IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11711

A STUDY ON DETECTING PHISHING WEBSITE USING MACHINE LEARNING

Yashwanth G R, Chinmaya S C, Vasudha J, Raghavendra Prasad Shetti, Neha R

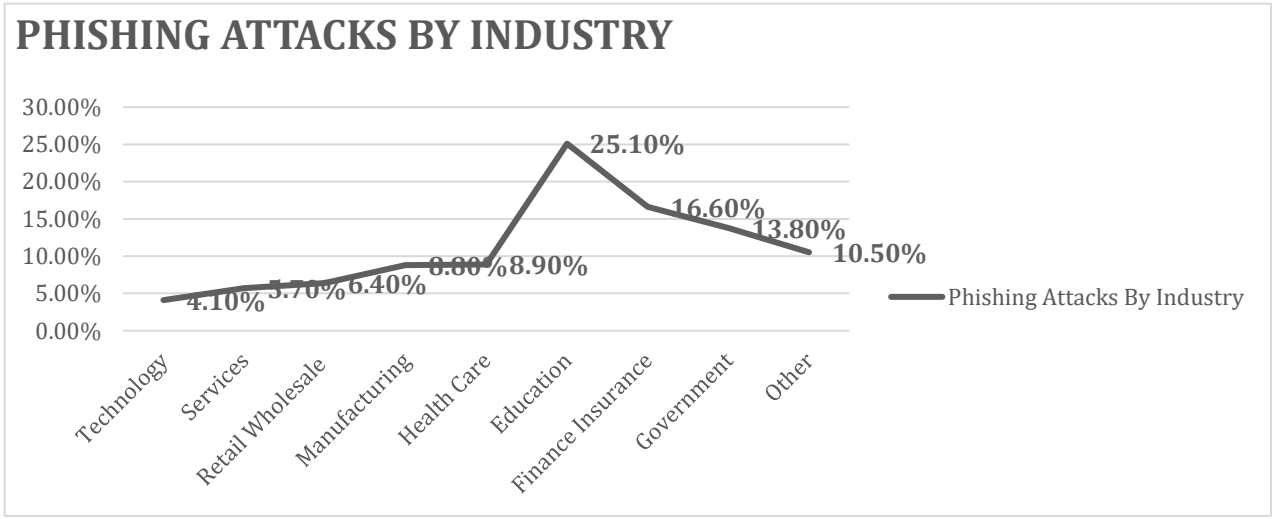
Department of AIML, Dayananda Sagar Academy of Technology and Management

Abstract: Phishing is a fraud attempt in which a scammer acts as a trusted person or reality to gain sensitive information from an internet user. In this Methodical Literature check (SLR), different phishing discovery approaches, videlicet Lists Grounded, Visual Similarity, Heuristic, Machine Learning, and Deep Learning grounded ways, are studied and compared. For this purpose, several algorithms, data sets, and ways for phishing website discovery are revealed with the proposed exploration questions. A methodical Literature check was conducted on 80 scientific papers published in the last five times in exploration journals, conferences, leading shops, the thesis of experimenters, book chapters, and from high- rank websites. The work carried out in this study is an update in the former methodical literature checks with further focus on the rearmost trends in phishing discovery ways. This study enhances compendiums' understanding of different types of phishing website discovery ways, the data sets used, and the relative performance of algorithms used. Machine literacy ways have been applied the most, i.e., 57 as per studies, according to the SLR. In addition, the check revealed that while gathering the data sets, explorationers primarily penetrated two sources 53 studies penetrated the PhishTank website (53 for the phishing data set) and 29 studies used Alexa's website for downloading licit data sets. Also, as per the literature check, utmost studies used Machine literacy ways; 31 used Random Forest Classifier, which, as per different studies, achieved the loftiest Accuracy, 99.98, for detecting phishing websites.

Keywords: Phishing, URL, Hyperlinks, Machine Learning, Random Forest, K-means, SVM.

INTRODUCTION

Phishing is a social engineering attack [1] is considered the most common system used by cybercriminals to pierce, for illustration, particular information of an Internet stoner credit card details, usernames and watchword [2] .occasionally bushwhackers appear phishing attacks to distribute malware online [3](Gupta etal., 2021). There are numerous types of phishing attacks and they're known and not limited to fraud, malware grounded attacks. phishing, DNS- grounded phishing, data theft, dispatch/ spam, web- grounded delivery and phishing by phone as shown in Figure 1 [4](Kathrine etal., 2019). Phishing attacks come by numerous forms and generally involve a different communication channels similar as dispatch, instant messaging, QR canons [5](Geng etal., 2018) and communication mass media bushwhackers impersonate well- known banks, credit cards



IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11711

agencies or well- knowne-commerce spots to scarify or move druggies to log into a phishing point and give credentials they may latterly lament. For illustration, the stoner can admit instant communication indicating a problem with their bank account and you'll be taken to a website that looks like a bank's website The client adds their credentials without vacillation applicable fields are captured by bushwhackers. culprits examiner this information and use it to gain access to the stoner's licit information accounts [6](Liu etal., 2021). According to a report by the Internet Crime Complaint Center(IC3), the FBI entered 791,790 complaints of suspected Internet crimes in 2020, further than 300,000 complaints compared to 2019 data [7](FBI, FBI press releases Cybercrime Complaints Center, 2020). colorful styles have been proposed in the literature to descry a phishing point, List- grounded, visual similarity, heuristics, machine literacy [8](Somesha etal., 2020; Nakamura and Dobashi, 2019) and Deep Learning ways [9](Basit etal., 2020). List- grounded Cybersurfers similar as Microsoft Edge, Firefox and Google Chrome uses list- grounded styles to descry phishing spots. Whitelist and blacklist are two types of list- grounded is approaching A whitelist contains a list of valid URLs that cybersurfers can use. This means that if a URL is whitelisted, the cybersurfer can load the web runner. At the same time, the blacklist database contains phishing or fraudulent URLs that stop cybersurfers to load web runners. The biggest strike is that it only takes a small change in the URL to jump List- grounded ways and help new phishing URLs lists must be streamlined constantly [10](Yang etal., 2021). Visual Similarity This approach evaluates the suspect and authentic websites grounded on colorful visual characteristics. Because the phishing runner looks veritably analogous to its licit runner runner, these tools compare parallels This approach uses CSS, textbook layout, source law, website totem, website screenshots, and other visual rudiments. Because these ways are similar to preliminarily visited or saved websites of a suspicious website can not descry zero- hour phishing attacks [11](Jain and Gupta, 2018). Heuristics a heuristic approach uses deduced functions phishing point This strategy is grounded on several features what separates a phishing point from a real bone . These styles collect information from colorful sources similar as URLs, textual content, DNS, digital instruments and website business. set of features, training samples and bracket algorithms each contribute the success of this system. One of the advantages of this technology that it can descry zero- hour phishing attacks [12](Jain and Gupta, 2018). Machine literacy Machine literacy is common these days approach to descry phishing spots [13](Sindhu etal., 2020). General attributes similar as URL information, point structure, and JavaScript attributes are collected to represent phishing URLs. Figure 1. Types of phishing attacks.A. Safi andS. Singh Journal of King Saud University – Computer and Information lores 35(2023) 590 – 611 591 and related websites. also, grounded on those features, phishing data sets are attained. After that, Machine Learning classifiers are trained to descry the phishing website grounded on those features [14] Zhu etal., 2020). This fashion works veritably well with Big Data sets(having high haste, Variety, Volume, Value, and Veracity). Machine literacy- grounded classifiers achieved further than 99 curacy, which proved to be the most effective system [15](Alkawaz etal., 2021).

RELATED WORK

Numerous authors have explored the discovery of phishing websites. still, only a many have conducted a methodical literature review on the content, as described below. Qabajeh etal. [16](Qabajeh etal., 2018) lately worked on conven tional vs automated phishing discovery ways. The conven tionalanti-phishing styles include raising mindfulness, educating druggies, conducting periodic training or factory, and using a legal perspective. The Motorized or automated anti phishing approaches addresses about list- grounded and Machine Learning Grounded ways. More importantly, the paper compares these approaches ' parallels, positive and negative rudiments from the stoner and performance perspectives. According to this study, Machine literacy and rule induction are suitable for combating phishing attacks. The limitations of this work are the review is grounded on 67 exploration particulars, and the study doesn't include Deep Learning ways for phishing website discovery. Zuraiq & Alkasassbeh [17](Zuraiq and Alkasassbeh, 2019) carried out a comprehensive review of current phishing discovery styles. The study discussesanti-phishing ways similar as Heuristic, Content Grounded, and Fuzzy rule- grounded approaches. The study indicated that there are better styles for relating phishing websites. The background of the work is grounded on exploration conducted between 2013 and 2018. The downsides of this work are that it anatomized only 18 studies and didn't include Machine Learning, List Based and Deep Learning approaches for phishing website discovery. Kunju etal. [18](Kunju etal., 2019) used a check system to descry phishing attacks. The exploration provides several phishing attack discovery results and methodologies. According to the exploration, numerous of the proposed results were set up to be inadequate in furnishing results to phishing attacks. The

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11711

literature in this work includes only 14 studies which are in the period between 2007 and 2019. The study discusses only Machine Learning ways for phishing website discovery. Benavides etal. [19](Benavides etal., 2020) conducted a methodical review to dissect different approaches of other experimenters for detecting phishing attacks by applying Deep Learning algorithms.

In conclusion, there's still a significant gap in the area of Deep Learning algorithms for phishing attack discovery. The literature in this work includes only 19 studies published between 2014 and 2019. Only exploration papers with the essential motifs of phish ing and Deep Learning are considered in this paper. Athulya & Praveen [20](Athulya and Praveen, 2020) addressed dif ferent phishing attacks, phishers ' most recent phishing tactics, andanti-phishing strategies. In addition, the composition aims to raise mindfulness regarding phishing attacks and strategies used for phishing discovery. According to this study, the stylish way to pre articulation phishing attacks is to educate druggies about the different types of phishing attacks. druggies can choose the stylish security software tools or operations to descry phishing attacks, similar as anti phishing cybersurfer extensions.

The literature in this work is grounded on nine exploration particulars. The study doesn't include Deep Learning ways for phishing website discovery. Basit etal. [9](Basit etal., 2020) reported a check on artificial intelligence- grounded phishing discovery ways. The authors habituated statistical phishing reports to examine the detriment and trends of phishing attempts. In the paper, Antiphishing evaluations are classified into four orders Machine literacy, mongrel literacy, script- grounded and Deep literacy. The exploration shows that Machine literacy procedures produce the stylish results compared to other approaches. The work is grounded on literature published in the last ten times and anatomized only 21 exploration particulars.

Kathrine etal. [21](Kathrine etal., 2019) presented a frame to descry and help different types of phishing attacks. According to this study, Machine literacy grounded algorithms effectively descry true positive results. The limitations of this study are the literature in this work bandied only 11 studies, and the exploration does not include Deep Learning ways for mollifying phishing websites. Korkmaz etal [22].(Korkmaz, 2020) proposed a review work for opting features that can be used in URL- grounded phishing detec tion systems. This exploration aims to produce a general check resource for scientists who work on web runner bracket or net work security. This study's limitation is that the work bandied only five studies in the literature. Arshad etal. [23](Arshad etal., 2021) presented different types of phishing andanti-phishing ways in their study. The SLR eval uated that phone phishing, Dispatch Spoofing, shaft phishing, and Dispatch Manipulation are the constantly used phishing ways. According to this study, the loftiest Accuracy was achieved through Machine Learning approaches.

The exploration is limited by the fact that it's grounded on only 20 studies. Catal etal [24].(Catal etal., 2022) worked on a methodical literature review, which answered nine exploration questions. The study's main end is to identify, assess, and synthesize the results of Deep Learn ing approaches for phishing discovery. According to this study, Supervised ML algorithms were applied in 42 studies out of 43. The most habituated algorithm was DNN, and the stylish performance was given by DNN and Hybrid DL algorithms. The work only dis cusses Deep literacy related studies for phishing discovery. Table 1 shows only three SLRs published in the last five times about phishing website discovery ways in the five influential journals named in the current study.

COMPARATIVE STUDY

The comparative analysis between conventional, machine learning, and deep learning techniques for anti-phishing measures reveals distinct strengths and limitations. Conventional methods like user awareness training and legal measures offer simplicity but are constrained by user compliance. Machine learning techniques, such as rule-based and ensemble methods, exhibit high accuracy and adaptability to evolving threats, albeit demanding expertise for implementation. Conversely, deep learning methods, particularly Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs), hold promise for superior accuracy, yet their practical application requires further exploration due to limited research. Thus, while conventional methods are straightforward to implement with inherent limitations, machine learning and deep learning techniques offer increasingly sophisticated solutions demanding varying degrees of expertise and further investigation.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11711

COMPARISON OF METHODOLOGIES:

Random Forest, Naive Bayes, and Decision Tree are three popular algorithms used in machine learning for classification tasks. Random Forest belongs to the ensemble learning category, characterized by high model complexity, which helps in achieving high accuracy and robustness to overfitting. However, it can be computationally expensive and lacks interpretability due to its black-box nature. Naive Bayes, a probabilistic algorithm, has low model complexity and is efficient, making it suitable for large datasets. It assumes independence among features, which might not always hold true, leading to lower accuracy compared to other algorithms. Decision Tree, a classification algorithm, has medium model complexity and offers interpretable results, making it easy to understand. However, it is prone to overfitting and sensitive to irrelevant features during training. Overall, each algorithm has its strengths and weaknesses.

Table: comparative study

Feature	Conventional Techniques	Machine Learning Techniques	Deep Learning Techniques
Ref No.	[16]	[16, 18, 19, 21, 23]	[19, 24]
Research Work/Paper	A Comparative Analysis of Conventional and Machine Learning Based Anti-Phishing Techniques	Various research papers on Machine Learning for phishing detection	Reviews on Deep Learning for phishing detection
Author / Year	Qabajeh et al., 2018	-	-
Techniques	User awareness training, education, legal measures	Rule-based, statistical, ensemble methods	Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs)
Experiments/Observations	Compares effectiveness of conventional vs. automated approaches	High accuracy in identifying phishing attempts	Promising results, requires further research
Remarks	Easy to implement, limitations on user awareness	Adapts to evolving attacks, requires expertise	Potentially highest accuracy, limited research on applications

Table : Methodologies Comparison

Feature	Random Forest	Naive Bayes	Decision Tree
Algorithm Type	Ensemble Learning	Probabilistic	Classification
Model Complexity	High	Low	Medium
Training Time	Slower	Faster	Faster
Interpretability	Lower	Higher	Medium
Data Preprocessing	Less sensitive	More sensitive	Moderately sensitive
Strength	High accuracy, robust to overfitting	Efficient, good for large datasets	Easy to understand, interpretable results
Weakness	Black box (difficult to explain decisions), can be computationally expensive	Assumes independence of features (may not be realistic), lower accuracy compared to others	Prone to overfitting, sensitive to irrelevant features
Suitability for Phishing Detection	Excellent	Good for initial exploration	Good for initial exploration
Additional Notes	Often the best performing algorithm for phishing detection	Simple and fast, can be a good baseline	Can be a good building block for ensemble methods

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11711

METHODOLOGIES USED

Since social engineering is a problem with phishing, effective defences are developed for several facets of education, legal oversight as well as specific methods [25]. The primary focus of this inspection is on specific methods for identifying phishing websites. Three orders—list-based, heuristic, and machine literacy styles—of techniques for identifying phishing websites have been created [26].
The manually reported and validated whitelists and blacklists by systems are matched by the list- grounded techniques. A collection of verified licit URLs or disciplines is called a whitelist. A blacklist is, of course, a collection of legitimate phishing websites. The website will be added to the blacklists since a stoner exposed and confirmed it to be a phishing website. This could prevent other drug users from falling victim to the same fate. Heuristic techniques use a collection of features extracted from the website's textual content to identify phishing web runners, then compare those features to the licit bone. The theory behind the method is that bushwhackers typically fool drug users by pretending to be well-known websites.
The properties of the website also influence machine literacy styles, which enable the model to learn from a batch of data with structured information and predict whether a new website is a phishing website. The identification of phishing websites is a major problem in the field of machine literacy.

List- Grounded Approaches

In 2016, Jain and Gupta presented a whitelist-based, bus-streamlined method to protect against customer-side phishing assaults. The experimental findings show that it attained 86.02 delicacy and a false-positive rate of 1.48, indicating a false recommendation for phishing attempts. This approach's quick access time ensures a real-time terrain and products, which is another advantage [27].

Heuristic Strategies

Three phases make up the PhishWHO phishing discovery approach, which was presented by Tan et al. It first gathers identify keywords from the runner's HTML using a weighted URL commemorative system, then groups the N-gram model. Second, it uses the keywords to search the legal domain and the licit website using popular search engines. The target website's sphere and the lawful sphere are then compared to ascertain whether or not the target website is a phishing website [28]. To determine whether the website was legitimate, Chiew et al. used a totem image from the source [29].
In this work, the authors used machine learning methods to uproot a totem using web runner photos. They then used the Google search engine to query the sphere using the term "totem." Consequently, this order hunt machine- grounded strategy was also dubbed by some experimenters.

Machine literacy- Grounded styles

With improved delicacy performance and fewer false positive rates than other approaches, machine literacy grounded remedies are suggested as a means of mitigating dynamic phishing attempts [25]. As a result, the six components of the machine literacy strategy are data collection, point birth, model training, testing, and prognostication. The findings of the machine literacy-based phishing website discovery are based on this flowchart, which can be optimized to improve efficiency in one or more corridors.

3.3.1. Information Gathering and the Starting Point Every strategy starts with data, which also seems to be a crucial factor in performance. There are two methods for gathering data: downloading URLs straight from the Internet and using publicly available datasets. Each row's data object in these three publicly available datasets comprises multiple features that are extracted from a URL and a class marker. Using data mining programs or available APIs, websites' original URL strings could be retrieved.

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11711

In 2012, Mohammad et al. suggested an algorithmic method to rank the elements of phishing websites and determine the importance of each item [30].The phishTank library [31] was used by the authors of that paper to gather 2500 phishing URLs, from which they extracted 17 features. These features were categorized into three orders: address bar grounded features, anomalous grounded features, and HTML and JavaScript grounded features. The majority of the functionalities were automatically removed from the web runner's source code and URL without relying on outside services. However, the DNS record and the sphere's age were removed from the WHOIS database [32].The Alexa database was used to determine the web runner's rank [30]. In the interim, the writers established a weight for every point and explained an IF-ELSE rule.

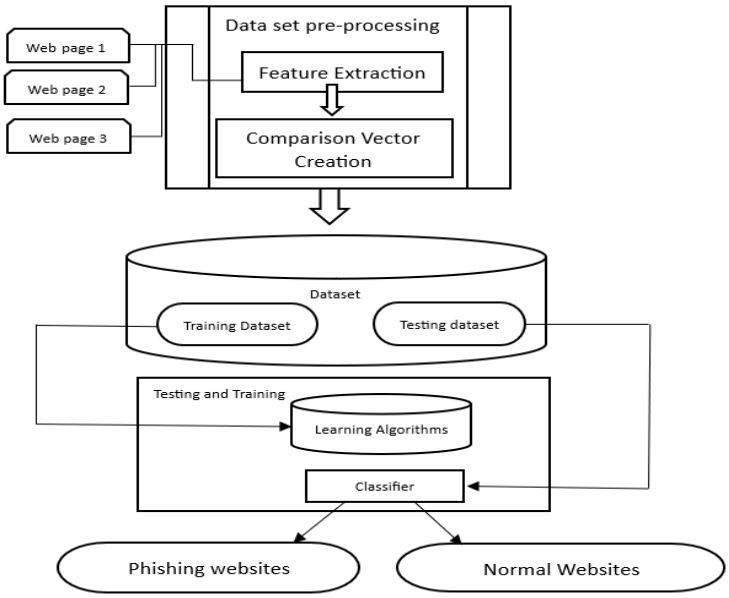
The value of each point might be expressed numerically as an independent member of the set{1, 0, −1}, with each representing licit, suspect, and phishing in turn [30]. A phishing website dataset with 30,525 examples and 30 features was published by Mohammad et al. in 2015 on the UCI Machine Learning Repository. This dataset served as a basis for machine literacy-based phishing discovery results and was widely utilized in various related investigation fields [31]. Similarly, in 2018, Choon released a phishing dataset on Mendeley that included 10,000 data rows and 48 features that were taken from 5000 websites each, phishTank and OpenPhish for phishing webpages, and Alexa and Common Bottleneck for licit webpages [32].

Evidently, in comparison to other machine literacy initiatives, the publicly available datasets are quite tiny. As a result, several resampling techniques are used in the process, such as N-foldcross-validation, which divides the data into N pieces and repeats the procedure N times, choosing one piece of data for testing and the other pieces for training. However, other experimenters gathered URLs from the Internet, such as those from licit websites like Common Bottleneck, Alexa, and Dmoztools.net, and phishing URLs like phishTank, OpenPhish, and Spamhaus.org, and they also independently parsed the features. Since natural language processing (NLP) has been developed so successfully, many researchers have taken character-position characteristics from URL strings based on NLP and fed them into deep literacy models to make them more delicate.

Inapplicable cybersecurity measures and not relying on outside network services are two of this system's key benefits [36]. Because the URL is made up entirely of characters, it lacks semantics and is difficult to discern between words. Character-position features are employed in a manner akin to that of TF-IDF features. Term frequency – Inverse Document frequency is known as TF–IDF. Every character is represented as a word by the character position. The program determines the TF-IDF score for every character in the URL string and creates a matrix containing those scores, which indicate how applicable a character is.

Using “https//www.google.com/”(visited on 18 July 2021) as a case, it consists of 17 characters(“ h ”, “ t ”, “ t ”, ” p ”, “ s ”, “”, “// ”, “ w ”, “ w ”, “ w ”, “. ”, “ g ”, “ o ”, “ o ”, “ g ”, “ l ”, “ e ”, “. ”, “ c ”, “ o ”, “ m ”) and is called character position 17- g in the corpus. thus, it'll induce a vector with 17 TF- IDF scores. One character’s TF- IDF score is calculated as in the calculation expression shown below TF(t, d) = (Number of times character t appears in a document d)/(Total number of characters in the document d)----(1) IDF(t, D) = loge ((Total number of documents D)/(Number of documents with character t in it))----(2) TFIDF(t, d, D) = TF(t, d) × IDF(t, D)

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11711



Point Selection :

Point selection is the process of automatically opting important features which contribute the most to the machine literacy model. Having nearly applicable features in the input can enhance the performance of the model, drop training time(especially in deep literacy models), and reduce overfitting issues. Generally, point selection system ologies could be classified into three orders the sludge system, wrapper system, and bedded system.

Information gain (IG), relief ranking, and recursive point elimination (RFE) were used by Zamir et al. to eliminate extraneous features for phishing

discovery. They also introduced principal component analysis (PCA) for attribute testing [37]. IG is an index that indicates the importance of features by computing class probability, point probability, and class probability under a point condition. RFE is a widely used point reduction algorithm that eliminates the least important features in the training process until the error rate meets prospects.

The closest neighbor hunt algorithm finds two adjacent data points; these point values are compared to determine the point value score. The point values are then sorted to determine the point value weight based on the score. This method is known as a relief ranking system. This technique was applied to the UCI dataset for phishing website bracket by Shabudin et al. Following the point selection procedure, eight extra features with zero scores were eliminated and features with weighted rankings were obtained.

Zabihimayvan and Doran applied Fuzzy Rough Set(FRS) proposition to elect important features from the UCI dataset and Mendeley dataset for phishing discovery operations [26].Fuzzy Rough Set(FRS) proposition is an extension of Rough Set(RS) proposition. RS is a system to find a decision boundary by calculating the equivalency of each data point grounded on certain features and the same classes, similar as websites A and B both being phishing websites and their features a and b having the same value. RS is suitable for the original UCI dataset in which the features are employed as a separate value; that is, they're an element of set { −1, 0, 1}. still, after the dataset executes the nominalization process, the value of the point is transferred to a nonstop number from 0 to 1, and the FRS strategy is applied.

El- Rashidy introduced a new fashion to elect features for a web phishing discovery model in 2021. The point selection system contains two phases. The first phase calculates each point's absence impact by training the arbitrary timber model with a new dataset that removes one point and figures out the delicacy. After the absence of each element in the circle, a point line ranked from high to low delicacy is attained. The alternate stage is to train and test the model by starting from one point, adding a new point from the ranked point list each time to form a dataset, calculate the delicacy of each time, and eventually find the point subset with the loftiest delicacy. This system workshop to elect the most effective point

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11711

subset. still, since each new dataset must go through the algorithm training and testing process, a high computational complexity and a long computation time are involved. For illustration, if the UCI dataset has 30 eigenvalues, also the first stage circles 30 times, the alternate stage circles 30 times, and the tree algorithm training must be performed each time. thus, this methodology is suitable for small point sizes and single classifiers.

Features Group	S.NO	Phishing websites features
URL and Domain Identity	1	Using the IP Address
	2	Long URL to hide the suspicious part
	3	Using URL shortening services 'Tiny URL'
	4	RL's having @ symbol
	5	Redirecting using '//'
	6	Adding prefix or suffix separated by(-)domain
	7	Sub domain and Multi-sub doamins
	8	HTTPS(HyperText Transfer Protocol with Secure Sockets Layer)
	9	Domain Registration Length
	10	Favicon
	11	Using Non-Standard Port
	12	The Existence of 'HTTPS' token in the Domain part of URL
Abnormal Based Features	13	Request URL
	14	URL of Anchor
	15	Links in<Meta>,<Script> and <Link> tags
	16	Server From Handler (SFH)
	17	Submitting information to E-mail
	18	Abnormal URL
HTML and JAVA Script- based Features	19	Website forwarding
	20	Status Bar Customization
	21	Disabling Right Click
	22	Using Pop-Up Window
	23	IFrame Redirection
Domain Based Features	24	Age of Domain
	25	DNS Record
	26	Website Traffic
	27	PageRank
	28	Google Index
	29	Number of links pointing to the page
	30	Statistical Reports based Feature

Modelling

Machine literacy- grounded models can be classified into three orders single classifier, mongrel models, and deep literacy. mongrel models combine further than one algorithm applied to the training process. Phishing website discovery is a double bracket problem. Some extensively used bracket algorithms are listed below. SVM A support vector machine(SVM) is a supervised literacy algorithm that classifies data points into two sections and predicts new data points belonging to each section. It is suitable for direct double bracket, which has two classes labeled, and the classifier is a

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11711

hyperplane with N confines applicable to the number of features. The core idea of this algorithm is to maximize the distance between the data point and the segmentation hyperactive aeroplane . For illustration, there are two classes — phishing and licit — and a 29- dimension hyperplane when we use the UCI dataset for training the SVM model. Decision Tree A decision tree is a popular machine learning algorithm, and the model sense is a tree structure. Each knot in the decision tree is a point; each stem presents a point value and a possibility, and the last knot presents the result. The further straightforward tree structure tends to have better performance. When trees grow veritably deep, it probably leads to overfitting training datasets. Random Forest A arbitrary timber is an ensemble of decision trees for bracket and retrogression. Random timbers reduce the overfitting problem by classifying or comprising the affair of individual trees in training processing. thus, arbitrary timbers generally have advanced delicacy than decision tree algorithms. k- NN A k- nearest neighbours' algorithm(k- NN) is anon-parametric bracket algorithm that makes prognostications by chancing analogous data points through calculating the distance between the target and the nearest neighbors. There are some styles to calculate the distance with respect to the Euclidean distance for nonstop data and the Hamming distance for separate values. In particular, it doesn't have a training process, and each vaticination will take a long time. thus, this algorithm is generally not suitable for real- time scripts. Bagging, also called bootstrap aggregating, is an ensemble meta- literacy algorithm for perfecting other machine learning algorithms' performance in bracket and retrogression. The bootstrapping procedure divides the original training dataset into N pieces and uses e-testing ways to induce the same size of the original dataset in each piece and also conducts bracket in N duplications that could be executed in parallelization. Eventually, the aggregating process combines N classifier labors by comprising or voting.

CONCLUSION

Phishing is a way where a stoner's private information can be fluently violated it may be through e-mail or be a website. As we all know the operation of the internet is vast and how all the effects are fluently available online it may be shopping for clothes or perhaps shopping for the ménage particulars. People prefer online styles rather than standing in a line for hours. Due to these reasons, the phisher has a wide compass to apply phishing. The experimenters have worked on this area, but there is not any single fashion that can descry all kinds of phishing attacks. As and how technology is adding in the world, so are the phishing bushwhackers are coming up with new styles to attack day by day. This enables us to figure out the effective classifier for the discovery of phishing. " Phishing is a major problem, which uses both social engineering and specialized deception to get druggies ' important information similar as fiscal

data, emails, and other private information. Phishing exploits mortal vulnerabilities, thus, utmost protection protocols can not help the whole phishing attacks. numerous of them use the black- list/ white- list approach, still, this can not descry zero hour phishing attacks, and they aren't suitable to descry new types of phishing attacks. thus, with the help of machine learning the discovery of phishing can be made effective as well as effective tools. To use the machine literacy approach, a lot of data as needed, and also the features of these data is important" [38]. The paper is aiming to identify and list the features of machine- literacy to help in the discovery of phishing websites.

FUTURE SCOPE

Compass for unborn Work There are numerous features that can be bettered in the work, for colorful other issues. The discovery can be made further like to develop for the discovery of all kinds of phishing attacks in the presence of objects like flash. Identity birth is an important operation and it was bettered with the Optical Character Recognition(OCR) system to prize the textbook and images. The effective rules to identify any given suspicious website/ webpage for discovering if the runner is a phishing target, it should be designed in a way to further ameliorate the performance and delicacy of the system. also, it has come a kind of challenge for the inventors to develop a discovery system, which can descry any website and give delicacy for the phishing. To add to this, the features of static and dynamic complement each other, and are considered as important in achieving high delicacy.

REFERENCES

[1] G. Palaniappan, S. Sangeetha, B. Rajendran, S. G. Sanjay and B. S. Bindhumadhava, "Malicious Domain Detection Using Machine Learning on Domain Name Features, Host-Based Features and Web-Based Features," Procedia Comput. Sci., vol. 171, no. 2019, pp. 654-661, 2020. https://doi.org/10.1016/j.procs.2020.04.071

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11711

[2] S. Paliath, M. Abu Qbeitah and M. Aldwairi, "PhishOut: effective phishing detection using selected features.," IEEE, 2020.

[3] A. V. Ramana, K. L. Rao and R. S. Rao, "Stop-Phish: an intelligent phishing detection method using feature selection ensemble," Social Network Analysis and Mining, vol. 11, 2021. https://doi.org/10.1007/s13278-021-00829-w

[4] 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," Computer Communications, vol. 175, pp. 45-47, 2021. https://doi.org/10.1016/j.comcom.2021.04.023

[5] J. W. Kathrine,G, P. M. Praise, A. A. Rose and E. C. Kalaivani, "Variants of phishing attacks and their detection techniques," ICOEI, pp. 225-259, 2019. https://doi.org/10.1109/ICOEI.2019.8862697

[6] G. G. Geng, Z. W. Yan, Y. Zeng and X. B. Jin, " RRPhish: Anti-phishing via mining brand resources request," IEEE International Conference on Consumer Electronics,, pp. 1-2, 2018. https://doi.org/10.1109/ICCE.2018.8326085

[7] D. J. Liu, G. G. Geng, X. B. Jin and W. Wang, "An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment," Computer Security, p. 110, 2021. https://doi.org/10.1016/j.cose.2021.102421

[8] FBI, "FBI Releases the Internet Crime Complaint Center 2020 Internet Crime Report, Including COVID-19 Scam Statistics. News, 2021," FBI, 2021. [Online]. Available: https://www.fbi.gov/news/pressrel/press-releases/fbi-releases-the-internet.

[9] M. P. A. R. R. R. V. Somesha, "Efficient deep learning techniques for the detection of phishing websites.," Sadhana – Acad. Proc. Eng.Sci. , 2020. https://doi.org/10.1007/s12046-020-01392-4

[10] L. Z. J. W. X. L. Z. L. Z. H. Y. Yang, "An improved ELM-based and data preprocessing integrated approach for phishing detection considering comprehensive features," Expert System. Application. , 2020. https://doi.org/10.1016/j.eswa.2020.113863.

[11] A. G. B. Jain, "PHISH-SAFE: URL features-based phishing detection system using machine learning," Advances in Intelligent Systems and Computing, pp. 467-474, 2018. https://doi.org/10.1007/978-981-10-8536-9_44

[12] A. G. B. Jain, "Two-level authentication approach to protect from phishing attacks in real time.," Journal of Ambient Intelligence and Humanized Computing , vol. 9, p. 1783–1796., 2018. https://doi.org/10.1007/s12652-017-0616-z

[13] S. P. S. S. A. R. F. S. A. Sindhu, "Phishing Detection using Random Forest, SVM and Neural Network with Back propagation," Proceedings of the International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2020, p. 391–394., 2020. https://doi.org/10.1109/ICSTCEE49637.2020.9277256

[14] E. J. Y. C. Z. L. F. F. X. Zhu, "DTOF-ANN: An Artificial Neural Network phishing detection model based on Decision Tree and Optimal Features," Applied Soft Computing, vol. 95, 2020. https://doi.org/10.1016/j.asoc.2020.106505

[15] M. S. S. H. A. R. Alkawaz, "A Comprehensive Survey on Identification and Analysis of Phishing Website based on Machine Learning Methods," ISCAIE 2021 - IEEE 11th Symposium on Computer Applications and Industrial Electronics, pp. 82-87, 2021. https://doi.org/10.1109/ISCAIE51753.2021.9431794

[16] I. T. F. C. F. Qabajeh, "A recent review of conventional vs. automated cybersecurity anti-phishing techniques," Computer Science Review, vol. 29, pp. 44-55, 2018. https://doi.org/10.1016/j.cosrev.2018.05.003

[17] A. A. M. Zuraiq, "Review: Phishing Detection Approaches," 2019 2nd international Conference on New Trends in Computing Sciences,IEEE, pp. 1-6, 2019. https://doi.org/10.1109/ICTCS.2019.8923069

[18] M. D. E. A. H. B. S. Kunju, "Evaluation of Phishing Techniques Based on Machine Learning," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 963-968, 2019. https://doi.org/10.1109/ICCS45141.2019.9065639

[19] E. F. W. S. S. S. M. Benavides, "Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review," Developments and Advances in Defense and Security, vol. 152, pp. 51-64, 2019.https:/doi.org/10.1007/978-981-13-9155-2_5

[20] A. P. K. Athulya, "Towards the Detection of Phishing Attacks.," Proceedings of the 4th international Conference on Trends in Electronics and Informatics, ICOEI 2020 , pp. 337-343, 2020. https://doi.org/10.1109/ICOEI48184.2020.9142967

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11711

[21] A. Z. M. J. A. J. Z. Basit, "A Novel Ensemble Machine Learning Method to Detect Phishing Attack," 2020 IEEE 23rd International Multitopic Conference (INMIC), 2020. https://doi.org/10.1109/INMIC50486.2020.9318210

[22] G. P. P. R. A. K. Kathrine, "Variants of phishing attacks and their detection techniques," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 255-259, 2019. https://doi.org/10.1109/ICOEI.2019.8862697

[23] A. R. A. J. S. A. T. S. J. A. M. Arshad, "A Systematic Literature Review on Phishing and Anti-Phishing Techniques," arXiv, 2021. https://doi.org/10.48550/arXiv.2104.01255

[24] C. G. G. T. B. K. S. S. Catal, "Applications of deep learning for phishing detection: a systematic literature review," Knowledge and Information Systems (KAIS) , vol. 64, pp. 1457-1500, 2022. https://doi.org/10.1007/s10115-022-01672-x

[25] M. S. O. D. B. Krkmaz, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020. https://doi.org/10.1109/ICCCNT49239.2020.9225561

[26] K. M, "Feature Selections for the Classification of Web pages to Detect Phishing Attacks," HORA 2020 - 2nd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, 2020.

[27] S. F. A. V. A. H. T. Kumar, "MLSPD - machine learning based spam and phishing detection," Springer International Publishing, 2018.

[28] Y. Alsariera, V. Adeyemo, A. Balogun and A. Alazzawi, " AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites," IEEE Access 2020, vol. 8, p. 142532–142542, 2020.

[29] D. D. Zabihimayvan M, "Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection," 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2019.

[30] A. Jain and B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list.," EURASIP Journal on Information Security, 2016.

[31] C. Tan, K. Chiew, K. Wong, K. Wong and S. Sze, "PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder.," Decision Support Systems, vol. 88, p. 18–27, 2016.

[32] K. Chiew, E. Chang, S. Sze and W. Tiong, "Utilisation of website logo for phishing detection," Computers & Security, vol. 54, pp. 16-26, 2015.

[33] R. Mohammad, F. Thabtah and L. McCluskey, " An Assessment of Features Related to Phishing Websites Using an Automated Technique.," 2012 International Conference for Internet Technology and Secured Transactions, London, UK , p. 10–12, 2012.

[34] "PhishTank," July 2021. [Online]. Available: https://www.phishtank.com/index.php. [Accessed 18 july 2021].

[35] "WHOIS Search, Domain Name, Website, and IP Tools," 2021. [Online]. Available: https://who.is/. [Accessed 18 july 2021].

[36] "Keyword Research, Competitive Analysis, & Website Ranking|Alexa," [Online]. Available: https://www.alexa.com/. [Accessed 18 july 2021].

[37] R. Mohammad, F. Thabtah and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, vol. 25, p. 443–458, 2014.

[38] R. Mohammad, L. McCluskey and F. Thabtah, " UCI Machine Learning Repository: Phishing Websites Data Set. Available online:," [Online]. Available: https://archive.ics.uci.edu/mL/datasets/Phishing+Websites. [Accessed 26 March 2015].

[39] C. Tan, "Phishing Dataset for Machine Learning: Feature Evaluation.," Mendeley Data, 2018.

[40] 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, vol. [Crossref], p. 1514, 2020.

[41] " URL 2016|Datasets|Research|Canadian Institute for Cybersecurity|UNB. Available online:," [Online]. Available: https://www.unb.ca/cic/datasets/. [Accessed 18 july 2021].

[42] A. Zamir, H. U. Khan, T. Iqbal, N. Yousaf, F. Aslam, A. Anjum and M. Hamdani, "Phishing web site detection using diverse machine learning algorithms.," Electron. Libr., vol. [CrossRef], 2020.

