





Phishing Detection Techniques: A review

Majid Abdolrazzagh-Nezhad^{1*} , Nafise Langarib² 

¹Faculty of Computer and Industrial Engineering, Birjand University of Technology, Iran

²Department of Computer Engineering, Faculty of Engineering, University of Birjand, Iran

*Corresponding Author: abdolrazzagh@birjandut.ac.ir

ARTICLE INFO

Article history:

Received 20 July 2024

Revised 15 August 2024

Accepted 26 November 2024

Available online 31 January 2025

E-ISSN: 2580-829X

P-ISSN: 2580-6769

How to cite:

M. Abdolrazzagh-Nezhad and N. Langarib, "Phishing Detection Techniques: A review," *Data Science : Journal of Computing and Applied Informatics*, vol. V9, no. 1, Jan. 2025, doi: 10.32734/jocai.v9.i1-19904.

ABSTRACT

Phishing remains one of the most pervasive and sophisticated threats to cybersecurity, exploiting human and system vulnerabilities to compromise sensitive information. This study systematically reviews and categorizes phishing detection techniques into four groups: anti-phishing tools, heuristic approaches, machine learning-based techniques, and metaheuristic algorithms. Each method is critically analyzed for its effectiveness, highlighting their strengths and limitations. The review identifies significant advancements in phishing detection, such as the adoption of hybrid techniques and real-time detection algorithms, while also addressing gaps, including handling zero-day phishing attacks and scalability in large datasets. The findings provide a roadmap for future research, encouraging the development of more robust, adaptive, and efficient solutions. This comprehensive analysis not only synthesizes the state-of-the-art in phishing detection but also lays the groundwork for designing next-generation defense mechanisms.

Keyword: Phishing, Anti-Phishing Tools, Heuristic, Machine Learning, Meta Heuristic

ABSTRAK

Phishing tetap menjadi salah satu ancaman yang paling luas dan canggih terhadap keamanan siber, mengeksploitasi kerentanan manusia dan sistem untuk membobol informasi sensitif. Studi ini secara sistematis meninjau dan mengkategorikan teknik pendeteksian phishing ke dalam empat kelompok: alat anti-phishing, pendekatan heuristik, teknik berbasis pembelajaran mesin, dan algoritme metaheuristik. Setiap metode dianalisis secara kritis untuk mengetahui efektivitasnya, dengan menyoroti kekuatan dan keterbatasannya. Tinjauan ini mengidentifikasi kemajuan signifikan dalam deteksi phishing, seperti adopsi teknik hibrida dan algoritme deteksi waktu nyata, sekaligus mengatasi kesenjangan, termasuk menangani serangan phishing zero-day dan skalabilitas dalam kumpulan data yang besar. Temuan ini memberikan peta jalan untuk penelitian di masa depan, mendorong pengembangan solusi yang lebih kuat, adaptif, dan efisien. Analisis komprehensif ini tidak hanya mensintesis teknologi mutakhir dalam pendeteksian phishing, tetapi juga meletakkan dasar untuk merancang mekanisme pertahanan generasi berikutnya..

Keyword: Phising, Alat Anti-Penipuan, Heuristik, Machine Learning, Meta Heuristik



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International.
<http://doi.org/10.32734/jocai.v9.i1-19904>

1. Introduction

Now with the growth of the Internet, the geographic distances and time differences are fading and online business and financial transactions could be easily done. One of the biggest obstacles, which impedes the development of e-commerce, is security and ensures the safety of business via the web. One of the security threats is phishing attacks [7] that nowadays are a very important subject for cyber attackers. The phishing's target is the extracting user's personal information, while the user believes to deal with a legitimate company or organization, but in fact he/she is dealing with the person/s who are illegitimate and criminal. To prevent misuse of personal information and spoof from information about them, it is necessary that phishing should be given special attention. So, attention to the phishing for personal data protection and reduce the damage Caused

by it is important. Expansion and increasing sophistication of phishing attacks, Existence various methods to detect phishing and lack of comprehensive and updated overview research, the motivation was to compile this review. In only the present review article [8] to review heuristic methods, machine learning, visual similarity and blacklisted has been discussed and all available methods have not been studied. Therefore, in this paper phishing detection techniques have been analyzed in four groups (Fig. 1). To achieve this goal, the paper is organized as follows was written. Describes the phishing problem and evaluation criteria of phishing detection rate have been discussed in Section 2. In Section 3 to review, categorize, analyze and compare the techniques presented in four categories has been on the agenda, and finally the conclusions are discussed in Section 4.

2. Problem description

Phishing is defined as practical for obtaining users sensitive information (such as ID, password, credit card number) via a fake website that looks exactly like the legitimate website. Phishing detection problem can be divided into three sub-problem (phishing websites detection [9] , phishing emails detection[10] , Social Phishing[11]). In conjunction with phishing problem can be recounted objective function such as increased accuracy in phishing website detection [12] , reducing the users visiting of phishing website[13] , increased rapidly phishing websites detection[5] , the minimum of phishing websites [14] , to date detect phishing websites [15] , a low false positive rate [15] ,the rate of true positive and true negative rates above [15] ,The evaluation criteria for identifying phishing are visible in "Table 1".

The limits of phishing websites detection can note be short lifespan phishing websites [16] that According to research, the average lifespan of phishing websites is a few days or even a few hours. So, the data set should be updated and online. The following section reviews existing approaches are discussed.

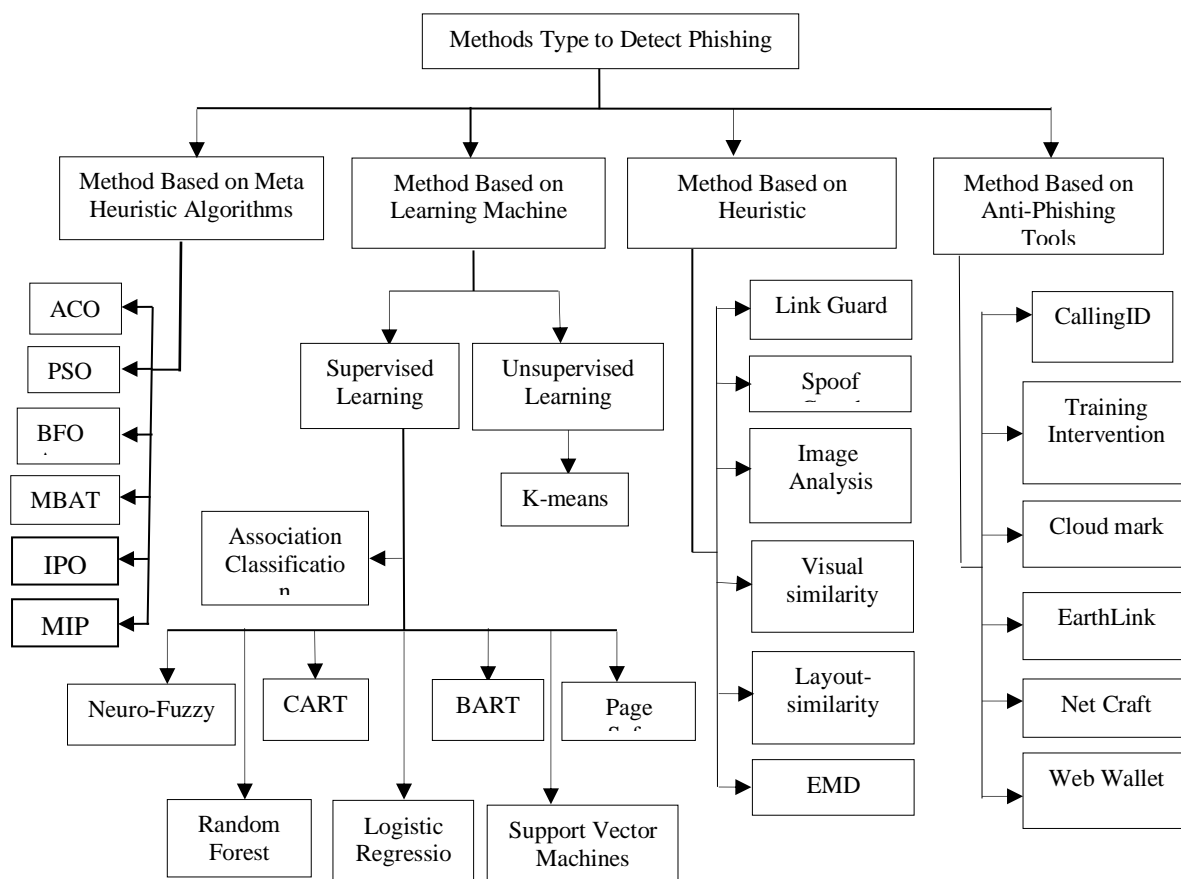


Figure 1. Methods type phishing detection

Table1. Phishing Detection Evaluation Criteria

Criteria	Phishing indicators
URL & Domain Identity [5, 12, 17, 18]	Using IP address Abnormal request URL Abnormal URL of anchor Abnormal DNS record Abnormal URL
Security & Encryption [5, 12, 17, 18]	Using SSL Certificate Certificate authority Abnormal cookie Distinguished names certificate
Source Code & Java script [5, 12, 17, 18]	Redirect pages Straddling attack Pharming attack On Mouse over to hide the Link Server Form Handler (SFH)
Page style & contents [5, 12, 17, 18]	Spelling Errors Copying website Using forms with Submit button Using pop-ups windows Disabling right-click
Web Address Bar [5, 12, 17, 18]	Long URL address Replacing similar char for URL Adding a prefix or suffix Using the @ Symbols to confuse Using hexadecimal char codes
Social human factor [5, 12, 17, 18]	Emphasis on security Public generic salutation Buying time to access accounts

3. Reviewed the proposed methods

Solutions that so far have been proposed to phishing websites detection can be classified and described in 5 groups (Figure 1). Continue to introduce these methods are discussed.

3.1 Method based on blacklist

The use of security tools anti phishing based on blacklists in browsers is a method for detecting phishing sites. These tools based on characteristics (such as a URL) where are applied them detect phishing sites and block the user's activities. Researchers have concluded that these tools alone are not effective for preventing phishing websites. Of the anti-phishing tools can point to the following Cases that in this review article have been studied.

3.1.1. Tools CallingID [19]: This tool is based on passive visual indicators. Change the indicator to green is indicating the legitimate site; to Yellow is indicating the suspicious site and to red color is indicates the phishing site. Some initiatives used this tool for phishing detection are reviewed site country of origin, duration of registration, site popularity and user reports and reviews have been blacklisted. CallingID toolbar runs in 98/NT/2000/XP Windows and Internet Explorer.

3.1.2. Tools Cloud mark [20]: when users visit the site, Cloud mark tools adapt it site with exist sites in blacklist, if availability it in blacklisted are displayed a warning to site as other methods. If it is not found in the black list, the site is assessed based on the feature popularity of the site. Points per site are calculated by gathering all the concessions be given to the site, the site that has a lower rating will be detected as fake site and blocked.

3.1.3. Tools Net craft [13]: This tool uses the blacklist to phishing site detection and if the site exists on the blacklist recommendation system will alert the user and the site will blocked. Net craft tool runs on most operating systems and in Internet Explorer are under Windows 2000 and XP.

3.1.4. Tools EARTHLINK [13] : This tool rely on combination of heuristic methods and manually and user ratings. EARTHLINK tool to detect phishing sites and adds to blacklist. Also review recorded information

such as the owner, and the age of the site. The indicator the tool is including three modes of green, red and yellow. EARTHLINK tool can be used in Internet Explorer and Firefox.

3.1.5. Tool Training Intervention [21] : This method uses of intervention messages to detect phishing website. This method gives information to end users and makes them aware of the mistakes that have been done to detect phishing sites. The main idea of this method is controlled a user, if the user Sensitive information published on the phishing website tool will display a message to users to help them to understand that website is phishing and how to detect it.

3.1.6. Tools GeoTrustWatch [19] : This tool also checks blacklist based certificate authentication, is identified phishing sites [19]. The indicator the tool is including three modes of green, red and yellow. This tool runs on Windows 98/NT/2000/XP and Internet Explorer.

Table 2. Advantage and disadvantage methods anti phishing tools based on blacklist

method	advantages	Disadvantages
Tool CallingID	No need to write any programming code	Low Accuracy and speed in detecting
Tool Cloud mark	Independent tool of browser	Low Detect Phishing Websites
Tool ARTHLINK	Good detection rates	Low speed in detection
Tool GeoTrustWatch	Easy implement	Low detection rates
Tool Net craft	Good detection rates	Depends on the particular Browser
Tool Web wallet	Reduced phishing attacks rates from 63% to 7%	False-positive rate in prediction
Tool B-APT	Ability to detect before they can be viewed items	High false negative rate
Tool Training Intervention	Capable to send of anti-phishing training to specific users via e-mail	low speed of proxy for Request Control.

3.2 Review methods based on heuristics

Another method to identify phishing is based on heuristic algorithms, heuristics be tested for Indicators such as obscure urls (hide real url destination), a strong visual similarity to the legitimate website, binary similarity to the legitimate website (including the discovery of obscure files), the code similarity to the legitimate website, using fake SSL certificates, records DNS suspected. continue will presented methods Based on heuristic.

3.2.1. Method based on Link Guard Algorithm [22] : a method to detect phishing sites is using Link Guard algorithms. Link Guard is based on detailed analysis of the characteristics of phishing hyperlinks. A hyperlink is structured as follows [22] :

` Anchor text <\ a>`.

This algorithm, hyperlinks are used in phishing sites are classified as follows [23].

1) Hyperlink in DNS domain names is provided in url text, but DNS name of the destination does not match with the actual link, such as the following hyperlink:

`
https://secure.regionset.com/EBanking/logon/ </ a>`

2) Decimal IP address is used directly in url instead of DNS name, as such as the following example:

` SIGN IN </ a>`

3) Hyperlink does not provide destination information in the url text and uses the DNS name in the url. DNS name in the url usually is similar with a popular company or organization. Like the hyperlink below:

` Click here to confirm your account </ a>`

Link Guard algorithm is character-based and has been implemented on Windows xp, experiments suggests that this method will consume little memory and in detecting phishing attacks with minimum false negative rate is very effective.

3.2.2. method based on visual similarity [3,4] : In this method, legitimate website is processed by the CPU module of legitimate website to obtain a moderate views and visual characteristics of the blocks. Discovery module a suspicious url, discovers suspicious url of emails. For any website with a suspicious URL, suspicious websites processor module, website with suspicious URL calls and displays. Evaluation module, the visual similarity between legitimate website with a fake website compares and calculates their visual similarity. If the visual similarity between legitimate and suspicious website is more than a threshold module reports website as phishing. Architecture method based on visual similarity is visible in "fig.2".

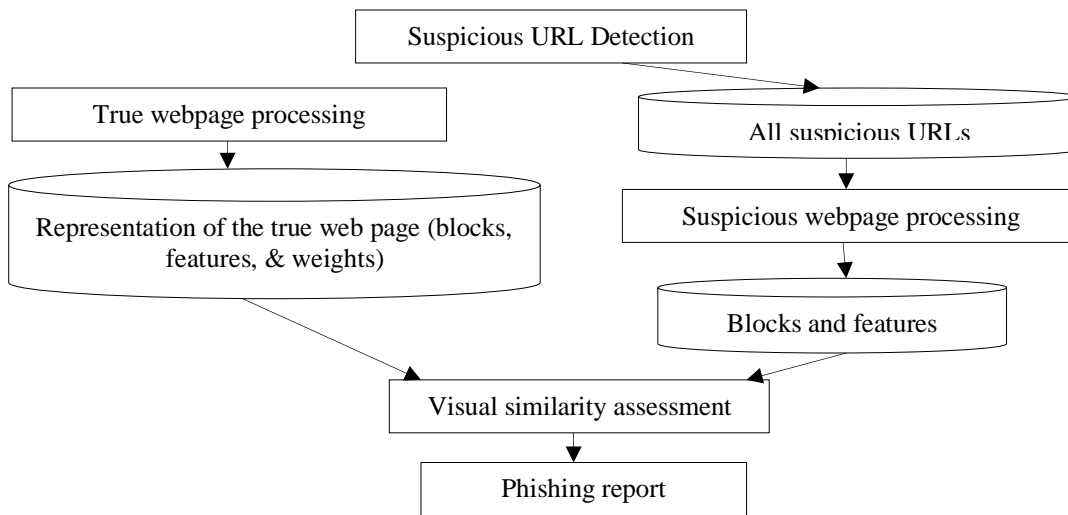


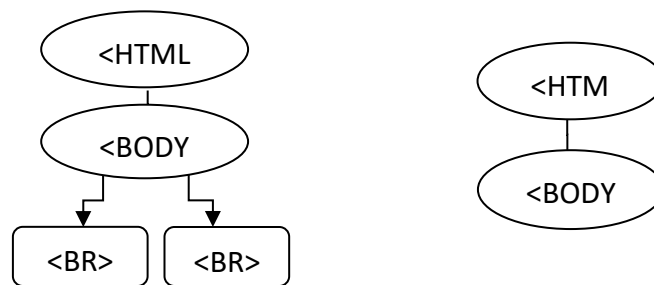
Figure2. system architecture of the proposed approach [3 ,4]

3.2.3. Method based on image analysis and site characteristics [24 ,23]: This method is relies on the collection and analysis of real-time URL posted on social media sites. Pages by Each URL will fetch and feature each page is measured to set of values as the number of images and links. This method also stored a page of provided images. Computes a hash function of image and use of hamming distance between these images is as a Compare apparent. A number of features marked include text Page title and number links, images, forms and labels are stored. In this method, the feature is intended as structural features of page. Image analysis is part of this work that can be done with set of fixed Dimension and quality settings for a page within a browser. Then a hash of the resulting images will be created and the hash values are compared using Hamming distance equation.

3.2.4. Method based on layout similarity [26 ,25]: This method can be calculated similarity between the current page and the page that is stored in the database. If layout similarity be more than the predefined threshold value, then the website is unreliable and a warning message is generated. Through Extraction DOM tree can be studied similarity of the two sites. DOM tree is an internal display that by browsers being used to display Web page. An example of DOM trees a legitimate website and a phishing website is visible in the figure below.

Table 3. Html code and DOM tree of phishing and legitimate website[25, 26]

Legitimate website	Phishing website
<HTML>	<HTML>
<BODY>	<BODY>
hello	 hello</BR>
</BODY>	</BODY>
</HTML	</HTML
Legitimate DOM tree	tree phishing DOM



3.2.5. Tools spoof Guard [27 ,13]: this tools the first is checking domain name and is compared with site that has been seen recently by the user to have stuck fake web site that has the same domain name .then for discover vague URL and non-standard port numbers is analyzed all the URLs. The next step is analysis of

page content, including passwords, images and embedded links. Tool Spoof Guard to analyze links on your web page using smart technologies or heuristic techniques. Finally, Will be examined the images on web pages. If the two images are identical, it is likely that fraudulent site, is copy image on the legitimate site.

3.2.6. EMD-based method with a linear programming model [29 ,28]: N.Revathy [2] of this method is used to detect phishing websites. The most important reason that Internet users fall victim to phishing attacks are similar to phishing pages and pages of legitimate website. EMD is a method to measure the distance (dissimilarity) between two signatures has been obtained. A signature is a set of features such as visual similarity of block designs; dominate colors, fonts, images. In this method, first are retrieved Suspicious and legitimate web pages, and then paid to the product signature for them. In this method, preprocessing of Web pages is a 3-step process: to obtain images of the web pages, normalization of images, display of Web page images as a visual signature that is includes the features of coordinates and color. In this method is used of the Graphics Device Interface (GDI) to obtain images of web page and save it as jpeg file. The architecture of this technique is visible in "Figure 3". If EMD amount is equal to 0, the two images are identical and if is equal to 1, the two image are completely different.

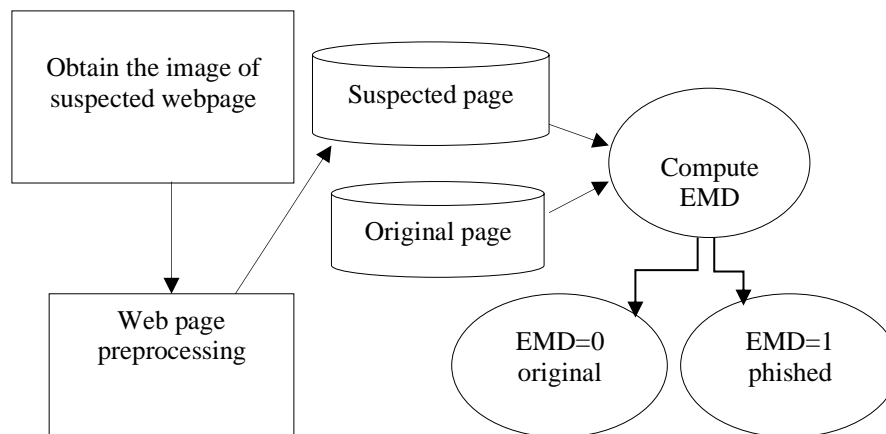


Figure 3. general architecture EMD [2]

Table 4. Advantage and disadvantage methods based on heuristic

method	advantage	disadvantage
link Guard	Low memory consumption, minimize the false negative rate	high false positive rate
method based on visual similarity	Intelligibility and simplicity	Speed time consuming
method based on image analysis and site characteristics	Low computational volume	Lack of proper performance when phishing web pages are different to real pages
method based on similarity layout	Creates problems for make phishing pages	Failure to properly detect when a phishing web page is similar to legal sites
Method based on EMD to linear programming model	Phishing classification with high accuracy	Failure to phishing detection at time different phishing web pages with the actual web page
spooof Guard	High speed in phishing detection	High false positive rate

3.3 Reviewing methods based on machine learning

Another method for phishing web sites detection is Machine Learning [31 ,30] that can be classified into two categories supervised learning and unsupervised learning that continues have been introduced.

3.3.1. Supervised learning: supervised learning methods can be noted to logistic regression, classification and regression trees, random forests, neural networks, support vector machines, Bayesian additive regression trees, Page Safe, logistic regression classification that continue has been investigated to the introduction of these methods in order to detect phishing [32].

3.3.2. association Classification[6 ,5] : Moh'd Iqbal AL Ajlouni is used of CBA and MCAR techniques to detect phishing websites in Internet banking and detect phishing sites based on 27 features that extracted from the site does . The architecture of this technique is visible in "Figure 4".

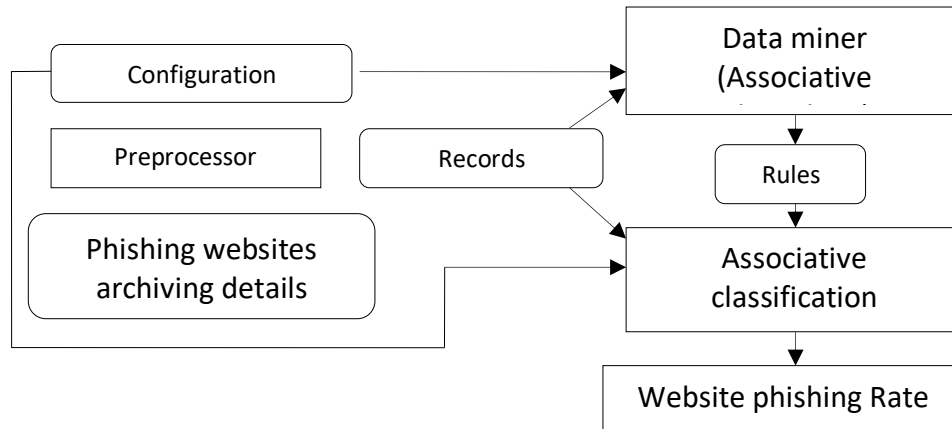


Figure 4. AC model for detecting phishing [5 ,6]

MCAR Model: MCAR model is the example of mining association rules, that right features of rules are represent classes. For example: $A, B \rightarrow Y$, where A, B are input feature and Y is output class. Output class features of rate internet banking phishing website is (phishing, Suspicious or legitimate).

3.3.2.1. Logistic regression [33] : this method largely used of a statistical model in many scope for the predict binary data. For Production organ of linear models, logistic regression commonly is used logic function, which is defined as follows:

$$\log \frac{p(x; \beta)}{1 - p(x; \beta)} = \beta^T \tag{1}$$

In equation (1), x is vector of P predicted $x = (x_1, x_2, \dots, x_p)$ and y is a binary response and β is a $p \times 1$ vector of regression parameters.

3.3.2.2. Support vector machines [34 ,35]: this method is a popular classifier that is being used today. The idea this method is to find optimum hyper plane (margin) between the two classes which by maximizing the margin between the together closest points of each class is done. Assumed to be which a linear discriminate function of two distinct classes with target value of +1 and -1 are there. A separating hyper plane is defined as \neg :

$$t_i = \begin{cases} 1 & w'x_i + w_0 \geq 0 \\ -1 & w'x_i + w_0 < 0 \end{cases} \tag{2}$$

Now distance every point x with the hyper plane is Equal to $|w'x_i + w_0|/\|w\|$ and the distance it to the origin is $|w_0|/\|w\|$.

3.3.2.3. Neuro-Fuzzy [18]: This method combines neural networks and fuzzy systems with 5 inputs, such as rules of legal sites, user behavior profile, Phish Tank, user-specific sites, Pop-Ups from the email (regular expressions are which by phisher used to appear on the screen) is to detect phishing in online transactions. Of 5 entries listed, 288 features are extracted which used as the training and testing data in the neural fuzzy systems to generate the fuzzy if- then rules and the expression of discrimination between phishing website and legal website in real-time Deals. If the website is a phishing, system offers a warning and process stops automatically. If a website is suspicious system produces red color with text-based descriptions of risk. The output of the neural fuzzy inference system by equation (3) is calculated.

$$y = \sum_{i=1}^n x_i^6 = \sum_{i=1}^n u_i [k_{i1} + k_{i2} \times x_1 + k_{i2} \times x_2] \tag{3}$$

3.3.2.4. Bayesian additive regression tree [34]: This method to discover unknown relationships between a continuous output Y and a p-dimensional vector of inputs $x = (x_1, \dots, x_p)$. Here assume that $Y = f(x) + \varepsilon$ and $\varepsilon \sim N(0, \sigma^2)$ is the random error. The main idea BART modeling or at least approximate $f(x)$ that takes place by the sum result of the regression tree:

$$f(x) = \sum_{i=1}^m g_i(x) \quad (4)$$

Each g_i refers to a binary regression tree with the desired structure, and when m be selected so large, value the low distribution in overall model as a weak learner is capable.

Table 5. Advantage and disadvantage methods based on machine learning

method	advantage	disadvantage
Logical regression	Simplicity and high ability to Interpretation	Loss of data from the database affects to the predicted rate
Classification and regression trees	The ability to explain the interaction between predicted \neg , the ability to interpret well	With the growth of a tree, considering the added impact is difficult
Random forests	Possibility of Setup a large number of variables in an educational set, the ability to Forecast missing nose	forest manufacturing process to form Random, the interpretation of the final model due to the large number of independent decision trees
Support vector machines	SVM is very powerful	High computational volume for train of data, sensitive to noise data
Bayesian additive regression tree	Does not require to the selection variables, generated tree automatically Convenient and quick response to inquiries, Minimum operating time	High false positive rate High Complexity
Neuro-Fuzzy	High accuracy in detecting phishing (98.5%)	High computational volume and much complexity
PageSafe	Remove a large percentage of phishing sites	false positive rate and false negative rates are relatively high
Logistic regression	High speed in phishing detection	high false negative rate Dependence of the final solution to the initial clusters, the lack of exist certain processes to calculate the initial clusters centers.
k-means	Low false positive rate	
Method association classification	High Precision and speed, flexible in large databases	High False positive rate

3.3.2.5. Method Page Safe [1] : This method performs an automated classification to detect phishing. Page Safe to examine the anomalies in web pages to perform automatic classification deals and decides about the legitimacy of a web page. Page Safe holds a white list (the list of domains with corresponding IP addresses), white list is encryption via a password to protect of malicious software. Page safe uses of artificial neural network for automatic classification after detected anomalies in a web site. Artificial neural network has the different architectures therefore require different types of algorithms. In this way, the algorithm Scaled Conjugate Gradient Back propagation (traincsg) has been used to test the neural network. This model is visible in "fig.5".

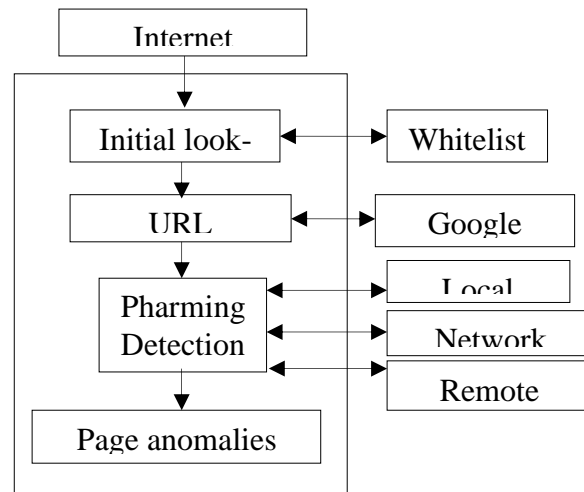


Figure 5. PageSafe model [1]

3.3.2.6. Classification and regression trees (CART) [36]: the model is which the distribution of conditional y with x to describe. The model consists of two components: a tree T with b output nodes and the vector $\Theta = (\theta_1, \theta_2, \dots, \theta_b)$ where θ_i is output node the i^{th} . If y is discrete, the model can be considered as a tree classification and if y is continuous, the model can be considered as tree regression. Maher Aburrous [12] used combination data mining algorithms and fuzzy systems in order to assess Internet banking that are in the risk of phishing websites and through extracting 27 feature detects phishing site. In this method is used of number classified techniques such as JRip, PART, Prism and C4.5 to is checked the different relationships between the characteristics of phishing.

3.3.2.7. Random forests[33 ,37]: random forests is a classifier which is combine of a large number of predictor trees, each tree independently is related to the values of a random vector sampled. In addition, all trees in the forest take advantage of the same distribution. Random forest can be setup with a large number of variables in a training set and the ability to forecast data is gone.

3.3.2.8. method based on logistic regression Classification[16 ,38]: In this method of heuristics used to detect phishing website and legitimate. These heuristics pay to model a linear regression classifier. In this method, the characteristics of phishing URL identified and is used the direction to model a logistic regression classifier. To train the model, blacklist and white list is used. To train blacklist, Google keeps a blacklist of URLs for protection Firefox users against phishing. The black list is updated by experts and will be removed each non-phishing URL from it.

3.3.3. Unsupervised learning: An unsupervised learning approach to phishing website detection is k-means algorithm which is introduced in the following.

3.3.3.1. Clustering with k-means[39]: in the k-means method start points are selected randomly, then the data according to the proximity similarity can be attributed to one of the clusters, and thus a new cluster is obtained. By repeating this procedure can in each repeat with Averaging of data calculating new centers and Again data attributed to new clusters. This process continues until the change in the data is not obtained. Under function is considered as the objective function.

$$J = \sum_{j=1}^k \sum_{i=1}^x \|x_i^{(j)} - c_j\|^2 \quad (5)$$

In the above equation $\| \|$ is Criterion the distance between points and C_j is the j^{th} cluster center.

Method based on Meta heuristic algorithms of meta heuristic algorithms that have been used to detect phishing website can be noted to algorithms Ant Colony Optimization (ACO) [40 ,41] , Particle Swarm Optimization (PSO) [40 ,41], Bacteria Foraging Algorithm (BFOA) [40], Modified Bat algorithm (MBAT)[40 ,42] and Inclined Planes Optimization Algorithm (IPO) [43, 44]. Ant colony algorithm (ACO) is inspired of studies and observations on colony of ants. Optimization with ACO algorithm is including restrictions such as random decisions sequence and non- Dependence between them and uncertain convergence time in phishing classification. To solve this problem, particle swarm optimization algorithm (PSO) that is inspired of social behavior of animals, such as a collective movement by birds and fish that have been used for detection of phishing sites, that Finds the best solution to optimize problem in the search space and detects phishing websites. However, PSO algorithm after the convergence cannot to increase found the accuracy of the found answer.

Bacteria Foraging optimization algorithm (BFO) is another method of optimization based on collective intelligence that has been used to detect phishing websites. BFO algorithm is inspired of way foraging a certain kind of flagellated bacteria in nature , this algorithm have been used to solve various engineering problems such as the harmonic approximation, training the neural networks and reduced losses in transmission lines .The main disadvantage of BFO algorithm in compare with other optimization methods is relatively slow convergence and Performance reduction of algorithm with increases the optimization problem Dimension (or alternatively, be larger the search space).

BAT algorithm has been developed in 2010 by Mr. Xin-She Yang. This algorithm is inspired of Reflection the voice of bats. Each bat is flying with a speed V_i randomly which is reached to position X_i or final solution. Frequency and wavelength of the sound of bats is Variable that during the search for Find hunting the frequency and wavelength changes. Bat algorithm for detecting phishing websites has less error rate in Compared with other optimization techniques, including ACO, PSO and BFOA.

Metaheuristic algorithms are optimization techniques inspired by natural phenomena, capable of solving complex problems by exploring large search spaces effectively. Their ability to optimize model parameters, select features, and identify patterns makes them particularly suitable for phishing detection. This section highlights key metaheuristic algorithms and their applications in the domain.

3.4.1 Genetic Algorithm (GA)

Genetic Algorithm (GA) is an evolutionary algorithm inspired by natural selection that has been employed for feature selection and rule optimization in phishing detection [45, 46]. GA iteratively applies crossover, mutation, and selection to evolve a population of candidate solutions toward an optimal feature subset or classification rules. For example, Kocyigit, E., et al. (2024) demonstrated the effectiveness of GA for optimizing URL-based features to enhance phishing detection accuracy [46]However, GA can suffer from premature convergence, especially in large and high-dimensional datasets.

3.4.2 Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO), inspired by the social behavior of bird flocks, has been widely applied for optimizing phishing detection models [47, 48]. PSO is particularly useful for feature selection and parameter tuning due to its simplicity and fast convergence. Aburrous et al. (2010) utilized PSO to optimize feature weights in a phishing detection model, achieving significant accuracy improvements [12]. Despite its advantages, PSO may get trapped in local optima when dealing with highly complex search spaces.

3.4.3 Ant Colony Optimization (ACO)

Ant Colony Optimization (ACO), inspired by the foraging behavior of ants, has been used for phishing detection, particularly in rule-based systems [49, 50]. ACO excels at solving combinatorial problems, such as constructing classification rules or identifying optimal feature subsets. Zareapoor et al. (2015) integrated ACO with machine learning algorithms for feature selection in phishing email classification, achieving better precision than standalone methods [51]. However, ACO can be computationally intensive, especially for large datasets.

3.4.4 Firefly Algorithm (FA)

The Firefly Algorithm (FA), inspired by the flashing behavior of fireflies, has shown potential in phishing detection for its ability to balance exploration and exploitation [52, 53]. FA has been successfully applied for optimizing classifiers and feature selection, as demonstrated in studies where it enhanced the detection rates of phishing websites by selecting relevant attributes from URL, HTML, and website content features.

3.4.5 Bat Algorithm (BA)

The Bat Algorithm (BA), based on the echolocation behavior of bats, has been explored for phishing detection due to its capability to navigate complex search spaces [54, 55]. In recent research, BA was used to optimize feature selection for machine learning classifiers, improving their accuracy and robustness against phishing attack.

3.4.6 Inclined Planes Optimization Algorithm (IPO)

Inclined Planes Optimization Algorithm (IPO), has been effectively applied to phishing detection tasks, demonstrating its capability to address complex classification challenges in this domain [43, 44]. Langhari and Abdolrazzagh-Nezhad (2015) utilized IPO for detecting phishing websites in e-banking, optimizing the feature selection process and improving detection accuracy by efficiently exploring and exploiting the solution space. The algorithm simulates the motion of objects on inclined planes, dynamically adjusting the inclination to balance global exploration and local exploitation, which is particularly effective in identifying the subtle patterns in phishing datasets. Furthermore, in a subsequent study, Abdolrazzagh-Nezhad (2017) incorporated fuzzy rules into a modified version of IPO, enhancing its ability to classify websites and detect phishing threats with high precision. This approach demonstrated superior performance in terms of accuracy and computational

efficiency compared to traditional metaheuristic algorithms, making IPO a promising tool for addressing the growing complexity of phishing detection.

3.4.7 Challenges and Future Directions

While metaheuristic algorithms offer promising results, they are not without limitations. Common challenges include computational overhead, slow convergence in highly complex search spaces, and sensitivity to parameter settings. Addressing these challenges requires:

- Developing hybrid approaches that leverage the strengths of multiple algorithms.
- Incorporating adaptive mechanisms to dynamically adjust parameters based on data complexity.
- Enhancing scalability to handle the growing volume of phishing attacks in real-time applications.

3.5 Comparison and Evaluation

In this section describe methods for detection of phishing websites is evaluated in one parameter Accuracy rates which in Table 6 are given these results.

Table 6. Compare between phishing detection methods

method	Number phishing website for evaluated	Detection accuracy rate
CallingID	100	50%
Cloud mark	100	50%
Net craft	100	75%
EARTHLINK	100	75%
Training Intervention	400	89%
GeoTrustWatch	100	60%
	C4.5	84.2%
	P.A.R.T	86.3%
Fuzzy	JRip	81.3%
Data Mining	R.I.P.E.R	84.9%
	PRISM	87.3%
	CBA	88.6%
)MCAR(association classification	1006	88.6%
Link Guard	210	96%
Visual similarity	140	93%
Image analysis and site characteristics	2.8 million	90%
Layout similarity	500	93%
spoof Guard	100	96%
EMD with Linear programming model	1500	95%
Logistic regression	1000	88.59%
Support Vector Machine	1000	87.07%
Neuro-Fuzzy	200	98.5%
Bayesian additive regression trees	1000	87.09%
PageSafe	200	97%
Classification and regression trees	1000	89.59%
Random Forest	1000	90.24%
Logistic regression classifier	1245	88%
k-mean	500	90.07%
ACO	1006	89%
PSO	1006	92%
BFOA	1006	97%
MBAT	1006	98%
IPO	1006	96.7%
MIPO	1006	97.3%

4. Discussion and Conclusion

The study provides a comprehensive analysis of phishing detection methods, categorizing them into five distinct groups and evaluating their applicability across different use cases. Each method offers unique strengths but also presents specific challenges that highlight areas for future improvement.

Strengths and Contributions of Existing Methods

- **Anti-Phishing Tools:** These tools, such as Netcraft and Cloudmark, provide user-friendly interfaces and straightforward implementation, making them accessible to non-technical users. However, they are heavily reliant on pre-existing blacklists, which limits their ability to detect zero-day phishing attacks. Their efficacy could be enhanced by integrating real-time heuristic or machine learning components to detect novel phishing techniques dynamically.
- **Heuristic Methods:** Heuristic approaches offer simplicity and speed, analyzing visual, structural, and URL-based features. Techniques like LinkGuard and visual similarity assessment have shown high accuracy for detecting phishing websites. However, their reliance on predefined rules makes them less effective against obfuscated or rapidly evolving phishing techniques. Incorporating adaptive rule-generation mechanisms could make these methods more robust.
- **Machine Learning Techniques:** Machine learning-based methods, including Random Forests and Neural Networks, excel in processing complex datasets and identifying nuanced patterns. While these methods demonstrate high accuracy, they often require extensive labeled training data, which may not always be available. Furthermore, their computational complexity poses challenges for real-time detection scenarios. The integration of semi-supervised learning and transfer learning techniques could address these limitations.
- **Metaheuristic Algorithms:** Metaheuristic methods, such as PSO and ACO, are highly effective for optimizing phishing detection models. Their ability to navigate large search spaces makes them particularly suitable for feature selection and model optimization. However, these methods may suffer from slow convergence and high computational costs, especially for large datasets. Hybridizing metaheuristic algorithms with machine learning techniques can balance efficiency and accuracy.

Challenges and Research Gaps

- **Handling Zero-Day Phishing Attacks:** Most existing methods rely on predefined datasets or rules, making them less effective for detecting novel phishing techniques. Developing adaptive algorithms capable of learning from real-time data streams is essential.
- **Scalability and Real-Time Performance:** As phishing attacks continue to grow in volume and complexity, scalability becomes a critical factor. Future methods must prioritize lightweight, computationally efficient models to ensure applicability in real-world scenarios.
- **Class Imbalance:** Phishing datasets often exhibit significant class imbalances, where legitimate samples vastly outnumber phishing samples. This imbalance can skew model performance, necessitating the development of techniques that maintain accuracy across both classes.
- **User-Centric Design:** Incorporating user behavior and feedback into detection mechanisms can enhance the practicality and adaptability of phishing detection tools. Methods that combine human and machine intelligence offer promising avenues for development.

Also, the existing methods were used of dataset (Anti-Phishing Working Group, 2007) [[56] and (http://www.phishtank.com/phish_archive.php 2008) [57] which are includes url lists of phishing web sites for testing. This review synthesizes the current state-of-the-art in phishing detection, offering valuable insights into the strengths and limitations of existing methods. While significant progress has been made, particularly in the development of hybrid techniques and advanced machine learning models, critical challenges remain. Addressing these challenges requires a multi-faceted approach that integrates adaptive, scalable, and user-centric solutions. The findings of this study provide a roadmap for future research, emphasizing the need for:

- Real-time detection systems that leverage adaptive learning mechanisms.
- Hybrid frameworks combining the strengths of multiple techniques, such as metaheuristics and machine learning.
- Scalable solutions that can handle the increasing volume and complexity of phishing attacks.

By addressing these areas, researchers and practitioners can contribute to the development of next-generation phishing detection mechanisms, ultimately enhancing cybersecurity and protecting users from evolving threats.

References

- [1] P. Sengar and V. Kumar, "Client-side defense against phishing with pagesafe," *International Journal of Computer Applications*, vol. 4, no. 4, pp. 6-10, 2010.
- [2] N. R. T. Guhan "Analyzing And Detecting Phishing Webpages Withvisual Similarity Assessment Based On Earth Mover's Distance With Linear Programming Model," *International Journal of Advanced Engineering Technology Research* vol. Vol.III, pp. 327-330, 2012.
- [3] E. Medvet, E. Kirda, and C. Kruegel, "Visual-similarity-based phishing detection," in *Proceedings of the 4th international conference on Security and privacy in communication netowrks*, 2008: ACM, p. 22.
- [4] W. Zhang, H. Lu, B. Xu, and H. Yang, "Web phishing detection based on page spatial layout similarity," *Informatica*, vol. 37, no. 3, pp. 231-244, 2013.
- [5] M. Aburrous, M. Hossain, K. Dahal, and F. Thabtah, "Associative classification techniques for predicting e-banking phishing websites," in *Multimedia Computing and Information Technology (MCIT), 2010 International Conference on*, 2010: IEEE, pp. 9-12.
- [6] M. d. I. A. Ajlouni, W. e. Hadi, and J. Alwedyan, "Detecting Phishing Websites Using Associative Classification," *European Journal of Business and Management*, vol. 5, no. 15, pp. 36-40, 2013.
- [7] K. Jansson and R. Von Solms, "Phishing for phishing awareness," *Behaviour & Information Technology*, vol. 32, no. 6, pp. 584-593, 2013.
- [8] M. Khonji, Y. Iraqi, and A. Jones, "Phishing Detection: A Literature Survey," 2013.
- [9] R. Dhanalakshmi, C. Prabhu, and C. Chellapan, "Detection of phishing websites and secure transactions," *International Journal Communication & Network Security (IJCNS)*. v1, pp. 15-21, 2011.
- [10] A. Bergholz, J. De Beer, S. Glahn, M.-F. Moens, G. Paaß, and S. Strobel, "New filtering approaches for phishing email," *Journal of computer security*, vol. 18, no. 1, pp. 7-35, 2010.
- [11] I. R. A. Hamid and J. H. Abawajy, "Profiling Phishing Email Based on Clustering Approach," in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on*, 2013: IEEE, pp. 628-635.
- [12] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," *Expert systems with applications*, vol. 37, no. 12, pp. 7913-7921, 2010.
- [13] Y. Zhang, S. Egelman, L. Cranor, and J. Hong, "Phinding phish: Evaluating anti-phishing tools," 2006: ISOC.
- [14] M. Wu, R. C. Miller, and G. Little, "Web wallet: preventing phishing attacks by revealing user intentions," in *Proceedings of the second symposium on Usable privacy and security*, 2006: ACM, pp. 102-113.
- [15] M. Blasi, "Techniques for detecting zero day phishing websites," Iowa State University, 2009.
- [16] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in *Proceedings of the 2007 ACM workshop on Recurring malware*, 2007: ACM, pp. 1-8.
- [17] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabatah, "Modelling Intelligent Phishing Detection System for e-Banking using Fuzzy Data Mining," in *CyberWorlds, 2009. CW'09. International Conference on*, 2009: IEEE, pp. 265-272.
- [18] P. Barraclough, M. Hossain, M. Tahir, G. Sexton, and N. Aslam, "Intelligent phishing detection and protection scheme for online transactions," *Expert Systems with Applications*, 2013.
- [19] P. Malathi and P. Vivekanandan, "An Efficient Framewo."
- [20] L. F. Cranor, S. Egelman, J. I. Hong, and Y. Zhang, "Phinding Phish: An Evaluation of Anti-Phishing Toolbars," in *NDSS*, 2007.
- [21] A. Alnajim and M. Munro, "An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection," in *Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on*, 2009: IEEE, pp. 405-410.
- [22] S. T. Kumar, V. Kumar, and A. Kumar, "Detection and Prevention of Phishing Attacks Using Linkguard Algorithm," 2008.
- [23] J. N. M. Joshua S. White, John L. Stacy, "A Method For The Automated Detection Of Phishing Websites Through Both Site Characteristics And Image Analysis," 2011.
- [24] J. Chhikara, R. Dahiya, N. Garg, and M. Rani, "Phishing & Anti-Phishing Techniques: Case Study," *International Journal*, vol. 3, no. 5, 2013.
- [25] A. P. Rosiello, E. Kirda, and F. Ferrandi, "A layout-similarity-based approach for detecting phishing pages," in *Security and Privacy in Communications Networks and the Workshops, 2007. SecureComm 2007. Third International Conference on*, 2007: IEEE, pp. 454-463.

- [26] J. Mao, P. Li, K. Li, T. Wei, and Z. Liang, "BaitAlarm: Detecting Phishing Sites Using Similarity in Fundamental Visual Features," in *Intelligent Networking and Collaborative Systems (INCoS), 2013 5th International Conference on*, 2013: IEEE, pp. 790-795.
- [27] M. Shrivastava, R. Sinha, and B. Shukla, "Panchâ ¼Vaktram (A Web Browser with a Spoof Guard Technology)," in *International Conference on Computer Technology and Development, 3rd (ICCTD 2011)*, 2011: ASME Press.
- [28] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD)," *Dependable and Secure Computing, IEEE Transactions on*, vol. 3, no. 4, pp. 301-311, 2006.
- [29] E. H. Chang, K. L. Chiew, S. N. Sze, and W. K. Tiong, "Phishing Detection via Identification of Website Identity," in *IT Convergence and Security (ICITCS), 2013 International Conference on*, 2013: IEEE, pp. 1-4.
- [30] T. Pitakrat, A. van Hoorn, and L. Grunske, "A comparison of machine learning algorithms for proactive hard disk drive failure detection," in *Proceedings of the 4th international ACM Sigsoft symposium on Architecting critical systems*, 2013: ACM, pp. 1-10.
- [31] A. Kalybayev, "Comparative study of machine learning algorithms in website phishing detection," Universiti Teknologi Malaysia, Faculty of Computing, 2013.
- [32] A. Khade and S. K. Shinde, "Detection of Phishing Websites Using Data Mining Techniques," in *International Journal of Engineering Research and Technology*, 2014, vol. 2, no. 12 (December-2013): ESRSA Publications.
- [33] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, 2007: ACM, pp. 60-69.
- [34] J. M. De Sa, *Pattern recognition: concepts, methods, and applications*. Springer, 2001.
- [35] H. M. Deylami and Y. P. Singh, "Cybercrime detection techniques based on support vector machines," *Artificial Intelligence Research*, vol. 2, no. 1, 2013.
- [36] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R news*, vol. 2, no. 3, pp. 18-22, 2002.
- [37] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [38] A. DeMaris and S. H. Selman, "Logistic regression," in *Converting Data into Evidence*: Springer, 2013, pp. 115-136.
- [39] R. Basnet, S. Mukkamala, and A. H. Sung, "Detection of phishing attacks: A machine learning approach," in *Soft Computing Applications in Industry*: Springer, 2008, pp. 373-383.
- [40] D. M. L. V. Radha Damodaram, "Experimental Study on Meta Heuristic Optimization Algorithms for Fake Website Detection " *International Association of Scientific Innovation and Research (IASIR)* vol. 2 pp. 43-53 2012.
- [41] M. Radha Damodaram and M. Valarmathi, "Phishing Website Detection and Optimization Using Particle Swarm Optimization Technique," *International Journal of Computer Science and Security (IJCSS)*, vol. 5, no. 5, p. 477, 2011.
- [42] M. Radha Damodaram and M. Valarmathi, "Bacterial Foraging Optimization for Fake Website Detection," *International Journal of Computer Science & Applications (TIJCSA)*, vol. 1, no. 11, 2013.
- [43] N. Langhari and M. Abdolrazzagh Nejad, "Phishing website detection for e-banking by inclined planes optimization algorithm," *Electronic and Cyber Defense*, vol. 3, no. 1, pp. 29-39, 2015.
- [44] M. Abdolrazzagh-Nezhad, "Classification and phishing websites detection by fuzzy rules and modified inclined planes optimization," *Nashriyyah-i Muhandisi-i Barq va Muhandisi-i Kampyutar-i Iran*, vol. 52, no. 4, p. 311, 2017.
- [45] W. Ali and A. A. Ahmed, "Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting," *IET Information Security*, vol. 13, no. 6, pp. 659-669, 2019.
- [46] E. Kocyigit, M. Korkmaz, O. K. Sahingoz, and B. Diri, "Enhanced Feature Selection Using Genetic Algorithm for Machine-Learning-Based Phishing URL Detection," *Applied Sciences*, vol. 14, no. 14, p. 6081, 2024.
- [47] W. Ali and S. Malebary, "Particle swarm optimization-based feature weighting for improving intelligent phishing website detection," *IEEE Access*, vol. 8, pp. 116766-116780, 2020.
- [48] S. M. Alshahrani, N. A. Khan, J. Almalki, and W. Al Shehri, "URL Phishing Detection Using Particle Swarm Optimization and Data Mining," *Computers, Materials & Continua*, vol. 73, no. 3, 2022.

- [49] M. M. Elsheh and K. Swayeb, "Phishing Website Detection Using a Hybrid Approach Based on Support Vector Machine and Ant Colony Optimization," in *2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, 2023: IEEE, pp. 402-406.
- [50] R. K. V. Penmatsa and P. Kakarlapudi, "Web phishing detection: feature selection using rough sets and ant colony optimisation," *International Journal of Intelligent Systems Design and Computing*, vol. 2, no. 2, pp. 102-113, 2018.
- [51] M. Zareapoor and K. Seeja, "Feature extraction or feature selection for text classification: A case study on phishing email detection," *International Journal of Information Engineering and Electronic Business*, vol. 7, no. 2, p. 60, 2015.
- [52] M. Hamed and J. Soyemi, "Classification Of Phishing Attacks In Social Media Using Associative Rule Mining Augmented With Firefly Algorithm," *GPH-International Journal Of Computer Science and Engineering*, vol. 6, no. 06, pp. 01-10, 2023.
- [53] O. A. Adewumi and A. A. Akinyelu, "A hybrid firefly and support vector machine classifier for phishing email detection," *Kybernetes*, vol. 45, no. 6, pp. 977-994, 2016.
- [54] M. Radha Damodaram and M. Valarmathi, "Phishing website detection and optimization using Modified bat algorithm."
- [55] V. Devika, A. Thushara, and M. J. Pillai, "Classification of phishing websites using extreme learning machine and hybrid bat algorithm," *Journal of Innovation in Computer Science and Engineering*, vol. 10, no. 1, pp. 23-29, 2020.
- [56] A.-P. W. Group. <http://www.Antiphishing.org> (accessed).
- [57] R. a. P. website. "phishing." <http://www.phishtank.com> (accessed).