International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 02 | Feb - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Real-Time Phishing Detection Using Machine Learning-Based URL Analysis

Dr.S.Gnanapriya, Ummu Habeeba k p
Assistant professor, Department of Computer Applications, Nehru College of Management, Coimbatore, Tamilnadu, India
gnanapriya_2006@yahoo.co.in
Student , II MCA, Department of Computer Applications, Nehru College of Management
Coimbatore, Tamilnadu, India
habeebahakim1334@gmail.com

Abstract

In latest years, improvements in Internet and cloud technology have brought about a huge boom in digital buying and selling wherein customers make on-line purchases and transactions. This increase results in unauthorized get admission to customers touchy records and damages the assets of an enterprise. Phishing is one of the acquainted assaults that trick customers to get admission to malicious content material and advantage their records. In phrases of internet site interface and uniform aid locator (URL), maximum phishing webpages appearance same to the real webpages. Various techniques for detecting phishing websites, consisting of blacklist, heuristic, Etc., had been suggested. However, because of inefficient safety technology, there may be an exponential boom withinside the wide variety of victims. The nameless and uncontrollable framework of the Internet is extra susceptible to phishing assaults. Existing studies works display that the overall performance of the phishing detection device is limited. There is a call for an sensible approach to defend customers from the cyber-assaults. A recurrent neural community approach is hired to locate phishing URL. Researcher evaluated the proposed approach with 7900 malicious and 5800 valid sites, respectively. The experiments final results suggests that the proposed approach`s overall performance is higher than the latest procedures in malicious URL detection. In latest years, with the growing use of cellular devices, there may be a developing fashion to transport nearly all real-international operations to the cyber international.

Index: Phishing, Phishing Attack, Machine Learning, Network Attack.

1. INTRODUCTION

Phishing is a fraudulent approach that makes use of social and technological hints to thieve consumer identity and monetary credentials. Social media structures use spoofed e-mails from valid agencies and companies to allow customers to apply faux web sites to reveal monetary information like usernames and passwords. Hackers set up malicious software program on computer systems to thieve credentials, regularly the usage of structures to intercept username and passwords of consumers` on line accounts. Phishers use a couple of methods, inclusive of email, Uniform Resource Locators (URL), immediately messages, discussion board postings, cell phone calls, and textual content messages to thieve consumer facts. The shape of phishing content material is much like the unique content material and trick customers to get admission to the content material so one can achieve their touchy data. The number one goal of phishing is to advantage positive non-public facts for monetary advantage or use of identification theft. Phishing assaults are inflicting excessive monetary harm across the world. Moreover, most phishing assault's goal monetary/charge establishments and webmail, in step with the Anti-Phishing Working Group (APWG) brand new Phishing sample studies.

In order to obtain personal data, criminals increase unauthorized replicas of a actual internet site and email, commonly from a monetary organization or different corporation managing monetary data. This electronic mail is rendered the usage of a valid corporation`s emblems and slogans. The layout and shape of HTML permit copying of snap shots or a whole internet site. Also, it's miles one

of the elements for the fast increase of Internet as a verbal exchange medium, and allows the misuse of brands, logos and different corporation identifiers that clients rely upon as authentication mechanisms. To entice customers, Phisher sends "spooled" mails to as many human beings as possible. When those e-mails are opened, the clients have a tendency to be diverted from the valid entity to a spoofed internet site.

Phishing is the maximum usually used social engineering and cyber-attack. Through such attacks, the phisher objectives naïve on-line customers via way of means of tricking the mint revealing exclusive facts, with the reason of the usage of it fraudulently. In order to keep away from getting phished, customers need to have cognizance of phishing web sites. Have a blacklist of phishing web sites which calls for the know-how of internet site being detected as phishing. Detect them of their early appearance, the usage of device getting to know and deep neural community algorithms of the above three, the device getting to know primarily based totally approach is tested to be best than the opposite methods. Even then, on-line customers are nonetheless being trapped into revealing touchy facts in phishing web sites. A phishing internet site isa not unusual place social engineering approach that mimics trustful uniform useful resource locators (URLs) and net pages. The goal of this venture is to educate device getting to know fashions and deep neural nets at the dataset created to are expecting phishing web sites. Both phishing and benign URLs of web sites are collected to shape a dataset and from them required URL and internet site content-primarily based totally capabilities are extracted. The overall performance degree of every version is measures and compared. The phishing internet site has advanced as a main cyber protection chance in current times. The phishing web sites host spam, malware, ransom ware, drive-via way of means of exploits, etc. A phishing internet site many a time look-alike a completely famous internet site and entice an unsuspecting consumer to fall sufferer to the trap. The sufferer of the scams incurs a economic loss, loss of personal facts and lack of reputation. Hence, it's miles vital to discover a answer that would mitigate such protection threats in a well-timed manner. Traditionally, the detection of phishing web sites is finished the usage of blacklists. There are many famous web sites which host a listing of blacklisted web sites, e.g. Phis Tank. The blacklisting method lack in aspects, blacklists won't be exhaustive and do now no longer stumble on a newly generated phishing internet site.

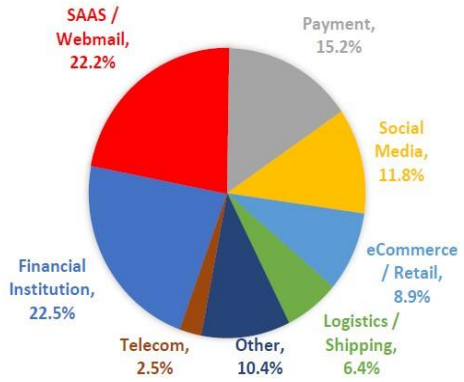

Fig. 1. Most targeted areas.

2. PROBLEM FORMULATION

Phishing assault is the handiest way to gain touchy statistics from harmless customers. The aim of the phishers is to gather crucial statistics like username, password and financial institution account details. Cyber protection individuals at the moment are seeking out sincere and consistent detection strategies for phishing website detection. This paper offers a device to get to know the era for detection of phishing URLs via a means of extracting and reading numerous functions of valid and phishing URLs. Decision Tree, random woodland and Support vector device algorithms are used to come across phishing websites. The aim of the undertaking is to come across phishing URLs in addition to slender right all the way down to a high-quality device, getting to know a set of rules via a way of means of evaluating fees, fake fines and fake poor fees of each set of rules. Nowadays, phishing has turned into a major problem for protection researchers due to the fact it is now no longer tough to create the faux internet site which seems so near a valid internet site. Experts can pick out faux websites. However, now, can not only pick out the faux internet sites and such customers emerge as the sufferers of phishing assault. The main intention of the attacker is to thrive the bank account credentials. In United States businesses, there may be a

lack of US$2billion step by year due to the fact that their customers emerge as sufferers of phishing. In the third Microsoft Computing Safer Index Report launched in February 2014, it became expected that the once-a-year global effect of phishing would be as excessive as $five billion. Phishing assaults have become a hit due to the loss of personal awareness. Since phishing assault exploits the weaknesses observed in customers, it's very tough to mitigate them. However, it's very vital to beautify phishing detection strategies.

categories	Predicted Phishing	Predicted Legitimate
Actual Phishing	TP	FN
Actual Legitimate	FP	TN

2.1 Key Metrics for Assessing Machine Learning Models

- $Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of PredictionsNumber}}$
- $Precision = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$
- $Recall = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$
- $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$
- $\text{Precision} = \frac{TP}{TP+FP}$
- $\text{Recall} = \frac{TP}{TP+FN}$

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

3. METHODOLOGY

In recent times machine learning techniques have been used in the classification and detection of phishing websites. In, this paper we have compared different machine learning techniques for the phishing website. In our daily life, we carry out most of our work on digital platforms. Using a computer and the internet in many areas facilitates our business and private life. It allows us to complete our transaction and operations quickly in areas such as trade, health, education, communication, banking, aviation, research, engineering, entertainment, and public services. The users who need to access a local network have been able to easily connect to the Internet anywhere and anytime with the development of mobile and wireless technologies. Although this situation provides great convenience, it has revealed serious deficits in terms of information security. Thus, the need for users in cyberspace to take measures against possible cyber-attacks has emerged.

The method of reaching target users in phishing attacks has continuously increased since the last decade. This method has been carried out in the 1990s as an algorithm-based, in the early2000s based on e-mail, then as Domain Spoofing and in recent years via HTTPs. Due to the size of the mass attacked in recent years, the cost and effect of the attacks on the users have been high. The average financial cost of the data breach as part of the phishing attacks in 2019 is $ 3.86million, and the approximate cost of the BEC (Business Email Compromise) phrases is estimated to be around $ 12 billion. Also, it is known that about 15% of people who are attacked are at least one more target. With this result, it can be said that phishing attacks will continue to being carried out in the ongoing years. Figure 1 also supports this idea and show the number of phishing sites in2019, and as can be seen from it, there is an increasing trend in this type of attack. In this regard, regular reports published by APWG (Anti Phishing Working Group) are an important guide for the researchers. According to the reports, the number of phishing sites is reached to approximately 640,000 sites were determined in 2018, and in the first three quarters of 2019, this number was reported as 629,611. Reports for the last quarter of 2019havenotbeen published yet. However, it can be said that the phishing attacks not only continue, but also there will be an increase in the number of attack types compared to the previous year.

This boom shows that phishing assaults are used greater via way of means of attackers. Because they may be smooth to design. Phishing assaults are primarily based totally at the attacker`s advent of a faux internet site, as depicted in Figure 2. First, a phisher makes faux websites,

which includes a phishing package. Then, the sufferer is directed to the faux internet site with the organized email. Believing that the email and URL are steady, the sufferer makes use of the faux internet site via way of means of clicking at the URL. After this moment, the Phishing package gets the sufferer`s credentials and sends it to the phisher. Finally, Phisher makes faux incomes from the valid internet site the use of the sufferer's credentials. These web sites normally have very comparable or maybe same visuals. In an e mail this is idea to be dispatched from a relied-on source, the goal is directed to this faux net site. The goal in this manner, the attacker receives data or earnings. Reliable e mail contents are created in distinct methods for the sufferer to believe. Previously, e-mails with low possibility offers, pressing texts, hyperlinks, or attachments that can be applicable and uncommon senders had been used. Today, dependable businesses or comparable hyperlinks to those businesses are preferred. Attackers opt for accomplishing to sufferers via way of means of the use of a steady verbal exchange protocol, and the actual URL is served via way of means of converting in a manner this is near the original. At this stage, if the sufferer is aware of the internet site is faux, he can defend himself from the attack.

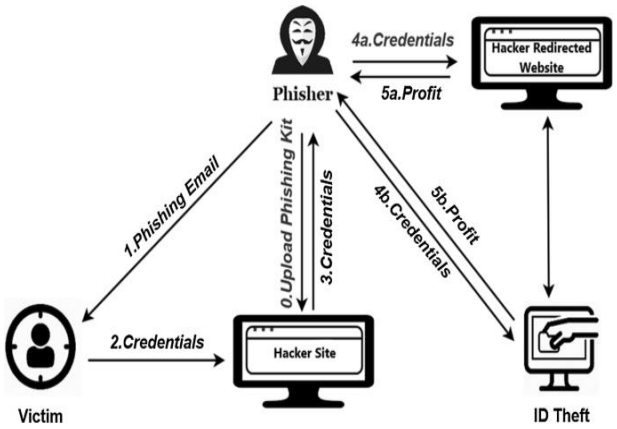

Fig2

3.1 Model Development and Evaluation

Several machine learning models were created and thoroughly assessed in order to meet the paper's aims. The selection of each model was based on its demonstrated performance in classification tasks, especially in the field of cybersecurity.

3.1.1 Machine Learning Models

- **Decision Tree Model:**

Decision Trees are essential to the sphere of device learning, recognized for his or her sincere and obvious method to type and regression tasks. These fashions function through growing a tree-like shape wherein every node represents a function of the dataset, and branches denote the choice policies of lead to exceptional outcomes. The simplicity of Decision Trees lies in their capacity to complicated choice-making strategies into a sequence of simpler, binary choices, making the model's choices clean to interpret and explain. This feature is specifically high-quality in phishing detection, because it lets in protection analysts to apprehend and hint the reasoning at the back of every type. Moreover, Decision Trees can control each numerical number and express data, making them flexible for diverse varieties of entry functions generally encountered in phishing datasets.

- **Random Forest Classifier:**

The Random Forest Classifier extends the idea of Decision Trees right into a greater effective ensemble technique that mixes more than one timber to enhance the predictive overall performance and decrease the hazard of overfitting. Each tree in a Random Forest works on a random subset of capabilities and information points, abilities to a numerous set of classifiers whose outcomes are aggregated to provide a very latest decision. This range makes Random Forests especially powerful in phishing detection, as they can seize a wide range of signs of malicious conduct without being overly touchy with noise and outliers with inside the information. The ensemble method additionally approaches that Random Forests are much less probable to be swayed via way of means of misleading strategies utilized by phishing attacks, supplying a strong protection towards quite a few phishing tactics.

- **Support Vector Machines (SVM):**

Support Vector Machines are powerful, supervised getting to know fashions used for type and regression tasks. SVMs are specifically referred to for his or her potential to create top-rated hyperplanes in a multidimensional area that relatively classifies the information points. This functionality is vital in phishing detection, wherein the difference between phishing and valid websites frequently lies in diffused and excessive-dimensional variations in features. SVMs are sturdy in opposition to overfitting, especially in excessive dimensional spaces, because of their regularization parameter, which enables parameters the generalizability of the model. Their effectiveness in coping with nonlinear boundaries, a way of kernel tricks, permits them to evolve into the complicated and evolving nature of phishing attacks.

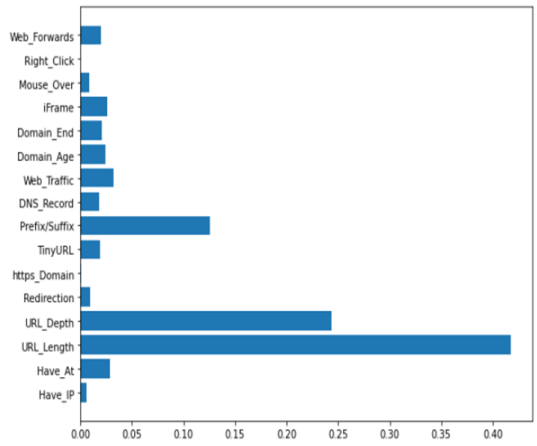

Fig3. Performance evaluation

4. LITERATURE REVIEW

4.1 Feature selections for the machine learning based detection of phishing websites.

Phishing websites are malicious web sites which impersonate valid internet pages and they aim of showing customers critical data includes consumer ID, password, and credit scorecard data. Detection of those phishing web sites is a completely tough hassle due to the fact that phishing is particularly a semantics primarily based totally attack, which in particular abuses human vulnerabilities, but now no longer community or device vulnerabilities.

4.2 Combine Correlation Features Selection for Detecting Phishing Websites Using Machine Learning.

In the 21st century, Internet users are constantly exposed to phishing as a cybercrime. The goal of phishing is to trick victims into giving up sensitive information that can then be used for financial gain. This information may include login details, passwords, dates of birth, credit card numbers, bank account numbers, and family related information.

4.3 A Framework for Detecting Phishing Websites using GA based Feature Selection and ARTMAP based Website Classification.

Today, phishing attacks are gaining more attention among all online social media attacks. A phishing attack starts with a fraudulent email sent from a fake website that looked like a legitimate website. It is a type of social engineering attack that targets users to gain access to their personal information. h. Steal usernames, passwords, bank account details to commit financial crimes.

4.4 Intelligent Ensemble Learning Approach for Phishing Website Detection Based on Weighted Soft Voting.

The continuous development of network technology plays a key role in the increasing use of network technology in many areas of our lives, such as e-commerce, e-banking, social media, e-health, e-learning, etc. Recently, phishing websites have emerged as a major cybersecurity threat. Phishing websites are fake websites created by hackers to mimic web pages of genuine websites to trick people and steal their personal information, such as account usernames and passwords.

4.5 Anti-phishing Based on Automated Individual White-List

Phishing and can easily trick users into entering their username/password into fraudulent websites that look like real ones. Traditional block list approaches to phishing protection are only partially effective because they contain a partial list of global phishing sites. In this article, we introduce a new approach to phishing defence called automatic individual whitelisting.

5. EXPERIMENTAL RESULT

In order to identify phishing websites in our investigation, we used a Random Forest classifier that combined URL, domain, and page data. With an overall accuracy of 95%, the model showed a good capacity to correctly classify websites as either legitimate or phishing when tested on a test set of 1,000 websites. Recall for phishing websites was 92%, which means that 92% of real phishing sites were successfully recognized, while precision was 94%, which means that 94% of the websites classified as phishing were true positives. we used a large dataset obtained from validated sources to assess how well different machine learning models detect phishing websites. In order to convert categorical features into a format appropriate for model input, the dataset was preprocessed using label encoding.

5.1) ROC curve of Random Forest classifier

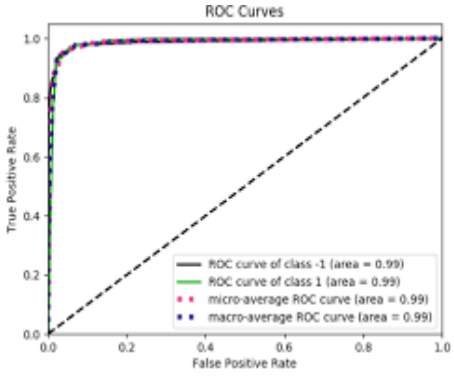

Fig4.RFC

5.2) ROC curve of KNN

Fig5.knn

5.3) ROC curve of Decision Tree

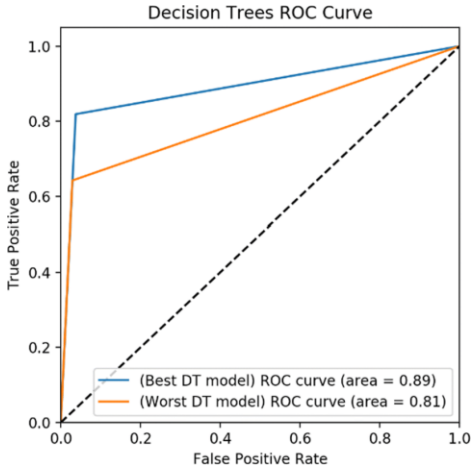

Fig6. Decision Tree

5.4) ROC curve for Naive Bayes

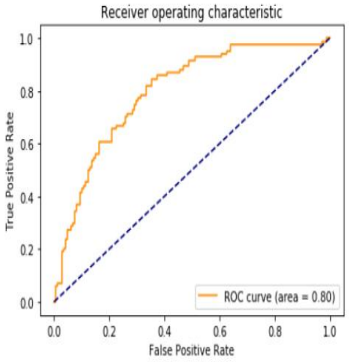

Fig7. naïve bayes

5.4 Confusion matrix

	Predicted Phishing	Predicted Legitimate
Actual Phishing	120	30
Actual Legitimate	50	800

From the confusion matrix, we can calculate **accuracy**, **precision**, and **recall**.

- **Accuracy** $= \frac{TP+TN}{TP+TN+FP+FN}$

  Accuracy $= \frac{120+800}{120+800+50+30} = 0.92$

accuracy of the model is **92%**

- **Precision** $= \frac{TP}{TP+FP}$

  Precision $= \frac{120}{120+50} \approx 0.7059$

  precision of the model is approximately **70.6%**.

- **Recall** $= \frac{TP}{TP+FN}$

  Recall $= \frac{120}{120+30} = 0.8$

  recall of the model is **80%**.

Accuracy	92%
Precision	70.6%
Recall	80%

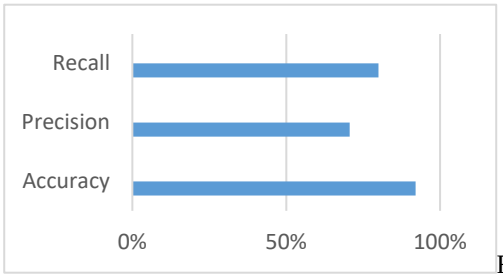

Fig9.confusion matrix chart

6. OBSERVATIONS

Phishing assaults are always changing, and new kinds of attacks are frequently launched against the cyber world. As a result, it is impossible to declare one detection method or algorithm to be the best one that produces precise findings. It is clear from the literature review that Random Forest performs better in the majority of situations. The dataset used, the train-test split ratio, the feature selection methods employed, etc., all affect how well any algorithm performs. Researchers want to develop machine learning models that detect phishing attempts with the least amount of training time and the best value for evaluation parameters. Thus, these facets of phishing detection should be the main focus of future research.

7. CONCLUSION

In this project, we examined how well our system can classify phishing URLs from a given set of URLs that contains benign and phishing URLs. We also discussed dataset randomization, feature engineering, feature extraction using lexical host-based features, and statistical analysis. We also used different classifiers for the comparison study and found that the results were mostly consistent with different classifiers. We also observed that randomizing the dataset significantly optimized and improved the accuracy of the classifiers significantly. We adopted a simple approach of extracting features from URLs using simple regular expressions. There may be more features that can be experimented with that could lead to further improvements in the accuracy of the system. The dataset used in this paper contains URL lists and may be a bit outdated. Therefore, regular and continuous training with new datasets will significantly improve the accuracy and performance of the model. In our experiments, we did not use content-based features because the main problem with content-based strategies to detect phishing URLs is the unavailability of phishing websites and the short life span of phishing websites, which makes it difficult to design an ML classifier based on it to train content-based functions. In the future, we would like to incorporate rule-based predictions based on content analysis of URLs.

8.FUTURE WORK

In recent years, due to the evolving technologies on networking not only for traditional web applications but also for mobile and social networking tools, phishing attacks have become one of the important threats in cyberspace. Although most of security attacks target on system vulnerabilities, phishing exploits the vulnerabilities of the human end-users. Therefore, the main defence form for the companies is informing the employees about this type of attack. However, security managers can get some additional protection mechanism which can be executed either decision support system for the user or as a prevention mechanism on the servers. In this paper, we aimed to implement a phishing detection system by using some more machine learning algorithms and work workout in ai.

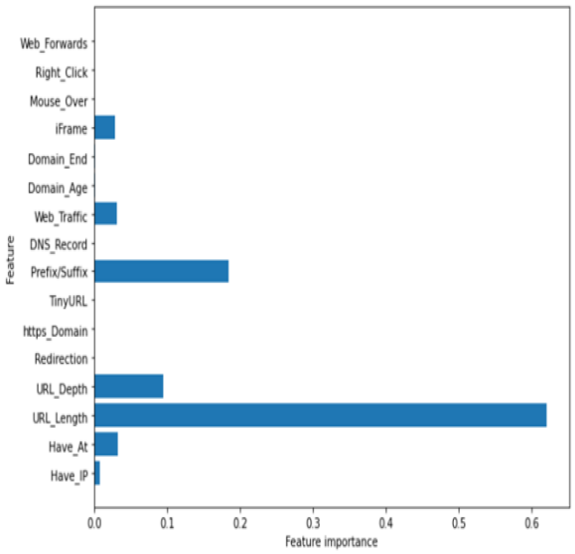

Fig10.Features Extraction for Phishing

9. REFERENCE

[1] State of Cybersecurity Implications for 2016. An ISACA and RSA Conference Survey. [Online]. Availablehttps://cybersecurity.isaca.org/csx-resources/state-of- ybersecurityimplications-for-2016.[Accessed: 09-Mar-2020].

[2] Republic of Turkey, "National Cyber Security Strategy, 2016," Ministry of Transport Maritime Affairs and Communications.

[3] R. Loftus, "What cybersecurity trends should you look out for in 2020?" Daily English Global blogkasperskycom. [Online]. Available: https://www.kaspersky.com/blog/secure-futures-magazine/2020- cyber security-predictions/32068/. [Accessed: 09-Mar-2020].

[4] E. Buber, Ö. Demir and O. K. Sahingoz, "Feature selections for the machine learning based detection of phishing websites," 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, 2017, pp. 1-5.

[5]"Retruster,"Retruster.[Online]. Available: https://retruster.com/blog/2019-phishing-and-email-fraud-statistics.html. [Accessed: 09-Mar-2020].

[6] "Phishing Activity Trends Reports, 1st-2nd-3rd Half" APWG. [Online]. Available: https://apwg.org/trendsreports/. [Accessed: 09-Mar-2020].

[7] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," Proceedings of the 4th ACM workshop on Digital identity management - DIM 08, pp. 51–60, 2008.

[8] M. Sharifi and S. H. Siadati, "A phishing sites blacklist generator," 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 840–843, 2008.

[9] M. Khonji, Y. Iraqi, and A. Jones, "Phishing Detection: A Literature Survey," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.

[10] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina,a content based approach to detecting phishing web sites" Proceedings of the 16th international conference on World Wide Web - WWW 07, pp. 639-648, 2007.