of data focuses on descriptive statistics and data visualizations (Sandfeld, 2024). Then, exploratory data analysis allows the data to suggest which quantitative models best fit the data, such as regression or analysis of variance (NIST/SEMATECH, 2019).

EDA was chosen for this study because there is little knowledge and research about Vietnam and e-commerce phishing websites. EDA is a systematic and scientific attempt to discover new information and generalizations about phenomena (Devezer et al., 2021). Research has shown common features and characteristics of phishing website analysis include URL character length, protocols, top-level domain, Levenshtein distance (similarity score), brand impersonation, VirusTotal score, and duration (Alshehhi et al., 2023; Choo et al., 2023; Lee et al., 2025; Linh &

Hung, 2025; Oest et al., 2020; Sabir et al., 2022). The duration is calculated based on the Vietnam reported phishing report date and the VirusTotal first submission date. First, descriptive statistics and EDA data visualizations (box plots, bar charts) are captured for each year for common features of e-commerce phishing websites. Excel with the Data Analysis Tool Pack was utilized to conduct the analysis.

Part 3: Hypotheses Generation and Testing

When EDA is completed, hypotheses were generated based on the initial descriptive statistics and visualizations.

Hypothesis 1

H_0 – The average lengths of Phishing URLs are the same

H_1 – The average lengths of the Phishing URLs are significantly different

Hypothesis 2

H_0 – The average phishing URL Levenshtein scores between 2022, 2023, and 2024 are the same

H_1 – The average phishing URL Levenshtein scores between 2022, 2023, and 2024 are significantly different

Hypothesis 3

H_0 – The average VirusTotal scores of phishing URLs between 2022, 2023, and 2024 are the same

H_1 – The average VirusTotal scores of phishing URLs between 2022, 2023, and 2024 are significantly different.

Hypothesis 4

H_0 – The average duration of phishing URLs between 2022, 2023, and 2024 are the same

H_1 – The average duration of phishing URLs between 2022, 2023, and 2024 are significantly different.

Analysis of variance (ANOVA) is one of the most common statistical methods used in research (Kim, 2017). ANOVA analysis focuses on the differences in variances. Specifically, ANOVA is parametric test designed for three or more groups with normal distributions. Kruskal-Wallis is a nonparametric test designed for three or more groups that are not normally distributed (McKnight & Najab, 2010). Microsoft Excel and Statisty were used to conduct ANOVA and Kruskal-Wallis tests. See Appendix B for the ANOVA equation and explanation.

Following a statistically significant result in an analysis of variance (ANOVA), post hoc Tukey's test is employed to further examine the specific nature of group differences. While ANOVA serves as an omnibus test that determines whether there is evidence to reject the null hypothesis of equal group means, it does not indicate which specific group comparisons are responsible for the observed effect. Conducting multiple pairwise comparisons without adjustment would substantially increase the likelihood of committing a Type I error (Edlund & Nichols, 2023). Vasavada (2016) web-based Post Hoc Tukey calculator is used to analyze whether the ANOVA is statistically significant between the groups.

Results

Part 1 Exploratory Data Analysis

The National Cyber Security Center (NCSC) of Vietnam reported 1467 phishing websites from 2022 to 2024. From that population, 607 phishing websites classified as e-commerce were chosen for this study. The phishing characteristics include URL character length, protocols, top-level domain, similarity score (Levenshtein distance), duration, and brand impersonation.

URL Character Length Descriptive Statistics

The overall average URL length for the sample size was 18.25 characters. In 2022, there were 150 reported e-commerce site impersonations with an average URL character length of 21.44. The distribution was characterized by a positive skewness of 0.88 and kurtosis of 2.06. In 2023, there were 170 e-commerce site impersonations reported with an average URL character length of 12.64. The distribution was characterized by a positive skewness of 1.42 and kurtosis of 2.03. In 2024, there were reported 287 e-commerce site impersonations with an average URL character length of 19.90. The distribution was characterized by a positive skewness of 0.34 and kurtosis of 0.25. The skewness and kurtosis suggest a normal distribution in each year.

Table 2 showcases the descriptive statistics between the years 2022, 2023, and 2024 for Phishing URL Character Length. Figures 3 illustrates the box plots for phishing URL character length for 2022, 2023, and 2024.

Table 2: Descriptive Statistics URL Character Length

Descriptive Statistics	2022	2023	2024
Mean	21.44666667	12.64117647	19.90243902
Standard Error	0.453525752	0.248707887	0.391412476
Median	20	12	20
Mode	20	10	20
Standard Deviation	5.554533388	3.242754112	6.630947862
Sample Variance	30.85284116	10.51545423	43.96946955
Kurtosis	2.063456275	2.029794916	0.248676359
Skewness	0.882409878	1.420184085	0.343081702
Range	36	16	43
Minimum	9	9	5
Maximum	45	25	48
Sum	3217	2149	5712
Count	150	170	287

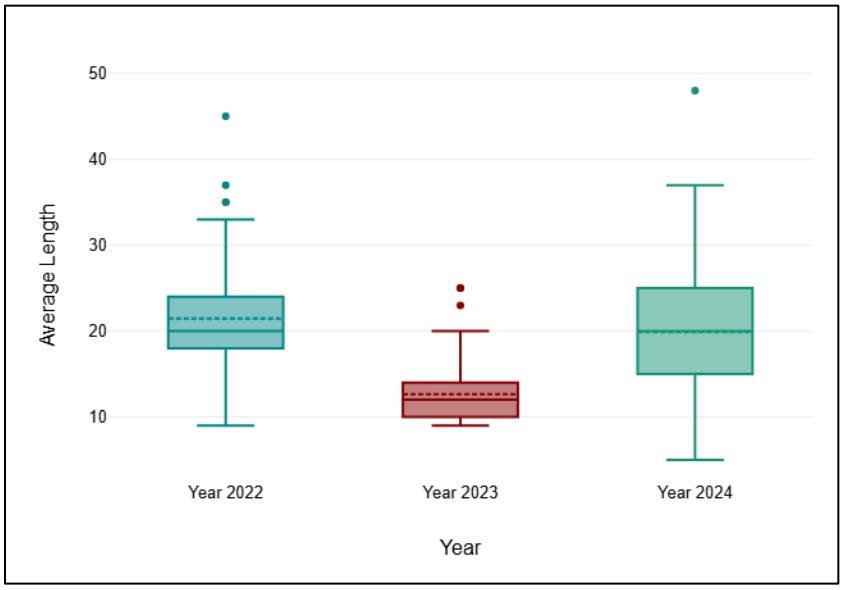


Figure 3 Average URL Length Boxplots

Reviewing the descriptive statistics and the box plots illustrates a potential difference in URL character length of e-commerce phishing websites between 2022, 2023, and 2024. An ANOVA will be used for further analysis.

Protocols

In 2022, there were 150 reported 150 e-commerce site phishing sites utilized the HTTP 8%, HTTPS 80.7%, and the relative protocol 11.3%. In 2023, there were 170 reported 170 e-commerce phishing site protocols used were HTTP 1.8%, HTTPS 0%, and the relative protocol 98.2%. In 2024, 287 reported e-commerce phishing sites used HTTP 0.7%, HTTPS 66.2%, and the relative protocol 33.1%. Table 3 and Figure 6 showcase the types of protocols for Phishing URLs in 2022, 2023, and 2024.

Table 3 Types of Protocols used in Phishing URLs

Type of protocol	2022	2023	2024
HTTP	12	3	2
HTTPS	121	0	190
Relative	17	167	95
Total	150	170	287

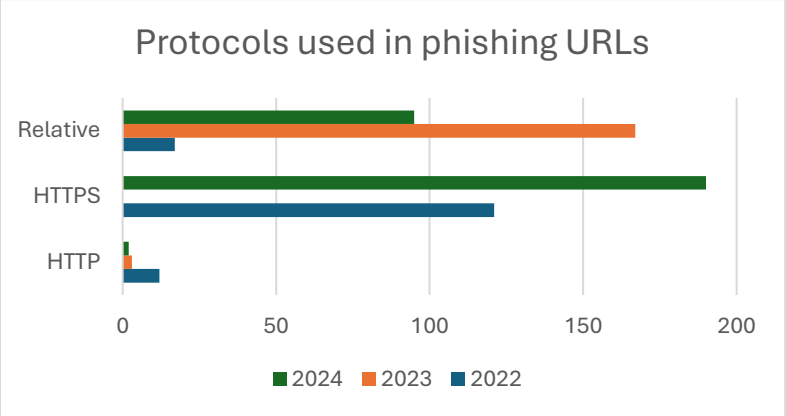


Figure 6 Protocols Bar Chart 2022, 2023, 2024

Top-Level Domains

In 2022, 2023, and 2024, the most widely used top-level domains for phishing websites aimed at Vietnam were ".com". In 2022, ".com" represented 66.66% of phishing ecommerce sites. In 2024, ".com" represented 81.76%. In 2024, ".com" appeared in 71.08% of phishing e-commerce websites. Figure 7 is a bar chart diagram showcasing the TLDs for 2022 to 2024. See Appendix C for the table of TLDs for each year.

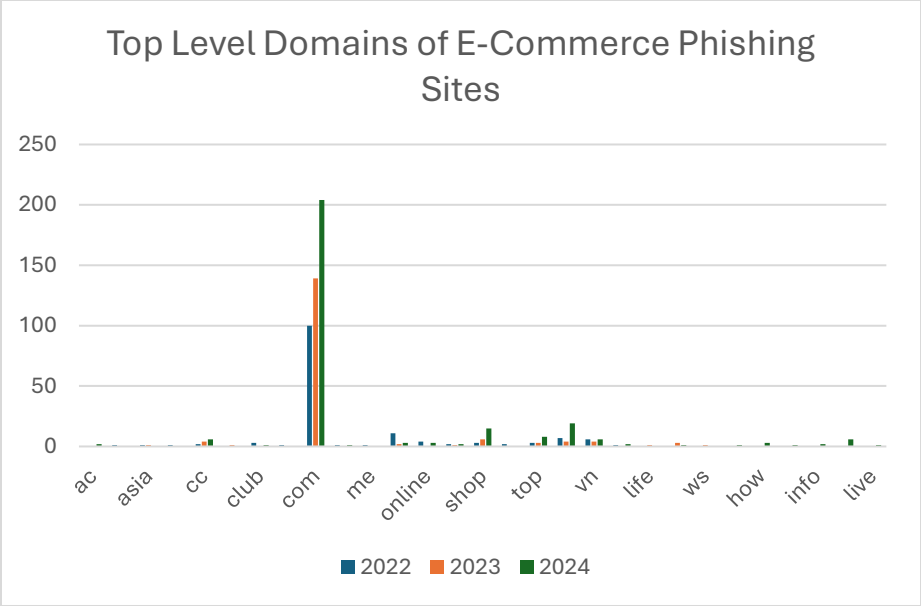


Figure 7 Phishing Site TLD 2002-2024

Levenshtein Distance

Overall the average Levenshtein Distance for the sample size was 13.74. In 2022, there were 150 e-commerce site impersonations reported with an average similarity score of 11.27. The distribution was characterized by a positive skewness of 1.42 and kurtosis of 4.49. This indicates a moderately positively skewed distribution and elevated kurtosis. The distribution deviates slightly from a normal distribution but remains acceptable for subsequent analyses.

In 2023, there were 170 e-commerce site impersonations reported with an average similarity score of 16.54. The distribution was characterized by a positive skewness of 0.32 and kurtosis of -0.93. In 2024, there were reported 287 e-commerce site impersonations with an average similarity score of 13.38. The distribution was characterized by a positive skewness of 0.61 and kurtosis of 0.97. This indicates a slight right skewed distribution and acceptable kurtosis that is considered normally distributed. Table 4 below shows the descriptive statistics for the Levenshtein Distance (similarity score). Figures 8 illustrate the boxplots for 2022, 2023, 2024.

Table 4: Descriptive Statistics Levenshtein Distance

Descriptive Statistics	***2022***	***2023***	***2024***
Mean	11.27333333	16.53529412	13.37630662
Standard Error	0.357332482	0.203885699	0.27742386
Median	11	16	13
Mode	13	15	12
Standard Deviation	4.376411246	2.658344276	4.699858235
Sample Variance	19.15297539	7.066794292	22.08866743
Kurtosis	4.490053498	-0.932157404	0.969285179
Skewness	1.405342819	0.321608107	0.609247188
Range	29	12	34
Minimum	4	11	3

Maximum	33	23	37
Sum	1691	2811	3839
Count	150	170	287

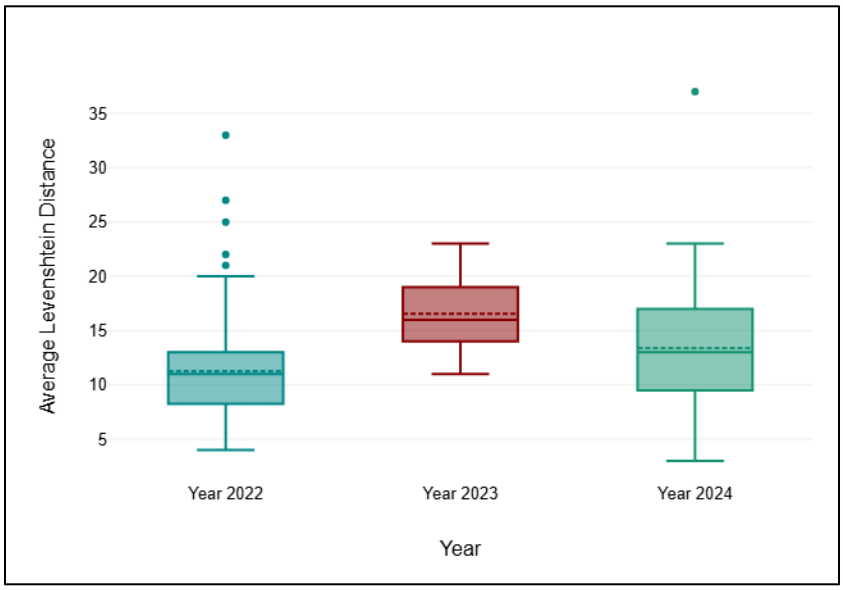


Figure 8 Levenshtein Distance Boxplots

Phishing Reporting Duration

The duration of websites was calculated based on the Vietnam reporting date and VirusTotal first submission date. Overall, the average duration for the sample size was 1.25 days. The average phishing website duration in 2022 was 132.60 days. This suggests that the average phishing websites that were reported in Vietnam were not submitted to VirusTotal for an average of nearly 133 days. The distribution was characterized by a skewness of -0.61 and kurtosis of 5.26. The average phishing website duration in 2023 was -33.22 days. This suggests that the average phishing website in 2023 was reported 33 days in VirusTotal before reporting occurred in Vietnam. The distribution was characterized by a skewness of -4.99 and kurtosis of 45.47. The average phishing website duration in 2024 was -46.99 days. This means that the average phishing website in 2024 was reported about 47 days in VirusTotal before being reported to the Vietnamese government. The distribution was characterized by a skewness of -4.99 and kurtosis of 45.47. This indicates a heavily skewed distribution and excess kurtosis that is not considered normally distributed.

Since the descriptive statistics show extreme skewness, kurtosis, and outliers, ANOVA is not appropriate for further investigation. Thus, a Non Parametric test was conducted with the Kruskal-Wallis Test. Table 5 below shows the descriptive statistics for the Phishing Reporting duration. Figure 9 illustrates a bar chart for 2022, 2023, and 2024.

Table 5: Descriptive Statistics Phishing Reporting Duration

Descriptive Statistics	***2022***	***2023***	***2024***
Mean	132.6	-33.21764706	-46.98606272
Standard Error	42.62609823	23.66594801	18.59776869
Median	2	-25.5	-11
Mode	5	-26	-4

Standard Deviation	522.060952	308.5662104	315.0661821
Sample Variance	272547.6376	95213.1062	99266.69911
Kurtosis	5.267916309	45.47479572	92.75245706
Skewness	-0.608041296	-4.988862451	-8.578021687
Range	3787	3586	4326
Minimum	-2595	-2885	-3910
Maximum	1192	701	416
Sum	19890	-5647	-13485
Count	150	170	287

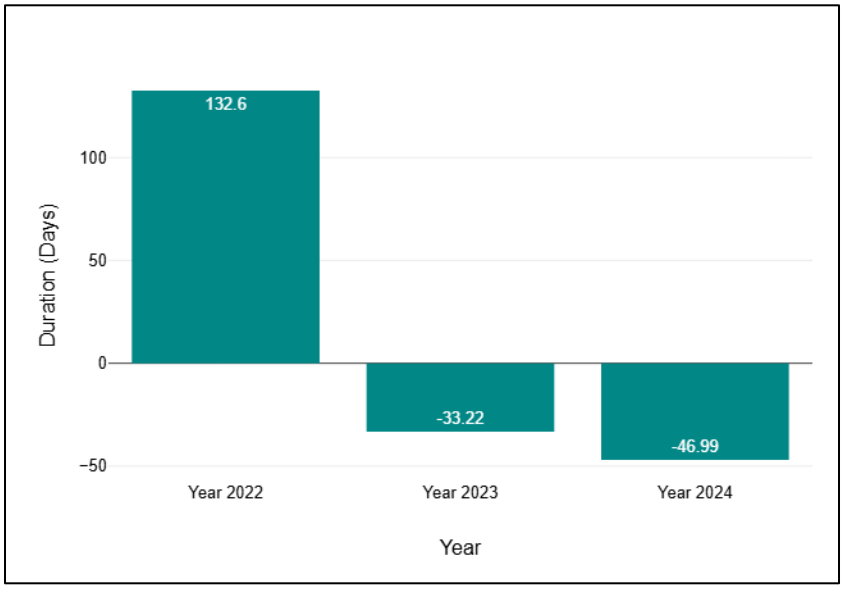


Figure 9 Phishing URL Duration Bar chart

Brand Impersonation

Out of 1467 phishing websites, 607 were e-commerce sites, representing about 41% of the total phishing websites reported by the Vietnamese government. In 2022, Shopee was impersonated 48 times. Dien May Xanh was impersonated 29 times. Tiki was impersonated 24 times, Lazada was impersonated 21 times, and Sendo Ecommerce was impersonated 16 times. In 2023, Tiki was impersonated 64 times, Shopee was impersonated 47 times, Lazada was impersonated 21 times, Amazon was impersonated 13 times, and Dien May Xanh was impersonated 10 times. In 2024, Tiki was impersonated 86 times, Shopee was impersonated 81 times, Lazada was impersonated 39 times, Amazon was impersonated 29 times, and Dien May Xanh was impersonated 21 times. Appendix D identifies the brands impersonated over the three years. Figure 10 illustrates the number of impersonated brand websites.

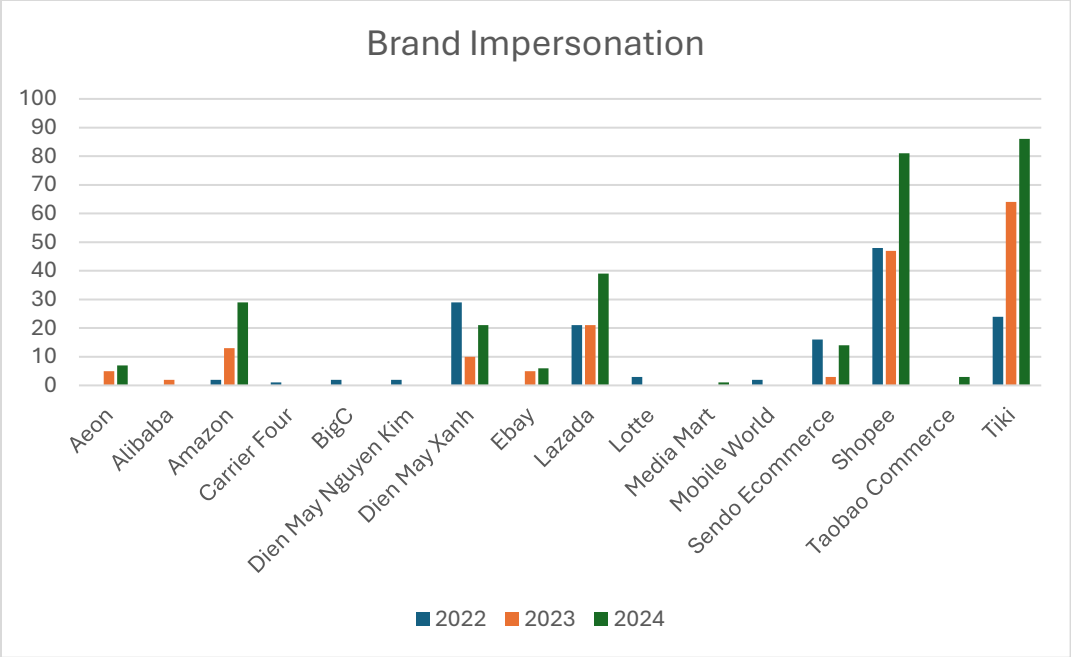


Figure 10 Impersonated Brand Websites

VirusTotal Community Score

The reported Vietnamese e-commerce phishing sites were submitted to VirusTotal, and each submission received a community score based on the number of security vendors that identified the website as malicious. The highest number of security vendors is 96. Overall, 215 out of the sample size of 607 were not flagged by the security vendors of VirusTotal. In 2022, 36 of reported websites were not flagged as malicious, 79 were not reported as malicious in 2023, and 100 were not reported as malicious in 2024.

The overall average VirusTotal score was 3.25. In 2022, the average malicious community score was 3.94. The distribution was characterized by a positive skewness of 0.55 and kurtosis of -0.76. In 2023, the average malicious community score was 3.15. The distribution was characterized by a skewness of -0.98 and kurtosis of 0.82. In 2024, the average malicious community score was 2.94. The distribution was characterized by a positive skewness of 1.35 and kurtosis of 0.83. The skew and kurtosis over each year is normally distributed. Table 7 below shows the descriptive statistics for the VirusTotal Community Score. Table 8 tabulates the number of malicious sites with a VT community score of 0. Figure 11 illustrate the boxplots for 2022, 2023, and 2024.

Table 7 VirusTotal Community Score Descriptive Statistics

Descriptive Statistics	***2022***	**2023**	***2024***
Mean	3.94	3.147058824	2.93728223
Standard Error	0.289600512	0.305952293	0.226604319
Median	3	1	1
Mode	0	0	0
Standard Deviation	3.546867415	3.989129852	3.838920612
Sample Variance	12.58026846	15.91315698	14.73731147
Kurtosis	-0.767874934	-0.983036725	0.827257379
Skewness	0.557704025	0.829854648	1.353334876

Range	13	13	16
Minimum	0	0	0
Maximum	13	13	16
Sum	591	535	843
Count	150	170	287

Table 8 VirusTotal # of malicious sites with a community score of 0

Year	# of URLs not flagged malicious by VT	Total for year	%
2022	36	150	24%
2023	79	170	46%
2024	100	287	35%
Total	215	607	35%

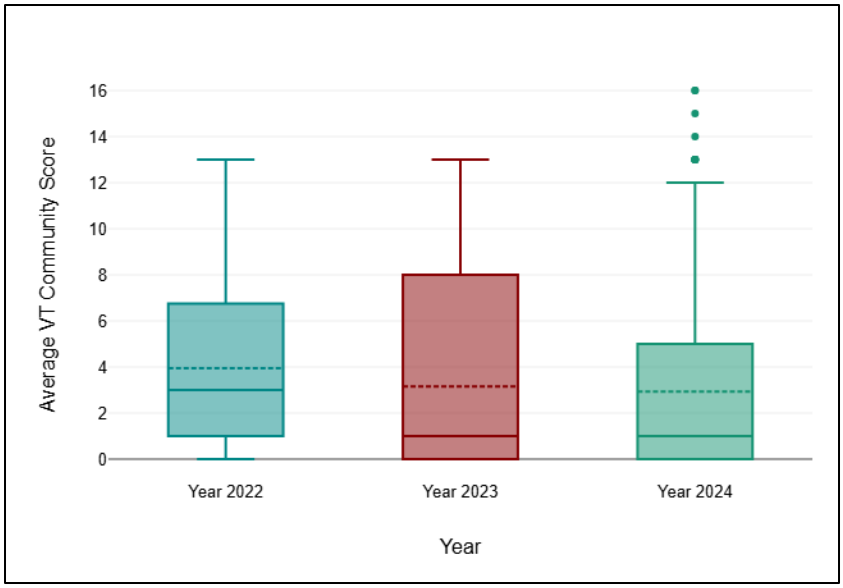


Figure 11 VirusTotal Score Boxplots

Part 2 Hypothesis Testing and Analysis

Hypothesis 1

A one-way ANOVA was conducted to examine differences in the URL character length across the three years (2022, 2023, and 2024). The results showed a statistically significant effect of year, $F [2, 604] = 121.68$, $p < .001$, indicating that at least one group pair mean differed significantly. See Table 9 for the one-way ANOVA. Further examination with the Post Hoc Tukey test indicated that all pairwise differences were statistically significant. Group pair 2022 vs 2023 was significantly different, $Q = 19.79$, $p < 0.01$. Group pair 2022 vs 2024 was significantly different, $Q = 3.83$, $p < 0.05$. Group pair 2023 vs 2024 was significantly different, $Q = 18.92$, $p < 0.01$. See Table 10 for the Post Hoc Tukey Test. Therefore, we reject the null hypothesis that the average phishing URL character lengths between 2022, 2023, and 2024 are the same.

Table 9 ANOVA Average Length of Phishing URLs

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	7651.719	2	3825.86	121.6783	3.96E-45	3.01064
Within Groups	18991.21	604	31.44241			
Total	26642.93	606				

Table 10 Post Hoc Tukey Test Average Length

Group Pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
2022 vs 2023	19.7946	0.0010053	** p<0.01
2022 vs 2024	3.8322	0.0189561	* p<0.05
2023 vs 2024	18.9224	0.0010053	** p<0.01

Hypothesis 2

A one-way ANOVA was conducted to examine differences in the URL Levenshtein distance across the three years (2022, 2023, and 2024). The results showed a statistically significant effect of year, F [2, 604] = 66.40, p < .001, indicating that at least one group pair mean differed significantly. See Table 11 for ANOVA. Further examination with the Post Hoc Tukey test indicated that all pairwise differences were statistically significant. Group pair 2022 and 2023 were statistically significant, Q = 16.04, p < 0.01. Group pair 2022 and 2024 were statistically significant, Q = 7.13, p < 0.01. Group pair 2023 and 2024 were statistically significant, Q = 11.14, p < 0.01. See Table 12 for the post hoc Tukey test. Therefore, we reject the null hypothesis that the average phishing URL Levenshtein Distance scores between 2022, 2023, and 2024 are the same.

Table 11 ANOVA Average Levenshtein Distance

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2278.952	2	1139.476	66.39789	8.6E-27	3.01064
Within Groups	10365.44	604	17.16133			
Total	12644.39	606				

Table 12 Post Hoc Tukey Test Levenshtein Distance

Group Pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
2022 vs 2023	16.0355	0.0010053	** p<0.01
2022 vs 2024	7.1256	0.0010053	** p<0.01
2023 vs 2024	11.1428	0.0010053	** p<0.01

Hypothesis 3

A one-way ANOVA was conducted to examine differences in VirusTotal community scores between 2022, 2023, and 2024. The data suggest that the average VirusTotal scores between 2022, 2023, and 2024 are ***significantly***

different, F [2, 604] = 3.48, P < 0.01. See Table 13. This indicates that at least one group pair differed significantly. Additional analysis with the Post Hoc Tukey test indicated that only one group pair, 2022 vs 2024, was statistically significant. The other group pairs were not statistically significant. Therefore, we reject the null hypothesis that the average VirusTotal community scores between 2022 and 2024 are the same. See Table 14.

Table 13 ANOVA Average VirusTotal Community Score

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	101.2597	2	50.62986	3.483499	0.031318	3.01064
Within Groups	8778.655	604	14.5342			
Total	8879.914	606				

Table 14 Post Hoc Tukey Test VirusTotal Community Score

Group Pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
2022 vs 2023	2.6258	0.1524932	insignificant
2022 vs 2024	3.6919	0.0250787	* p<0.05
2023 vs 2024	0.8041	0.818637	insignificant

Hypothesis 4

A Kruskal-Wallis test was conducted to examine the differences in phishing website reporting duration across 2022, 2023, and 2024. The non-parametric approach was used because the descriptive statistics exhibited extreme skewness and excess kurtosis. Kruskal-Wallis test revealed a *statistically significant difference* in phishing website duration between the three years, H (2) = 68.41, p < 0.001, $\eta^2 = 0.11$. Therefore, we reject the null hypothesis that the average duration of phishing URLs between 2022, 2023 and 2024 are the same. This is validated with the Dunn-Bonferroni Test. Table 15 is the summary of the duration ranks, Table 16 shows the result of the Kruskal-Wallis test, Table 17 shows the Dunn-Bonferroni test.

Table 15 Duration Ranks

Groups	n	Median	Mean Rank
Year 2022	150	2	398.92
Year 2023	170	-25.5	239.09
Year 2024	287	-11	292.84
Total	607	-12	

Table 16 Kruskal-Wallis Test

Chi^2	df	p
68.41	2	<.001

Table 17 Dunn-Bonferroni Test

Group Pair	Test Statistic	Std. Error	Std. Test Statistic	p	Adj. p
Year 2022 - Year 2023	159.83	19.64	8.14	<.001	<.001

Year 2022 - Year 2024	106.08	17.67	6	<.001	<.001
Year 2023 - Year 2024	-53.75	16.97	-3.17	.002	.005

Adj. p: Values adjusted with Bonferroni correction.

Discussion

Brand Impersonation

Malicious actors tend to target the top e-commerce brands in Vietnam. Shopee, Tiki, Lazada, Dien May Xanh, and Amazon were the most frequently impersonated e-commerce brands in Vietnam. Shopee had 176 phishing websites, Tiki had 174, Lazada had 81, Dien May Xanh had 60, and Amazon had 44 over three years, according to the descriptive statistics. Hoa (2024) noted that Shopee, Lazada, Tiki, TikTok Shop, and Sendo were the top e-commerce organizations operating in Vietnam. The top three align with expectations that the biggest e-commerce organizations operating in Vietnam also have the most impersonating websites. However, it is noted that reporting phishing websites and fraud is very low due to several reasons, including a lack of law enforcement. Law enforcement's lack of cybercrime knowledge and digital investigation skills leads to low investigation and conviction rates (Curis and Oxburgh, 2022). Additionally, malicious domains are difficult to deal with since the website owner could unknowingly be exploited by a malicious actor (Pritom et al., 2022).

Domain Spoofing and Obfuscation

The results suggest that there were differences between the characteristics over the three years. The differences potentially be attributed to changing tactics for spoofing and obfuscation. The findings suggest malicious actors tended to use the HTTPS and Relative protocols the most, as well as the ".com" top-level domain in their fake websites. Most likely trying to deceive victims with common website taxonomies. Most organizations utilize HTTPS for security reasons rather than HTTP. However, malicious actors will also employ this protocol to appear more legitimate; a study by Skula and Kvet noted that phishing sites utilized the HTTPS protocol between 70-80% of the time (2024). This study noted HTTPS was used 80.7% in 2022, 0% in 2023, then 66.2% in 2024. However, it appears that the malicious actors may have tried a different tactic in 2023 by using the relative protocol to focus more on obfuscation rather than increasing the credibility of their fake website.

The ".com" top-level domain is still the most popular (IANA, 2025; Agarwal & Vasek, 2025). This study showed ".com" was used 66.66% in 2022, 81.76% in 2023, and 71.08% in 2024. Agarwal and Vasek (2025) study showed that ".com" was used in 55.18% of domains they reviewed. The differences can be potentially explained by malicious actors that were experimenting with different TLDs or wanted to purchase cheaper domains (Agarwal & Vasek, 2025).

The average phishing URL character length of this study was approximately 18 characters. Ghalechyan et al. (2024) noted that it is harder to find phishing patterns with short URLs (less than 50 characters) than those with long URLs (more than 50 characters). This study also found that the average Levenshtein Distance score was 13.74. This finding was higher than Zouina and Ottaj (2017) study of 1000 phishing websites, which suggested the average Levenshtein distance score was eight characters more than legitimate websites. The average VirusTotal score was 3.24 with 215 phishing URLs that were not flagged as malicious.

Duration

The average duration between reporting by the Vietnamese government and the VirusTotal submission was 1.25 days over the three years 2022 through 2024. This finding aligns with the range that the academic literature suggests of 0.5 days to 11.94 days (Agarwal & Vasek, 2025; Drury et al., 2022; Lee et al., 2025; Mohammed et al., 2015; Moore & Clayton, 2007; Oest et al., 2020). However, each year varied immensely, with 2022 averaging 132.6 days

after it was reported by the Vietnamese government. But in 2023 and 2024, it was submitted to VirusTotal on average 33 days and 47 days before reported by the Vietnamese government. Norris et al. (2019) noted the difficulties that local governments have with identifying and reporting cyber threats and attacks due to budget and skills to support this effort. However, Lanz (2022) suggested that governments and organizations are increasing alerts and reports regarding cyber-related threats and activities, most likely due to stronger information sharing in the cyber-security community.

Generative AI and Phishing

Generative AI has been a consistent trend in organizations and society over the past three years. Developing a generative AI literate workforce can facilitate informed decision-making with ethical and legal awareness (Annapureddy et al., 2025). However, malicious actors have already utilized generative AI to create deepfake videos, audio, images, and text for phishing and social engineering attacks (Annapureddy et al., 2025; Schmitt & Koutroumpis, 2025). Schmitt and Koutroumpis (2025) specify that large language models (LLMs) pose a notable threat to the cybersecurity landscape. The authors suggest cybersecurity threats are enhanced and amplified with generative AI. Specifically, generative AI can enhance and amplify content generation, personalization, and scalability of cyber-attacks. This may increase the velocity of phishing websites, along with making it more difficult to identify malicious websites. On the other hand, generative AI has also been adopted by cybersecurity researchers and practitioners. Patel et al. (2025) specify that generative AI models have been used for anomaly detection, threat simulation, and creating synthetic data for cybersecurity. Beydemir et al. (2024) employed generative AI for a phishing email detection research study.

Future Research

The findings of this exploratory study open several avenues for further investigation into the evolving landscape of digital fraud in Southeast Asia. Future research can focus on regional and other sector comparisons, evaluation of real-time phishing detection systems, and generative AI. This research focused exclusively on e-commerce. Future studies could expand this methodology to other high-risk sectors in Vietnam, such as banking, social media, or government services, to see if malicious actors employ different tactical patterns when targeting different types of sensitive data. This could include collaboration with private brand protection firms or utilizing global repositories like PhishTank or OpenPhish to conduct comparative studies between Vietnam and other rapidly growing digital economies in Southeast Asia. This study found that many phishing sites were not initially flagged by scanners like VirusTotal. Future work should evaluate the effectiveness of real-time, region-specific detection algorithms that account for the unique characteristics of the Vietnamese digital landscape, such as the frequent use of the ".com" TLD and very short URL lengths. Generative AI is significantly increasing the velocity, personalization, and scalability of cyber-attacks. Future research should empirically analyze how these tools are being used to create more convincing Vietnamese-language phishing content, moving beyond URL structure to examine the quality of deepfake images, audio, and text used to deceive consumers.

Limitations

While this exploratory study provides significant insights into the e-commerce phishing landscape in Vietnam, several limitations must be acknowledged. The study relied exclusively on weekly reports from the National Cyber Security Center (NCSC) of Vietnam. Consequently, the data is restricted to a specific timeframe (February 2022 to December 2024) and does not capture the full universe of phishing activity in the country. Furthermore, due to Vietnamese government restructuring in 2025, public reports became unavailable, preventing a longer-term longitudinal analysis. Second, the sample may be subject to reporting bias, as phishing and fraud reporting in Vietnam is disproportionately low compared to more developed countries like Singapore and the United States. This can be attributed to a few factors, including a lack of specialized cybercrime knowledge and digital investigation skills within local law enforcement, which leads to lower investigation and conviction rates. Third, the data sources

were limited to Vietnamese government reports. Alternatives data sources can be used such a PhishTank or private Vietnamese organizations to continue longitudinal work.

Conclusion

This study provides a comprehensive exploratory analysis of the e-commerce phishing landscape in Vietnam from 2022 to 2024. As Vietnam continues its rapid transition into a digital economy, with e-commerce growth reaching 25% in 2023, the findings highlight that malicious actors are actively exploiting this expansion.

Malicious actors primarily target the most prominent e-commerce brands in the country, specifically Shopee, Tiki, and Lazada. To deceive consumers, phishers frequently utilize the ".com" top-level domain and employ HTTPS and relative protocols to mimic the legitimacy of established website taxonomies. Furthermore, the study found that phishing URLs in Vietnam are notably short, averaging approximately 18 characters, which may be a strategic choice to evade detection patterns that typically look for longer URLs.

Statistical analysis revealed significant differences across 2022, 2023, and 2024 regarding average URL character lengths, Levenshtein distance (similarity scores), and VirusTotal community scores. These variations suggest that phishers are continuously adapting their tactics to potentially prioritize obfuscation over fake credibility.

The findings underscore a critical gap in current cybersecurity defenses. Such as 100 reported cases in 2024 were not initially flagged as malicious by third-party scanners highlights the limitations of existing technical countermeasures. With the advent of generative AI, which increases the velocity and sophistication of content generation and personalization, the difficulty of identifying these threats will only intensify.

This study emphasized that defending Vietnam’s digital economy requires a multi-faceted approach. Technical measures must be complemented by enhanced security awareness, targeted training for employees and consumers, and a more robust threat information sharing framework. Furthermore, addressing the low rates of cybercrime investigation by improving the digital investigation skills of local law enforcement is essential to deterring malicious actors in the region. This foundational research serves as a starting point for future longitudinal studies to monitor the ongoing evolution of social engineering in Vietnam's increasingly digital society.

References

Agarwal, S., & Vasek, M. (2025). Examining Newly Registered Phishing Domains at Scale. In Proceedings of 24th Workshop on the Economics of Information Security (WEIS2025). https://discovery.ucl.ac.uk/id/eprint/10209951/1/CDA_Domains___WEIS_25.pdf

Akram, M., Barker, W., Clatterbuck, R., Dodson, D., Everhart, B., Gilbert, J., Haag, W., Johnson, B., Kapasouris, A., Lam, D., Pleasant, B., Raguso, M., Souppaya, M., Symington, S., Turner, P., & Wilson, C. (2020). Securing Web Transactions TLS Server Certificate Management Includes Executive Summary (A); Security Risks and Recommended Best Practices (B); Approach, Architecture, and Security Characteristics (C); and How-To Guides (D). https://doi.org/10.6028/NIST.SP.1800-16

Afroz, S., & Greenstadt, R. (2011). PhishZoo: Detecting Phishing Websites by Looking at Them. 2011 IEEE Fifth International Conference on Semantic Computing. https://doi.org/10.1109/icsc.2011.52

Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing Attacks: a Recent Comprehensive Study and a New Anatomy. Frontiers in Computer Science, 3(1), 1–23. https://doi.org/10.3389/fcomp.2021.563060

Almomani, A., Alauthman, M., Shatnawi, M. T., Alweshah, M., Alrosan, A., Alomoush, W., Gupta, B. B., Gupta, B. B., & Gupta, B. B. (2022). Phishing Website Detection With Semantic Features Based on Machine Learning Classifiers: A Comparative Study. International Journal on Semantic Web and Information Systems (IJSWIS), 18(1), 1-24. https://doi.org/10.4018/IJSWIS.297032

Alshehhi, S., Aldhaheri, H., Alzaabi, A., Alhammadi, S., Alhosani, A., & Koci, A. (2023, December). Analytical Approach for Mitigating Phishing Attempts. In International Conference on Robotics and Networks (pp. 195-206). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64495-5_15

Annapureddy, R., Fornaroli, A., and Gatica-Perez, D. 2025. Generative AI Literacy: Twelve Defining Competencies. Digit. Gov.: Res. Pract. 6, 1, Article 13 (March 2025), 21 pages. https://doi.org/10.1145/3685680

Ayeni, R. K., Adebiyi, A. A., Okesola, J. O., & Igbekele, E. (2024). Phishing Attacks and Detection Techniques: A Systematic Review. 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG), 1–17. https://doi.org/10.1109/seb4sdg60871.2024.10630203

Beydemir, A. B., Sezgin, U., Doğan, U., Aşıklar, B. E., Yerlikaya, F. A., & Bahtiyar, Ş. (2024). A Dynamically Selected GPT Model for Phishing Detection. 2024 14th International Conference on Advanced Computer Information Technologies (ACIT), 481–484. https://doi.org/10.1109/acit62333.2024.10712553

Braun, O., Hörnemann, J., Pohlmann, N., Urban, T., & Grosse-Kampmann, M. (2025). Different Seas, Different Phishes – Large-Scale Analysis of Phishing Simulations Across Different Industries. Proceedings of the 20th ACM Asia Conference on Computer and Communications Security, 1520–1534. https://doi.org/10.1145/3708821.3733905

Burda, P., Luca Allodi, & Zannone, N. (2023). Cognition in Social Engineering Empirical Research: a Systematic Literature Review. ACM Transactions on Computer-Human Interaction. https://doi.org/10.1145/3635149

"Cảnh báo tuần | Trung tâm Giám sát an toàn không gian mạng quốc gia - National Cyber Security Center (NCSC)," Soc.gov.vn, 2022. https://soc.gov.vn/canhbaoattt/

Carella, A., Kotsoev, M., & Truta, T. M. (2017). Impact of security awareness training on phishing click-through rates. 2017 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/bigdata.2017.8258485

Charmet, F., Morikawa, T., Tanaka, A., & Takahashi, T. (2024). VORTEX : Visual phishing detectiOns aRe Through EXplanations. ACM Transactions on Internet Technology, 24(2), 1–24. https://doi.org/10.1145/3654665

Choo, E., Nabeel, M., Kim, D., Ravindu De Silva, Yu, T., & Khalil, I. (2023). A Large Scale Study and Classification of VirusTotal Reports on Phishing and Malware URLs. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 7(3), 1–26. https://doi.org/10.1145/3626790

Curtis, J., & Oxburgh, G. (2022). Understanding Cybercrime in "Real World" Policing and Law Enforcement. The Police Journal: Theory, Practice and Principles, 96(4), 573–592. https://doi.org/10.1177/0032258x221107584

Dawkins, S., Jacobs, J. (2023). NIST Phish Scale User Guide. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Series TN 2276. https://doi.org/10.6028/NIST.TN.2276

Desolda, G., Ferro, L. S., Marrella, A., Catarci, T., & Costabile, M. F. (2022). Human Factors in Phishing Attacks: A Systematic Literature Review. ACM Computing Surveys, 54(8), 1–35. https://doi.org/10.1145/3469886

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3). https://doi.org/10.1098/rsos.200805

Dodson, D., Montgomery, D., Polk, T., Ranganathan, M., Souppaya, M., Johnson, S., Kadam, A., Pratt, C., Thakore, D., Walker, M., Lear, E., Weis, B., Barker, W. C., Coclin, D., Hojjati, A., Wilson, C., Jones, T., Baykal, A.,

Cohen, D., & Yeich, K. (2021). Securing Small-Business and Home Internet of Things (IoT) Devices: Mitigating Network-Based Attacks Using Manufacturer Usage Description (MUD). *Securing Small-Business and Home Internet of Things (IoT) Devices*. https://doi.org/10.6028/nist.sp.1800-15

Drury, V., Lux, L., & Meyer, U. (2022). Dating Phish: An Analysis of the Life Cycles of Phishing Attacks and Campaigns. Proceedings of the 17th International Conference on Availability, Reliability and Security. https://doi.org/10.1145/3538969.3538997

Dwight, J., Nguyen, L. P., & Ingram, G. (2025). E-Commerce Offender Script Analysis: A Novel Approach to Profile Cyber Offenders in E-Commerce. Deviant Behavior, 1–20. https://doi.org/10.1080/01639625.2025.2526646

Edlund, J. E., & Nichols, A. L. (2023). The Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences. In A. L. Nichols (Ed.), Cambridge University Press eBooks (1st ed., Vol. 1). Cambridge University Press. https://doi.org/10.1017/9781009010054

Ghalechyan, H., Israyelyan, E., Arakelyan, A. et al. Phishing URL detection with neural networks: an empirical study. Sci Rep 14, 25134 (2024). https://doi.org/10.1038/s41598-024-74725-6

Hamid, I. R. A., & Abawajy, J. H. (2014). An approach for profiling phishing activities. Computers & Security, 45, 27-41. https://doi.org/10.1016/j.cose.2014.04.002

He, D., Liu, Z., Lv, X., Chan, S., & Guizani, M. (2024). On Phishing URLs Detection Using Feature Extension. IEEE Internet of Things Journal, 1–1. https://doi.org/10.1109/jiot.2024.3446894

Hoa, P. (2024, February 3). Revenue at top 5 e-commerce platforms in Vietnam up 53.4% in 2023. Nhịp Sống Kinh Tế Việt Nam & Thế Giới. https://vneconomy.vn/revenue-at-top-5-e-commerce-platforms-in-vietnam-up-53-4-in-2023.htm

Internet Assigned Numbers Authority. (2025). *IANA — Root Zone Database*. Public Technical Identifiers. https://www.iana.org/domains/root/db

Johnson, C. S., Badger, M. L., Waltermire, D. A., Snyder, J., & Skorupka, C. (2016). Guide to Cyber Threat Information Sharing. *Guide to Cyber Threat Information Sharing*, *800-150*. https://doi.org/10.6028/nist.sp.800-150

Kim, T. K. (2017). Understanding one-way ANOVA Using Conceptual Figures. Korean Journal of Anesthesiology, 70(1), 22–26. NCBI. https://doi.org/10.4097/kjae.2017.70.1.22

Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y. (2016). Exploratory Data Analysis. In: Secondary Analysis of Electronic Health Records. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_15

Kustiawan, Y. A., & Ghauth, K. I. (2025). Evaluating the Impact of Feature Engineering in Phishing URL Detection: A Comparative Study of URL, HTML, and Derived Features. IEEE Access, 13, 126756–126768. https://doi.org/10.1109/access.2025.3579223

Lanz, Z. (2022). Cybersecurity Risk in U.S. Critical Infrastructure: An Analysis of Publicly Available U.S. Government Alerts and Advisories. International Journal of Cybersecurity Intelligence & Cybercrime: 5(1), 43-70. https://doi.org/10.52306/FWOZ7041

Lee, K., Lim, K., Kim, H., Kwon, Y., & Kim, D. (2025). 7 Days Later: Analyzing Phishing-Site Lifespan After Detected. 945–956. https://doi.org/10.1145/3696410.3714678

Linh, D. M., & Hung, T. C. (2025). A feature-engineered dataset of benign and phishing URLs for machine learning and large language models evaluation. Data in Brief, 63, 112162. https://doi.org/10.1016/j.dib.2025.112162

McKight, P.E. and Najab, J. (2010). Kruskal-Wallis Test. In The Corsini Encyclopedia of Psychology (eds I.B. Weiner and W.E. Craighead). https://doi.org/10.1002/9780470479216.corpsy0491

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Tutorial and critical analysis of phishing websites methods. Computer Science Review, 17, 1–24. https://doi.org/10.1016/j.cosrev.2015.04.001

Moore, T., & Clayton, R. (2007, October). Examining the impact of website take-down on phishing. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit* (pp. 1-13). https://doi.org/10.1145/1299015.1299016

Nguyen, B. H., & Le, H. N. Q. (2024). Investigation on information security awareness based on KAB model: the moderating role of age and education level. Information & Computer Security. https://doi.org/10.1108/ics-09-2023-0152

Nguyen, N. A. (2022). Understanding the socialist-market economy in Vietnam. *Emerging Science Journal*, *6*(5), 952-966. https://doi.org/10.28991/ESJ-2022-06-05-03

NIST/SEMATECH e-Handbook of Statistical Methods. (2019). National Institute of Science and Technology. https://doi.org/10.18434/M32189

Norris, D. F., Mateczun, L., Joshi, A., & Finin, T. (2019). Cyberattacks at the Grass Roots: American Local Governments and the Need for High Levels of Cybersecurity. Public Administration Review, 79(6), 895–904. https://doi.org/10.1111/puar.13028

Oest, A., Zhang, P., Wardman, B., Nunes, E., Burgis, J., Zand, A., ... & Ahn, G. J. (2020). Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. https://www.usenix.org/conference/usenixsecurity20/presentation/oest-sunrise

Patel, A., Ram Chandra Sachan, Hariharan Ragothaman, Sheth, A., Pandey, P., & Saai Krishnan Udayakumar. (2025). Leveraging Generative AI for Proactive Cybersecurity Threat Detection in Cloud Environments. 80–85. https://doi.org/10.1109/icict64582.2025.00019

Pritom, M. M. A., & Xu, S. (2022). Supporting Law-Enforcement to Cope with Blacklisted Websites: Framework and Case Study. 2022 IEEE Conference on Communications and Network Security (CNS), 181–189. https://doi.org/10.1109/cns56114.2022.9947260

Rose, S., Liu, C., & Gibson, R. (2025). *Secure Domain Name System (DNS) Deployment Guide*. https://doi.org/10.6028/nist.sp.800-81r3.ipd

Sabir, B., Babar, M. A., Gaire, R., & Abuadbba, A. (2022). Reliability and Robustness analysis of Machine Learning based Phishing URL Detectors. IEEE Transactions on Dependable and Secure Computing, 1–18. https://doi.org/10.1109/tdsc.2022.3218043

Sandfeld, S. (2024). Exploratory Data Analysis. In: Materials Data Science. The Materials Research Society Series. Springer, Cham. https://doi.org/10.1007/978-3-031-46565-9_9

Schmitt, M., & Pantelis Koutroumpis. (2025). Cyber Shadows: Neutralizing Security Threats with AI and Targeted Policy Measures. IEEE Transactions on Artificial Intelligence, 1–9. https://doi.org/10.1109/tai.2025.3527398

Skula, I., & Kvet, M. (2024). URL and Domain Obfuscation Techniques - Prevalence and Trends Observed on Phishing Data. IEEE Xplore. https://doi.org/10.1109/SAMI60510.2024.10432841

Subramani, K., Melicher, W., Starov, O., Vadrevu, P., & Perdisci, R. (2022). PhishInPatterns. Proceedings of the 22nd ACM Internet Measurement Conference. https://doi.org/10.1145/3517745.3561467

Tian, K., Jan, S. T. K., Hu, H., Yao, D., & Wang, G. (2018). Needle in a Haystack. Proceedings of the Internet Measurement Conference 2018. https://doi.org/10.1145/3278532.3278569

Tran, D.V., Nguyen, P. V., Linh Phuong Le, & Thi, S. (2024). From awareness to behaviour: understanding cybersecurity compliance in Vietnam. International Journal of Organizational Analysis. https://doi.org/10.1108/ijoa-12-2023-4147

WHOIS. (2025). WHOIS Search, Domain Name, Website, and IP Tools - Who.is. Who.is. https://who.is/

Van, L. T. H., Guzikova, L., & Nguyen, A. T. (2021, October). Evaluation of e-commerce impact on sustainable economic growth: The case of Vietnam. In *International Scientific Conference on Innovations in Digital Economy* (pp. 101-111). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-14985-6_7

van Liebergen, K., Caballero, J., Kotzias, P., Gates, C. (2023). A Deep Dive into the VirusTotal File Feed. In: Gruss, D., Maggi, F., Fischer, M., Carminati, M. (eds) Detection of Intrusions and Malware, and Vulnerability Assessment. DIMVA 2023. Lecture Notes in Computer Science, vol 13959. Springer, Cham. https://doi.org/10.1007/978-3-031-35504-2_8

Vasavada, N. (2016). ANOVA with post-hoc Tukey HSD Test Calculator with Scheffé, Bonferroni and Holm multiple comparison - input k, the number of treatments. Astatsa.com. https://astatsa.com/OneWay_Anova_with_TukeyHSD

Vietnam E-commerce Association (VECOM). *Vietnam E-Commerce Business Index Report 2024*. (2024). Viet Nam E-Commerce Association. http://en.vecom.vn/vietnam-e-commerce-business-index-report-ebi-2024

Vietnam eCommerce and Digital Economy Agency. (2024). Idea.gov.vn. https://en.idea.gov.vn/default.aspx?page=news&do=detail&category_id=4f89e92d-e344-49f6-9abc-0ffa134f88a1&id=80b29be9-3b23-4601-959d-d11d01327e26

Zouina, M., & Outtaj, B. (2017). A novel lightweight URL phishing detection system using SVM and similarity index. Human-Centric Computing and Information Sciences, 7(1). https://doi.org/10.1186/s13673-017-0098-1

Appendix A APWG 2024 reports

APWG			
	# Reported phishing	Percentage E-commerce	Estimated # (rounded)
Q1	963,664.00	5%	52,038
Q2	877,536.00	6%	51,775
Q3	932,923.00	8%	74,634
Q4	989,123.00	10%	98,912
Total	3,763,246.00	7%	277,359

Appendix B Equations

ANOVA

For this study we have three different years to compare the differences of the specific differences in phishing websites. The experimental unit is the object of measurement. Y is the dependent variable of interest in the study. $\bar{Y}$ is the overall mean. K is the number of groups. N is the sample size. $\bar{Y}_i$ is the average of the group i. n_i is the number of observations in the group i.

The equations for one-way ANOVA are as follows (Edlund & Nichols, 2023; Kim, 2017; NIST/SEMATECH, 2019):

The sum of Squares (Treatment) $$SST = \sum_{i=1}^{k} n_i \left(\bar{Y}_i - \bar{Y}\right)^2 \quad (1)$$

The sum of Squares (Error) $$SSE = \sum_{ij=1}^{n} \left(Y_{ij} - \bar{Y}_i\right)^2 \quad (2)$$

Degree of freedom (Treatment) $$(K-1) \quad (3)$$

Degree of freedom (Error) $$(N-K) \quad (4)$$

Mean sum of Squares (Treatment) $$MSS = \sum_{i=1}^{K} n_i \left(Y_i - \bar{Y}\right)^2 / (K-1) \quad (5)$$

Mean sum of Squares (Error) $$MSE = \sum_{Yij}^{n} \left(Y_{ij} - \bar{Y}_i\right)^2 / (N-K) \quad (6)$$

F Statistic $$F = \frac{Treatment\ variance}{Error\ Variance} = \frac{\sum_{i=1}^{K} n_i (\bar{Y}_i - \bar{Y})^2 / (K-1)}{\sum_{ij=1}^{n} (\bar{Y}_i - \bar{Y})^2 / (N-K)} \quad (7)$$

Appendix C Top-Level Domains (TLD)

Type of TLD	2022	2023	2024
ac	0	0	2
am	1	0	0
asia	1	1	0
bet	1	0	0
cc	2	4	6
cloud	0	1	0
club	3	0	1
co	1	0	0
com	100	139	204
html	1	0	1
me	1	0	0
net	11	2	3
online	4	0	3
org	2	1	2
shop	3	6	15
store	2	0	0
top	3	3	8
vip	7	4	19
vn	6	4	6
xyz	1	0	2
life	0	1	0
site	0	3	1
ws	0	1	0
gg	0	0	1
how	0	0	3
in	0	0	1
info	0	0	2
link	0	0	6
live	0	0	1
Total	150	170	287

Appendix D Brand Impersonation

Brand Impersonation	2022	2023	2024
Aeon	0	5	7
Alibaba	0	2	0
Amazon	2	13	29
Carrier Four	1	0	0
BigC	2	0	0
Dien May Nguyen Kim	2	0	0
Dien May Xanh	29	10	21
Ebay	0	5	6
Lazada	21	21	39
Lotte	3	0	0
Media Mart	0	0	1
Mobile World	2	0	0
Sendo Ecommerce	16	3	14
Shopee	48	47	81
Taobao Commerce	0	0	3
Tiki	24	64	86
Total	150	170	287

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- VNPhishingURLDatasetCopy.xlsx