# Introduction

In today's digital age, the internet has become a crucial communication tool, enabling people to create an online environment that complements their offline business management or even set up fully functional online businesses. While the internet offers many benefits, it also has a downside that users must be aware of. One of the significant risks associated with the online environment is online fraud through phishing attacks. Phishing involves an attacker luring users to visit fake or malicious websites to obtain their confidential information, such as passwords, credit card numbers, usernames, and so on [1]. Unfortunately, phishing attacks have become rampant and comprise over half of all internet fraud cases affecting ordinary users [2]. In response to this, Google issues up to ten million warnings every day to users who visit known websites for their phishing scams. The list of dangerous sites is growing by about 10,000 daily, emphasizing the need for online users to be vigilant and cautious while browsing.

Phishing offences have been increasing and have resulted in billions of dollars in losses. Phishers use various attack vectors, such as search engines, fake websites, advertisements, email, instant messaging, or phone calls [3]. However, the most widely used technique by phishers is URL obfuscation [4]. This technique lures users by misleading them into visiting fake websites through URLs or genuine websites that are familiar to the victim [5]. In these attacks, users enter their critical information, such as credit card details and passwords, into a forged website that appears legitimate. Software-as-a-Service (SaaS) and webmail sites are the most common targets of phishing attacks [6]. Phishers create websites that look very similar to benign websites, and then they send the phishing website link to millions of internet users via emails and other communication media. These types of cyber-attacks are usually activated by emails, instant messages, or phone calls [7]. The aim of the phishing attack is not only to steal the victims' identity, but it can also be used to spread other types of malwares, such as ransomware, exploit approach weaknesses, or to receive monetary profits [8].

During the third quarter of 2022 [6], there was a noticeable increase in phishing attacks compared to earlier in the year. According to the Anti-Phishing Working Group (APWG) report, a total of 1,270,883 unique phishing sites were detected between July and September [6]. Furthermore, wire transfer Business E-mail Compromise (BEC) attacks demanded an average of $48,000, down from $80,000 in the previous quarter and $54,000 in the first quarter. In the fourth quarter of 2022, it was discovered that phishing attacks against financial institutions were the most prevalent. Attacks against Software-as-a-Service (SaaS) and webmail sites decreased, while attacks against E-commerce sites increased. However, attacks against media companies declined slightly from 12.6% to 11.8% [6]. The prevailing pandemic situation has also seen a significant increase in phishing attacks. Cybercriminals are exploiting the global focus on Covid-19, and according to the World Health Organization (WHO), many hackers and cyber scammers are sending fraudulent emails and WhatsApp messages to people [9]. These attacks come in various forms, including fake job offers, fabricated messages from health organizations, Covid-19 vaccine-themed phishing, and brand impersonation.

Phishing is a type of cyber-attack that is difficult to detect and prevent because phishers are constantly coming up with new methods to evade existing anti-phishing techniques. Even well-educated and experienced users can fall victim to these attacks. In order to combat phishing, software-based detection techniques are preferred. Some common methods for detecting phishing include blacklists/whitelists [10], natural language processing [11], visual similarity [12], rules [13], and machine learning [14], [15]. However, blacklists/whitelists can fail to detect new or previously unlisted phishing sites, while machine learning techniques rely on heuristic features such as URL, webpage content, website traffic, search engine, WHOIS record, and Page Rank. While these heuristic features can improve detection efficiency, they are not guaranteed to be present in all phishing websites and can sometimes be present in benign websites, leading to classification errors.

Some of the heuristic features used in phishing detection are difficult to access and dependent on third-party services. These services, such as page rank, search engine indexing, and WHOIS, may not be enough to accurately identify phishing websites that are hosted on compromised servers. Such websites may be falsely identified as benign because they appear in search results, and may even be misclassified as phishing sites due to their lack of domain age. Visual similarity-based heuristic techniques compare new websites to pre-stored signatures that include screenshots, font styles, images, page layouts, logos, and other visual elements. However, these techniques have a high false-negative rate for new phishing websites. URL-based techniques [16]–[18], which extract features such as the number of dots in a URL, the presence of special symbols, URL length, and brand names, also have limitations. These features are often manually extracted, which requires time and extra maintenance costs. As a result, a hybrid approach that combines multiple methods is recommended for phishing detection, rather than relying on a single technique.

Phishing attacks pose a significant threat as they can be hard to identify. Cybercriminals use social engineering tactics to create convincing phishing emails or messages that deceive victims. To make matters worse, the fake websites used in these attacks are often designed to look like legitimate sites, making it challenging for victims to differentiate between them. Therefore, there is a pressing need for effective phishing detection methods that can help individuals and organizations safeguard against such attacks. One approach to detecting phishing is URL-based detection techniques, which involve examining website URLs to determine their authenticity.

URL-based phishing detection techniques are designed to scrutinize website URLs to differentiate between legitimate and fake sites. These techniques are particularly useful in identifying phishing websites that are created to resemble legitimate sites. The most widely used method for URL-based phishing detection is the analysis of the domain name. Phishing websites often use domain names that are similar to legitimate websites but with minor changes. For instance, a phishing website may replace one letter in the domain name with another similar-looking letter. The main objective of this study is to provide an extensive overview of phishing attacks, highlight the importance of detecting such attacks, and explore different URL-based detection techniques that can be utilized to identify phishing websites.

## AIMS AND OBJECTIVES OF THE PROJECT

Phishing refers to a fraudulent technique used to trick individuals into revealing their sensitive information, such as login credentials, credit card details, and other personal information, to fake websites that appear authentic. The attackers then use this information to gain unauthorized access to important accounts, leading to identity theft and financial harm. Among the various types of cyber-attacks, phishing is the most frequently used method. Therefore, it is crucial for computer and internet users to safeguard their personal information, ensure its security, and minimize the risk of falling victim to fraudulent activities while browsing different websites.

Phishing has been recognized as a persistent issue since the early days of the internet, and it is one of the most challenging issues to address and control. The goal of this research is to examine the dangers of phishing and to create a detection system to mitigate the issue. In related studies, deep learning (a branch of machine learning) and natural language processing have been utilized and have exhibited significant progress in detecting phishing. Despite this, several cybersecurity experts have advocated for hybrid or mixed methods rather than a single strategy. To this end, this research proposes an automated phishing responder (APR) that employs a mixed machine learning approach that combines natural language processing to identify phishing attempts. Furthermore, the effectiveness of the proposed approach will be evaluated by comparing it to other techniques.

Thus, the research objective of this project encompasses, but is not restricted to, the following:

* To develop a novel approach to detect malicious URLs and alert users.
* To apply ML techniques in the proposed approach to analyze the real-time URLs and produce effective results.
* To implement the concept of different ML technique that has the capability to handle huge amount of data.
* Developing a phishing detection system
* Studying previous work on the proposed topic and looking for ways to improve them.

# References

[1] A. Solanki and S. R. Dogiwal, “Implementation of an anti-phishing technique for secure login using USB (IATSLU),” *Smart Innovation, Systems and Technologies*, vol. 31, pp. 221–231, 2015, doi: 10.1007/978-81-322-2205-7\_21/COVER.

[2] H. Orman, “The compleat story of phish,” *IEEE Internet Comput*, vol. 17, no. 1, pp. 87–91, 2013, doi: 10.1109/MIC.2013.16.

[3] G. Liu, B. Qiu, and L. Wenyin, “Automatic detection of phishing target from phishing webpage,” *Proceedings - International Conference on Pattern Recognition*, pp. 4153–4156, 2010, doi: 10.1109/ICPR.2010.1010.

[4] R. S. Rao and S. T. Ali, “PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach,” *Procedia Comput Sci*, vol. 54, pp. 147–156, Jan. 2015, doi: 10.1016/J.PROCS.2015.06.017.

[5] M. Cova, C. Kruegel, and G. Vigna, “There is No Free Phish: An Analysis of ‘Free’ and Live Phishing Kits”.

[6] “APWG | Phishing Activity Trends Reports.” https://apwg.org/trendsreports/ (accessed Feb. 22, 2023).

[7] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J. P. Niyigena, “An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL,” *Electronics 2020, Vol. 9, Page 1514*, vol. 9, no. 9, p. 1514, Sep. 2020, doi: 10.3390/ELECTRONICS9091514.

[8] R. Dhamija, J. D. Tygar, and M. Hearst, “Why Phishing Works”, Accessed: Feb. 22, 2023. [Online]. Available: www.paypa1.com

[9] “Cyber security .” https://www.who.int/about/cyber-security (accessed Feb. 22, 2023).

[10] A. K. Jain and B. B. Gupta, “A novel approach to protect against phishing attacks at client side using auto-updated white-list,” *EURASIP J Inf Secur*, vol. 2016, no. 1, pp. 1–11, Dec. 2016, doi: 10.1186/S13635-016-0034-3/FIGURES/5.

[11] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, “Machine learning based phishing detection from URLs,” *Expert Syst Appl*, vol. 117, pp. 345–357, Mar. 2019, doi: 10.1016/J.ESWA.2018.09.029.

[12] S. Haruta, H. Asahina, and I. Sasase, “Visual Similarity-Based Phishing Detection Scheme Using Image and CSS with Target Website Finder,” *2017 IEEE Global Communications Conference, GLOBECOM 2017 - Proceedings*, vol. 2018-January, pp. 1–6, Jul. 2017, doi: 10.1109/GLOCOM.2017.8254506.

[13] D. L. Cook, V. K. Gurbani, and M. Daniluk, “Phishwish: A stateless phishing filter using minimal rules,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5143 LNCS, pp. 182–186, 2008, doi: 10.1007/978-3-540-85230-8\_15/COVER.

[14] A. K. Jain and B. B. Gupta, “A machine learning based approach for phishing detection using hyperlinks information,” *J Ambient Intell Humaniz Comput*, vol. 10, no. 5, pp. 2015–2028, May 2019, doi: 10.1007/S12652-018-0798-Z/TABLES/6.

[15] Y. Li, Z. Yang, X. Chen, H. Yuan, and W. Liu, “A stacking model using URL and HTML features for phishing webpage detection,” *Future Generation Computer Systems*, vol. 94, pp. 27–39, May 2019, doi: 10.1016/J.FUTURE.2018.11.004.

[16] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, “CANTINA+,” *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 2, Sep. 2011, doi: 10.1145/2019599.2019606.

[17] W. Zhang, Q. Jiang, L. Chen, and C. Li, “Two-stage ELM for phishing Web pages detection using hybrid features,” *World Wide Web*, vol. 20, no. 4, pp. 797–813, Jul. 2017, doi: 10.1007/S11280-016-0418-9/TABLES/4.

[18] R. S. Rao, T. Vaishnavi, and A. R. Pais, “CatchPhish: detection of phishing websites by inspecting URLs,” *J Ambient Intell Humaniz Comput*, vol. 11, no. 2, pp. 813–825, Feb. 2020, doi: 10.1007/S12652-019-01311-4/FIGURES/4.