Implementation of Machine Learning to Automate the Phishing Websites Detection

Angeline Tandri[1], Intan Farahana Binti Kamsin[2], Zety Marlia Zainal Abidin[3], Hemalata Vasudavan[4]

[1'2'3'4]Asia Pacific University of Technology and Innovation, Technology Park Malaysia, Bukit Jalil, KualaLumpur, Malaysia.

[1]tp058954@mail.apu.edu.my, [2]intan.farahana@staffemail.apu.edu.my, [3]zety@staffemail.apu.edu.my, [4]hemalata@staffemail.apu.edu.my

***Abstract*— The number of gadgets linked to the internet has increased dramatically in recent years. In comparison to other types of cyber-attacks, phishing has become the most popular in cyberspace because it leverages human flaws rather than technology vulnerabilities. In phishing assault, an internet user is duped into providing personal information, such as login credentials or credit card information, by an apparently trustworthy organization. Many researchers have recently proposed alternative ways to phishing assaults. However, they are still reliant on user engagement to progress. Therefore, this research aims at studying on how Machine learning may be used to automate the identification of phishing websites. This research will use the stratified sampling method which is applied for 150 users raging from normal users, organizations' IT team, as well as the browser team by distributing the Online Survey. In addition, three random participants are selected to attend the interview session to validate the reliability of the online survey result. At the end of this research, an automatic detection tool is developed to achieve the aim. Furthermore, this research motivates the future work in terms of adding more functional features and also the available platforms.**

***Index Terms*— Automation, Machine Learning, Phished Website Phishing**

1. Introduction

In today's borderless world, we are all able to connect and get in touch with everyone around the world seamlessly. This circumstance can only be achieved due to the presence of technology such as computer, smartphone, internet, and social media. The technology advancements as well as the raising number of users have made the cyber-attack threat rate even higher. In addition, these cyber-attacks can only be successful due to the vulnerabilities of one or even both of the following factors, namely human as the users, and the implemented security strategies.

The most extremely common attacks which has occurred and kept existing in these last few years is Phishing which uses the social engineering technique by impersonating a legitimate person and sending emails, messages, or even calls to persuade the victim. The phisher will ask the victim to hand over sensitive data such as password, financial data. Besides that, the victim will also be asked to perform certain tasks such as clicking the given web link or downloading the attached file that contains the malware [1].

It is reported that the attackers are mostly put their target on businesses and companies either small or big and according to CISCO's 2021 Cybersecurity Threat Trends Report, in around 86% of the company, there will be at least one person who click the phishing link where 90% of the data breach cases was caused by this phishing attack [2]. We are able to see that even the educated people can still fall into this attack and lead the business or the company to the great loss. Not only about the data loss, but also the financial aspect as well as the trust and reputation from its stakeholders.

Consequently, to avoid this attack, most of the business' owner has taken several security strategies starting from implementing the security software, backing up the data, until training employees to raise their security awareness for instance, not to open the unknown or suspicious emails and click the attached link. But, even though those security measures have been implemented, there are still many cases of this attack that keep happening every year, no matter it is because of human error or the security system failure such as the filter. Hence, there have been many experts who are conducting research on how to overcome this attack where one of the ways is by using the machine learning approach.

2. Literature Review

2.1. Research Domain

2.1.1. Machine Learning and Automation

Machine Learning is a machine which is developed to be able to learn and improve by itself via learning and experience without being explicitly programmed [3]. The primary goal of machine learning is to use previous data to construct a model that can be improved, recognize complex patterns, and train itself to find solutions to new problems. [4]. Because machine learning provides flexibility, the data required by the computer may be obtained from a variety of sources dependent on its role. As a result, it may be used in any sector.

According to [5], Machine Learning in cyber-security is critical for assisting in the fight against security events.

"Machine learning may be used to automate some sorts of misuse-based detection by letting a system to "learn" what different types of assaults look like," according to the article. Furthermore, the supervised learning classifier may be trained to recognize the tell-tale indications of various sorts of assaults without the need for human intervention, particularly when compiling the list of criteria to trigger the alarm [5]. Therefore, in this research, machine learning will be implemented to automate the phished websites detection and prevention.

2.1.2. Phished Website and Its Solution

Phished Website is a pre-built false website that appears to be authentic and has been discovered to be the most common medium used by phishers to start phishing attacks in order to obtain the victims' personal data such as their profile, login, password, and even financial information. By only obtaining a tiny portion of the victim's information, the phisher has already created a tailored and credible email to be utilized for their own profit.

Since, phishing is a social engineering assault, it does not rely on a powerful infection system such as virus or malware, but rather, it is fully an easy approach because it merely depends on the users' willingness to provide their data. Besides that, this cyber-attack cannot be readily mitigated since its methods and approaches are always changing. As a result, it has been a huge security risk for a long time, with no viable remedy in place.

With these concerns in mind, the most often used anti-web phishing detection technology integrates a blacklisting or whitelisting solution to the suspected phished websites [6]. The blacklisting and whitelisting methods are utilized as detection tools because they are simple to develop and may yield a low false-positive rate while operating at high speed. According to a statistic, the blacklisting strategy may place 47% - 83% of phishing websites on the blacklist after 12 hours. However, the lifespan of a phished website is usually only two hours. As a result, the blacklist list is no longer reliable [6].

Machine learning is another prominent anti-web phishing tool solution that is now in use. This program can detect phishing websites based on their URL, IP address, or other factors. According to [7], the accuracy of machine learning is 99.57 percent, with a false positive rate of roughly 0.53 percent. Not only that, but this strategy is also seen to be an excellent way to tackle phishing websites since it can grow with them [8].

In a word, it can be stated that using the machine learning technology is the most effective way to stop the spread of phished websites by speeding up the detection and prevention of the phished website existence on the internet which is the same as what has been proposed in this research.

2.1.3. Machine Learning in Identifying The Phished Website and its Algorithms

Machine Learning is a machine that has been designed to do a certain job without the need for a programmed answer, instead determining the solution on its own by training with the task's example data. Machine Learning approaches have been shown to be an effective tool for discovering patterns in data since it could detect as well as recognize the typical features of phished websites [9].

Even if a phishing website is created uniquely and has distinct features, the majority of them will exhibit some similarities and patterns [9]. The website aspects being evaluated might range from significant to minor, such as URL consistency, brand name, and even hidden or limited content. Furthermore, Machine Learning is categorized intro four groups which are reinforced, unsupervised, supervised, and semi-unsupervised. The most fundamental learning approach is supervised learning, in which the machine is given the sample input-output and it will learn those given sample and start to build a model on how to convert an input to achieve the desired output. This category is then broken into two parts: Regression and Classification. Classification is the distribution of data into groups specified on the data set based on their individual properties, whereas regression is the prediction or conclusion of additional data features based on certain accessible attributes [4].

Unsupervised learning, on the other hand, will work without requiring human intervention to examine the labeled information. This is commonly used for detecting generative characteristics, discovering relevant patterns, and extracting generating features.

The next one is the Semi-supervised machine learning which is a blend of supervised and unsupervised approach. Since, it stands between learning "with supervision" and "without supervision". Semi-supervised learning aims at delivering a better predictive result rather than just the labeled data model. This approach is frequently used for labeling, fraud detection, machine translation, text categorization, and other applications [3].

Finally, the reinforcement machine learning type would allow the machines and software agents to evaluate the ideal behavior of certain case environment in order to enhance the performance. This strategy will make use of environmentalists' ideas to take action that minimizes risk or increases rewards. Besides that, it is also a useful method for developing AI models, increasing automation, and sophisticated systems' efficiency like robots and driverless activities [3].

According to the overview above, machine learning offers a range of approaches with different functionality and purposes. As a result, machine learning will be employed in this study to detect and prevent phishing websites.

2.2. Similar System

2.2.1. SpoofGuard

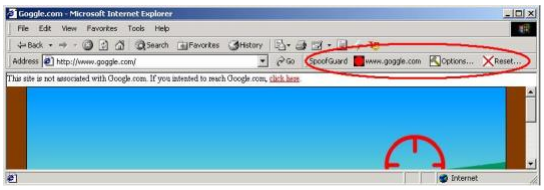

Figure 1: *SpoofGuard*

SpoofGuard is an anti-phishing toolbar that uses a variety of heuristic techniques to categorize phishing websites and calculate the weighted total score for each website from each heuristic collection. This score is characterized by a high false positive rate. However, because SpoofGuard does not employ a blacklist, it must wait for the web page to fully load before authenticating and categorizing whether or not it is valid. [10].

2.2.2. NetCraft Toolbar

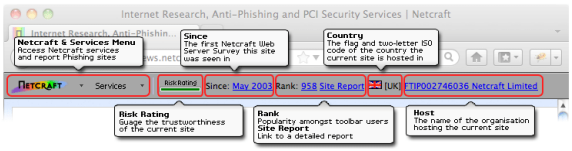

Figure 2: *NetCraft Toolbar*

Netcraft is an anti-phishing browser addon that determines the validity of a website using heuristic algorithms, sniffer techniques, and banned URLs maintained in Netcraft's database. Netcraft works by scanning the URL for specific characters and blocking pop-ups that are often used to mask the navigation control. On the other hand, The Netcraft toolbar plugin is simply available for Microsoft Internet Explorer and Firefox. Furthermore, after recognizing the counterfeit sites, it will just issue a warning, which users may disregard and end up providing their personal information even though they are using the tool. Aside from that, Netcraft toolbar is highly dependent on connection speed, thus if the connection is slow, the alert may be delayed and may be too late to notify the users. [10].

2.2.3. BrandShield

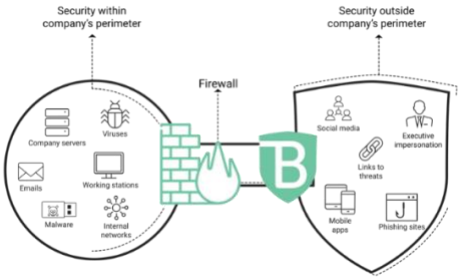

Figure 3: *BrandShield*

BrandShield is an anti-phishing dashboard product that is driven by a combination of AI and machine learning [11]. BrandShield will provide a detailed digital hazard map by tracking the Internet, along with the social media, phishing websites, identity theft, as well as online scams [12]. It also has a Website Duplication Detector, automatic takedown notifications, blacklists, and many more features. All of this is done by the powerful technology created by BrandShield to track and safeguard the company's trademarks and visual assets. [13].

2.2.4. Comparison Table

Similar Systems / Features	SpoofGuard	NetCraft Toolbar	BrandShield
Machine Learning	✘	✘	✓
Identify Spoofed Site	✓	✓	✓
Automatic Spoofed Sites Takedown	✘	✘	✓
Type	ToolBar	Browser Extension	-
Availability	Microsoft Internet Explorer	Microsoft Internet Explorer and Firefox	Social Media

Table 1: *Similar system comparison table*

The comparison table above displays the systems that have the attributes that are comparable to the suggested system in this study. Machine Learning, spoofed sites identification capability, automatic takedown, an embedded system, and available for all types of internet browsers are all the aspects that are taken into account.

2.2.5. Conclusion

In a nutshell, not all of the comparable systems described above have the capabilities and features shown in the table above. As a result, this research is proposing an automatic phished website detection solution through the help of Machine Learning, which is able to be embedded in all types of browsers. Thus, by investigating another method to solve the phished websites with the aid of Machine Learning, it is expected to solve the phishing website issues as well as maximizing the anti-phishing tool role and would lead to the greater benefits for users, companies, and world.

3. Problem Statement

In this digitalization era, most of all our activities have shifted into the online base which makes our life easier and comfortable. The level of user convenience can be achieved through a high level of security. But, unfortunately, the security issues keep popping up such as phishing, malware, and identity theft, [14].

Among all the existing security incidents, 93% of them were caused by the phishing attacks which also identified as the root

cause of another cyber incidents [15]. Those attacks are launched by tricking the users to visit the phishing webpage by sending the emails, real-time messages, and so on. [16]. According to the most recent Anti-Phishing Working Group Report, in the fourth quarter of 2021, it found an increment in the total number of unique web by 21.67% while compared to the previous quarter in the same year [17].

Even though various anti-phishing tools have been widely available, most of them still cannot run automatically. Thus, it requires the users' involvement actively [18]. Consequently, the continuous massive attacks are inevitable and made the victims to keep suffering from various forms of losses [19]. One of them is about the monetary which was reported to be the second largest cause of financial loss in 2021 with $4.65 million [20].

Therefore, investigating a method to overcome the phishing webpages automatically with the aid of Machine Learning is crucial. Not only, in reducing the victims' number and side effects, but it also benefits the advancement of the cyber security field and the future world.

4. Research Aim and Objectives

4.1. Aim

This research aims at proposing the Machine Learning approach for detecting and preventing the spread of phishing webpages on the internet by incorporating Machine Learning into a platform to identify the phishing webpage patterns while focusing on the below objectives:

4.2. Objectives

i) To identify the phishing webpages patterns.
ii) To collect the phishing webpages' pattern information in the database
iii) To enable automatic phishing webpage recognition by the system.
iv) To implement the automated phishing webpage detection in the browser.

5. Research Significance

The findings of this research would make some contributions either to normal internet users or companies. They can start to surf the browser freely and open any webpages safely. By doing so, they are all can be protected from the chances of being suffered from phishing attacks. Meanwhile, for the researchers, this research will provide them some insights about the significance role of a machine learning in the cyber security field to detect or prevent the attacks as well as for automating tools. Thus, there will be more and more automated tools to combat the cyber-attacks in the future.

6. Research Methodology

This research will be using the mixture of both qualitative and quantitative methods to receive the valid as well as reliable results' support for the set aim and objectives. For the quantitative method, the stratified sampling method will be used which is applied for 150 target users who are raging from organizations' IT team, browsers' IT team, and normal users. This sampling will be conducted through an online survey. On the other hand, the qualitative method will use the interview technique with 3 persons who will be randomly selected from each category to verify the accuracy and reliability of the online survey.

6.1. Sampling Method

Stratified sampling method is a complex probability sampling which falls under probability sampling technique. This stratified sampling will divide the population who shares the similar characteristics into a different subgroup [21]. After that, from each subgroup, equal size of the respondents will be selected as their group representative. As a result, it can help in improving the accuracy and prevent the sampling bias [22]. In addition, the respondents who will be involved for this particular research will come from multiple backgrounds, namely, normal users, organizations' IT team, and browser's owners.

6.2. Survey

Survey will be used for the first data collection method since it doesn't require a lot of cost and time, but can reach a large number of respondents [23]. This survey will be conducted online through a Google form which is contributed through email and social media. The survey questions will be the mixture of closed and open questions where the open questions aims at gaining the respondents' explanation, opinion, and facts, while close questions will be in the Likert scale-based. However, those prepared questions will go through the pilot test to assess its simplicity, reliability, and suitability. So that, the survey will produce good quality results.

6.3. Interview

Interview is one of the effective data collection methods to receive the live answer from the participants through face-to-face conversation. Currently, there are three available category of the interview which are structured, semi-structured, and unstructured. In addition, there are two ways on how to conduct the interview session which are through one to one or one to many.

For this research, three random participants from each category will be invited to attend the one to one structured interview. This method is chosen than others because one to one session will give the participants privacy as well as allow them to answer all the questions from their own perspective without hesitation and minimize the probability of one participant to

copy others participant response or wait for other participants to answer a particular question. Besides that, the structured interview requires less time than other method since questions have been prepared in advanced and will also minimize the mistakes that may happen during the interview. Aside from that, the results from structured interview are easier to be compared and analyzed because they are asked the same questions.

7. Proposed System Overview

Figure 4 below shows the flowchart of the proposed system overview when the user wants to upload a website on the internet. Firstly, when the user uploads a website, this tool will automatically be triggered to scan through all the website element such as URL, the contents, IP domain, logo, etc. If the study finds one or more elements are suspected then, it will automatically takedown the website and send the notification to the browser along with the website details and suspected elements. The study then will be stored to the database as a reference later on. However, if there is no suspected elements are found then, the website will be released on the internet.

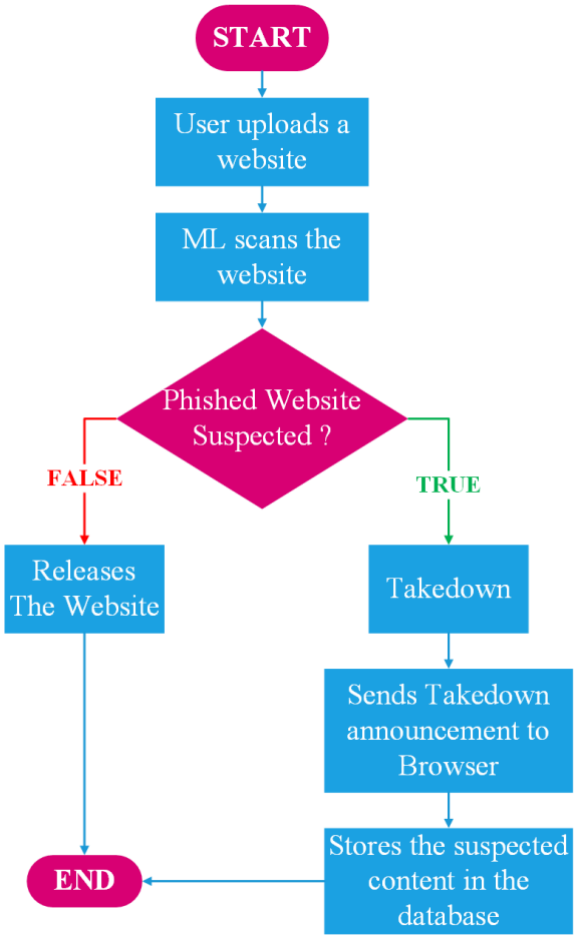

Figure 4: *Flowchart of Automated Phishing Webpage Detection and Prevention*

On the other side, this proposed system will also work to scan through all the current existed websites in the browser. The flow of this function is similar to the previous function which is shown in Figure 4 above. However, for this function, the system will directly scan the current existed website one by one and decide whether it should be takedown or let it be.

8. Conclusion

In a nutshell, phishing has been the most popular and successful cyber-attacks for years as it can be the root cause of other security incidents and cause severe damage towards the victims. However, solutions to combat this issue are still limited and not capable of following the trends as the phishers techniques keep evolving from time to time. Therefore, in this research, we proposed an anti-phishing tool to combat the phished websites by using the machine learning approach to detect and prevent them from circulating on the internet automatically as well as predict the future phished website's appearance. Aside from that, it will be embedded and available for all browsers. Thus, the number of spoofed websites which circulates in browsers will be reduced and slowly diminished. As a result, the cyber incidents related to phishing can be suppressed and users can surf online safely. This research motivates future work in adding more functionalities and available platforms.

References

[1] S. Cook, "Phishing statistics and facts for 2019–2022," 27 January 2022. [Online]. Available: https://www.comparitech.com/blog/vpn-privacy/phishing-statistics-facts/.

[2] J. Wise, "HOW MANY PHISHING EMAILS ARE SENT DAILY IN 2022? 11+ STATISTICS," 11 March 2022. [Online]. Available: https://earthweb.com/how-many-phishing-emails-are-sent-daily/#:~:text=Phishing%20Email%20Statistics-,1.,that's%20quite%20massive%20and%20scary!.

[3] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," 2021.

[4] Ö. ÇELİK and S. S. ALTUNAYDIN, "A Research on Machine Learning Methods and Its Applications," 2018.

[5] M. Musser and A. Garriott, "Machine Learning and Cybersecurity," 2021.

[6] P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," 2018.

[7] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione and X. Chang, *A novel approach for*

phishing URLs detection using lexical based machine learning in a real-time environment, 2021.

[8] I. Vayansky and S. Kumar, "Phishing – challenges and solutions," 2018.

[9] V. Shahrivari, M. M. Darabi and M. Izadi, "Phishing Detection Using Machine Learning Techniques," 2020.

[10] G. Nduati, J. B. Rubaiza, S. Mabarani and L. Svotwa, "Enhancing Phishing Detection Tools: A Machine Learning Approach," 2020.

[11] BrandShield, "Brand-Oriented Digital Risk Protection," [Online]. Available: https://www.brandshield.com/brand-oriented-digital-risk-protection/.

[12] BrandShield, "Anti-Phishing Solutions," [Online]. Available: https://www.brandshield.com/products/anti-phishing/.

[13] Softprom, "ANTI-PHISHING - PROTECTING YOUR BRAND ONLINE," [Online]. Available: https://softprom.com/vendor/brandshield/product/anti-phishing.

[14] I. Ghafir, M. Hammoudeh, V. Prenosil, L. Han, R. Hegarty, K. Rabie and F. Aparicio-Navarro, "Detection of advanced persistent threat using machine-learning correlation analysis," *Future Generation Computer Systems*, 2018.

[15] APWG, "Phishing activity trends report," Anti-Phishing Working Group, United States, 2018.

[16] J. Feng, L. Zou, O. Ye and J. Han, "Web2Vec: Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning," 2020.

[17] APWG, "Phishing Activity Trends Report 4th Quarter 2021," Anti-Phishing Working Group, United States, 2022.

[18] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," 2020.

[19] M. Cooper, Y. Levy, L. Wang and L. Dringus, "Heads-up! An Alert and Warning System for Phishing Emails," 2021.

[20] IBM, "Cost of a Data Breach Report 2021," IBM, United States of America, 2021.

[21] A. E. Berndt, "Sampling Methods," 2020.

[22] P. Lynn, "The Advantage and Disadvantage of Implicitly Stratified Sampling," 2019.

[23] D. A. Story and A. R. Tait, "Survey Research," 2019.

[24] S. M. Thaler, "Automation for Information Security using Machine Learning," 2019.

[25] Y. D. Bari, "The Future of Tomorrow: Automation for Cybersecurity," 2019.

[26] W. Jin, "Research on Machine Learning and Its Algorithms and Development," 2020.

[27] Y. Li, Z. Yang, X. Chen, H. Yuan and W. Liu, "A stacking model using URL and HTML features for phishing webpage detection," 2019.