Transactions on Machine Intelligence 7(1) (2024) 38-50

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/index.php/TMI

Detection of Phishing Website Attacks in Electronic Banking Using a Principal Component Analysis Algorithm and Multi-Layer Perceptron Neural Network Algorithm

M. Ghasemi [1,*]

[1] Masters student, Department of Computer Engineering, Faculty of Technology and Engineering, Yasouj branch, Islamic Azad University, Yasouj, Iran

ARTICLE INFO

Article History:
Received 20 November 2023
Received in revised form 1 February 2024
Accepted 18 March 2024
Available online 21 March 2024

Keywords:
Detection, Phishing Website Attacks, Electronic Banking, Dimensionality Reduction Algorithm, Multi-Layer Perceptron Neural Network Algorithm

ABSTRACT

Phishing, commonly known as the unauthorized acquisition of personal information from users of online platforms and clients of digital stores and financial institutions, has witnessed a notable surge in recent years. This surge has fueled the growth of a thriving criminal enterprise, particularly targeting financial service providers. Given the magnitude of this threat, we adopted a dual approach involving the application of a Principal Component Analysis (PCA) algorithm and a multi-layer perceptron neural network algorithm to identify and combat phishing attacks within the realm of electronic banking. Initially, we employed the PCA algorithm to streamline the identification process, reducing the number of features from an initial 30 to a more manageable 14. Following this feature reduction step, we fine-tuned the accuracy of detecting phishing website attacks using the multi-layer perceptron neural network algorithm. This algorithm, functioning as a binary classification technique, adeptly determines whether an input vector belongs to a specific class. Acting as a linear classifier, it relies on the weighted linear combination of input factors to make predictions. To further fortify our defenses, we implemented the Waka tool, an online algorithm capable of meticulously examining individual inputs. Through the strategic integration of the PCA and multi-layer perceptron neural network algorithms, we achieved a substantial enhancement in the accuracy of detecting phishing website attacks in the electronic banking domain, reaching an impressive 91.64%.

1. INTRODUCTION

Phishing is a form of social engineering attack aimed at tricking individuals into disclosing sensitive information, like credit card details, usernames, and passwords, or installing harmful software, such as ransomware [1]. The main aim of phishing is to acquire information related to online stores, such as usernames, passwords, and bank account information, by creating counterfeit websites [2]. Phishing lures users into exposing their passwords through deceitful means. The primary concept behind a phishing attack is to entice the user with a hook in hope that they

* Corresponding Author: Majid1355@gmail.com
Masters student, Department of Computer Engineering, Faculty of Technology and Engineering, Yasouj branch, Islamic Azad University, Yasouj, Iran

http://dx.doi.org/10.47176/TMI.2024.38

will take the bait and be caught, just like a fish. This bait usually manifests as an email, website, or short message, which then redirects the user to a phishing website. Uninformed internet users are particularly susceptible to this sort of deceptive tactic [3].

Phishing has long been recognized as a major cybersecurity threat. Its initial conceptualization can be traced back to 1987, with the first documented phishing attack occurring in 1995 [4-5].

The widespread emergence of phishing began around 2005. According to the Anti-Phishing Working Group (APWG), 163,333 phishing attacks were recorded in the third quarter of 2014 a slight 5% decrease from the previous quarter [6]. By 2018, Microsoft reported a 250% increase in phishing incidents, with over 470 billion malicious emails circulating globally [7]. The FBI's 2020 cybercrime report identified phishing as the most prevalent form of cyberattack, accounting for 32.35% of all reported cases 241,342 incidents marking a more than tenfold increase since 2015 [8].

Cybersecurity Ventures has forecasted that by 2021, global cybercrime damages would reach $6 trillion annually, with phishing contributing significantly [9]. A 2021 report by Barracuda Networks, based on data from 17,000 organizations, revealed over 12 million phishing and social engineering attempts targeting 3 million email accounts [10]. Approximately 43% of these attacks involved spoofing Microsoft services [11]. Notably, more than 80% of phishing attempts were directed at employees outside of finance and executive roles, with CEOs and IT staff experiencing 57 and 40 targeted attacks per year, respectively. Furthermore, phishing attacks related to cryptocurrency platforms rose by 192% from October 2020 to April 2021, an increase largely attributed to the rising value of digital assets [12]. In early 2022, phishing attacks exceeded one million cases in a single quarter for the first time [13].

This study proposes a hybrid detection model that integrates Principal Component Analysis (PCA) with a Multi-Layer Perceptron (MLP) neural network to enhance phishing website detection in electronic banking. The model was implemented using the WEKA data mining tool. By combining dimensionality reduction with supervised learning, the proposed approach improves detection accuracy and response effectiveness.

The article is structured into six sections: Introduction and background, PCA methodology, MLP architecture, implementation details, experimental results, and conclusion.

2. RESEARCH BACKGROUND

In this section, we have discussed the previous articles for the accuracy of detection of phishing website attacks.

2.1. Detection of phishing websites using logistic model tree algorithm

The study utilizes a dataset containing 1,353 samples and 9 features to evaluate the accuracy of phishing website detection. The classification model employed is the Logistic Model Tree (LMT), which integrates the structure of a standard decision tree with logistic regression models at its leaf nodes. This hybrid approach is conceptually similar to regression trees, where regression functions are used at the terminal nodes. In the LMT model, each internal node performs a univariate test based on one of the features, thereby enabling efficient decision-making through a hierarchical structure. To implement and assess the model's performance, the WEKA data mining software was employed. Experimental results demonstrated a classification accuracy of 89.357%, indicating that the Logistic Model Tree is a promising method for phishing website detection in terms of both interpretability and predictive performance.

2.2. Detection of phishing websites using functional tree algorithm

In this study, a dataset consisting of 1,353 samples and 9 features was utilized to evaluate the accuracy of phishing website detection. The classification model applied is the Functional Tree (FT) algorithm, which represents a flexible and generalized learning framework. Unlike traditional decision trees, FT models support multivariate classification and regression, allowing for the use of combinations of features in both decision nodes and leaf nodes. This enables the construction of more expressive and adaptive models. The implementation and evaluation of the model were

carried out using WEKA, a well-known data mining platform. The results demonstrated that the FT algorithm achieved a detection accuracy of 88.9135%, indicating its effectiveness in identifying phishing websites within the given dataset.

2.3. Detection of phishing websites using decision tree j48 algorithm

This study investigates the performance of phishing website detection by analyzing a dataset comprising 1,353 samples and 9 distinct features. All technical terms and abbreviations are defined upon first appearance to ensure clarity. The study strictly adheres to academic standards, including appropriate formatting, citation practices, and the use of objective, unbiased language. Where subjective interpretations are presented, they are explicitly identified. The J48 decision tree algorithm, an enhanced version of the ID3 algorithm, was employed for classification. J48 offers several improvements over its predecessor, including the ability to handle missing values, perform pruning to avoid overfitting, manage continuous attributes, and derive classification rules from the decision tree structure. The model implementation and evaluation were conducted using the WEKA data mining platform. Experimental results demonstrated that the J48 algorithm achieved a classification accuracy of 91.569% in detecting phishing website attacks, highlighting its effectiveness for this application.

3. PRINCIPAL COMPONENT ANALYSIS ALGORITHM

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique that helps in identifying the underlying patterns in high-dimensional data by transforming it into a smaller set of uncorrelated components. In the context of fault diagnosis, PCA has proven to be effective in reducing the complexity of feature sets, making it easier to detect and analyze critical information. For example, Haddadnia, Seryasat, and Rabiee (2013) applied PCA in the diagnosis of thyroid diseases by reducing the data dimensions and enhancing the model's accuracy when combined with a probabilistic neural network [14]. Similarly, Seryasat et al. (2013) utilized PCA for the fault diagnosis of ball bearings, where it was paired with a support vector machine to improve the diagnostic performance[15]. By extracting the most significant features from vibration signals, PCA assists in improving the accuracy and efficiency of predictive models in both medical and mechanical applications, highlighting its versatility as a preprocessing technique in various domains.

Machine learning algorithms rely on input variables, commonly referred to as features or factors, to perform classification tasks. However, when the number of features becomes excessively large, it can hinder the interpretability and visualization of the training dataset. Moreover, the presence of redundant or irrelevant features may introduce noise, complicate model training, and reduce overall performance. To address these challenges, dimensionality reduction techniques are often employed to eliminate redundancy, enhance computational efficiency, and improve the model's generalization capabilities [16].

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique in machine learning and data mining. It helps address the problem of high-dimensional data by transforming it into a lower-dimensional representation, while preserving as much of the original data variance as possible. PCA can be approached through two primary strategies [17]:

3.1. Feature Selection

Feature selection involves identifying a subset of the original variables that are most relevant to the learning task. The goal is to retain informative features while eliminating redundant or irrelevant ones. There are three commonly used methods for feature selection:

- Filter methods, which evaluate feature relevance based on statistical properties of the data (e.g., correlation, mutual information).

- Wrapper methods, which assess feature subsets by training and evaluating a model on different combinations.

- Embedded methods, which perform feature selection during model training, such as decision tree algorithms or Lasso regression.

3.2. Feature Extraction

Unlike feature selection, feature extraction creates new variables by transforming or combining the original features into a new feature space. The objective is to project data from a high-dimensional space to a lower-dimensional one, while preserving the intrinsic structure. PCA accomplishes this by identifying the directions (principal components) along which data variance is maximized, and projecting the data onto those directions.

4. MULTILAYER PERCEPTRON NEURAL NETWORK

The Perceptron algorithm, introduced in 1957 by Frank Rosenblatt at the Cornell Aeronautical Laboratory, was among the earliest artificial neural networks developed for binary classification tasks [18]. It operates by computing a weighted sum of the input features and applying a threshold-based activation function. The original perceptron is a linear classifier and is typically trained using an online learning algorithm, which updates weights sequentially for each training instance.

A more advanced form of this model is the Multilayer Perceptron (MLP), which is a class of feedforward artificial neural networks. An MLP consists of at least three layers of nodes:

- An input layer that receives the features,

- One or more hidden layers that perform intermediate computations,

- An output layer that produces the final classification or regression output.

Each node, or artificial neuron, in these layers acts as a computational unit that processes the weighted input from the previous layer and applies an activation function to produce its output. The outputs of one layer become the inputs for the next, forming a network capable of learning complex, nonlinear mappings.

4.1. Feedforward Process in MLP

The feedforward algorithm is the first stage in training a multilayer perceptron. It defines how input data propagate through the network layer by layer, ultimately producing an output. Unlike the single-layer perceptron, where computations are relatively simple, the feedforward process in an MLP is more intricate due to the presence of multiple hidden layers.

During feedforward propagation, the output of each layer is computed and passed to the next layer as input. These computations involve weighted sums of inputs and the application of activation functions, and they continue until the final output layer produces the prediction. This entire process can be mathematically formalized through a set of equations that define how neuron outputs are calculated at each layer.

4.2. Feedforward Activation Function

For each input feature vector $\mathbf{I} = (I_1, I_2, \ldots, I_n)$, the output y_i of a neuron is computed using a weighted sum of the inputs plus a bias term, followed by a nonlinear activation function:

$$y_i = f\left(\sum_{i=1}^{n} w_{ij} I_i + \text{Bias}_i\right) = f\left(\mathbf{w}^T \mathbf{I} + \text{Bias}_i\right) \tag{1}$$

Several activation functions are commonly used in practice, including:

- **Hyperbolic tangent (tanh)**:

$$f(\mathbf{w}^T \mathbf{I} + \text{Bias}) = \tanh(\mathbf{w}^T \mathbf{I} + \text{Bias}) \tag{2}$$

- **Sigmoid function**:

$$f(\mathbf{w}^T \mathbf{I} + \text{Bias}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{I} + \text{Bias})}} \tag{3}$$

These functions introduce non-linearity, enabling the network to learn complex patterns in data.

4.3. Calculation of Network Error Using the Loss Function

Once a training input is propagated through the network and a prediction is made, the result is compared with the known output (target). The final layer's computed value is referred to as the predicted output, while the actual label provided in the dataset is called the expected output.

In supervised learning, the loss function quantifies the difference between these two values. The result, known as the loss value, is critical in evaluating the network's performance. This value is subsequently used during backpropagation to adjust the weights of the network to minimize prediction errors.

4.4. Backpropagation Algorithm

The backpropagation algorithm is the central learning rule in training multilayer perceptron neural networks. Once the prediction error is computed, the error is propagated backward from the output layer through the hidden layers toward the input layer. This process adjusts the weights to minimize the loss using the gradient descent optimization method.

Step 1: Forward Pass Output

Given an input vector with n features and a network with J hidden layers, the output of each node in the network can be computed as:

$$y_j = f\left(\sum_{i=1}^{n} w_{ji} y_i\right) \tag{4}$$

Where y_j is the output of node j, w_{ji} are the weights connecting the layers, and f is the activation function.

Step 2: Error Calculation at Output Layer

The delta error δ_i for each node i in the output layer is calculated as follows:

$$\delta_i = (y_{predicted_i} - y_{expected_i}) \cdot y_{expected_i} \cdot (1 - y_{expected_i}) \quad (5)$$

Where:

- $y_{expected\ i}$ is the expected output for node *i*,
- $y_{predicted\ i}$ is the predicted output for node *i*

Step 3: Error Calculation at Hidden Layer

For each node *j* in a hidden layer, the delta value is computed using:

$$\delta_j = O_j(1 - O_j)\sum_i w_{ji}\delta_i \quad (6)$$

Where:

- O_j is the output of node *j*,
- w_{ji} are the weights from node *j* to each output node *i*,
- δ_i is the delta from the next layer,

Step 4: Weight Update Using Gradient Descent

Weights are updated using the **gradient descent** method. For adjusting weights from the input layer to a hidden layer:

$$\Delta w_{nj} = \eta \delta_j x_n \quad (7)$$

Where:

- η is the learning rate,
- δ_j is the delta of the hidden node,
- x_n is the value of input node *n,*

For adjusting weights from the hidden layer to the output layer:

$$\Delta w_{ji} = \eta \delta_i o_j \quad (8)$$

Where:

- δ_i is the delta of the output node *i,*
- o_j is the output of the hidden node *j,*

5. RESEARCH FINDINGS

This section outlines the implementation methodology and evaluation criteria employed in the research.

5.1. Implementation Method

The Weka data mining tool was used to execute the experimental process, which consisted of the following five sequential steps:

1. Data Preparation
2. Dimensionality Reduction
3. Data Sampling
4. Implementation of the Multilayer Perceptron Neural Network Algorithm
5. Data Classification

Each of these steps is explained in more detail below.

5.1.1. Data Preparation (Stage One)

In the first step, a dataset containing 11,055 samples was used. Each sample comprised 30 features. This dataset was obtained from the Kaggle platform. After acquiring the dataset, it was imported into the Weka software environment for subsequent analysis.

Table 1. Important features of phishing websites [18].

Number	Features	Description
1	Use of IP address*	Using an IP address instead of a domain name in a URL, such as http://125.98.3.123/fake.html, is a clear indication that an individual is attempting to steal personal information. In some cases, the IP address may even be converted into a hexadecimal code, as seen in http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html.
2	Long address to hide suspicious part*	Phishing attackers can mask the malicious portion of a URL by utilizing an elongated web address. For example, http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html.
3	Use URL shortening services	URL shortening is a technique employed on the World Wide Web to significantly reduce the length of a web address while still directing users to the intended web page. This is achieved through the use of an HTTP redirect on a short domain name, linking to a web page with a longer address. A prime example of this is the URL http://portal.hud.ac.uk/, which can be shortened to "bit.ly/22bcdx".
4	Web address with "@" symbol	The "@" symbol in a URL instructs the browser to disregard anything preceding the "@" symbol, with the URL commonly appearing after the "@" symbol.
5	Redirect using "//"	When a URL path contains "//", it signifies that the user is being redirected to another website. To determine the position of "//", we examine the URL closely. For instance, http://www.legitimate.com//http://www.phishing.com is an example of such URLs. If the URL begins with "HTTP", the "//" must appear in the sixth position. However, if the URL uses "HTTPS", "//" must appear in the seventh position.
6	Add a prefix or a separated suffix to the web address (-)	The dash symbol is seldom utilized in authentic URLs. Rather, cybercriminals usually affix prefixes or suffixes separated by (-) to domain names to present a façade of legitimacy to unsuspecting users. For instance: /http://www.Confirme-paypal.com.
7	Subdomains and multiple domains	Let's consider the following link: http://www.hud.ac.uk/students/. The domain name includes a country code top-level domain (ccTLD), which in this case is "UK". The "ac" component denotes "university", while "ac.uk" is referred to as a combined second-level domain (SLD), and "hud" is the actual domain name. It should be noted that a valid URL contains two dots due to the fact that we can reference it by including "www". URLs with three dots are considered "suspicious" as they include a subdomain. However, if there are more than three dots, the website is classified as "phishing" because it will have multiple subdomains.
8	Cloud text transfer protocol with secure connection layer	The secure connection layer ensures encryption of data transmission. In browsers, this is indicated by the lock symbol displayed when an SSL certificate is present. Trusted certification authorities include GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster, and VeriSign. Obtaining an SSL certificate is crucial for websites as phishers tend to target sites lacking SSL certificates and relying solely on HTTP. Additionally, after conducting a test on the dataset, it was discovered that a valid certificate must be at least two years old.
9	Length of domain registration	Given that phishing websites have a short lifespan, reliable domains are likely to pay for themselves several years beforehand. The longest fraudulent domains in our dataset were active for only a year.
10	favicon	A favicon is a graphic icon linked to a particular webpage. Numerous user agents including graphical browsers and newsfeeds display the favicon in the address bar as a visual cue of the website's identity. However, downloading the favicon from a domain other than the one displayed in the address bar might suggest a phishing attempt.
11	Using a non-standard port	This feature proves valuable in verifying a particular service, such as HTTP, or on a specific server. For intrusion management purposes, it is advisable to open only essential ports. Having all ports open would

		give phishers the ability to launch just about any service, putting user information at risk.
12	The presence of the "HTTPS" indicator in the domain field of the web address	Phishers may deceive users by incorporating the "HTTPS" indicator in the domain portion of a URL. For instance, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/ serves as an example.
13	Request a web address	For authentic websites, the majority of elements within the webpage are linked to the identical domain. For instance, if the URL entered in the address bar is http://www.hud.ac.uk/students/portal.com, we extract the keyword <src=> from the source webpage and verify if the domain in the URL differs from the address in <src=>. If the outcome is affirmative, the website is categorized as "phishing".
14	Anchor web address	The anchor element is defined by the <a> tag and functions similarly to a URL request.
15	Links in <Meta>, <Script> and <Link> tags	Our research shows that credible websites frequently employ the <Meta> tag to offer metadata about the HTML file. <Script> tags are utilized to generate client-side <Script> elements, while <Link> tags are used to fetch other resources. These tags are assumed to be related to the same domain as the webpage. This enables us to provide supplementary details about our webpage for search engines and indexing "robots." The description section should provide a concise overview of the page's content, while the keywords section should include a list of comma-separated words for indexing purposes.
16	Server form controller	SFHs containing a string or "about:blank" are considered suspicious and require action based on the information submitted. Furthermore, if the domain name in SFHs differs from the domain name of the web page, it suggests the web page is suspicious as external domains infrequently handle submitted information.
17	Send information to email	The webpage's form permits the user to input personal data and sends it to the server for processing. A phisher can redirect the personal information to their own email by utilizing a server-side scripting language like the PHP "mail()" function or a client-side function such as the ":mailto" function.
18	Unusual web sign	This information can be obtained from the WHOIS database. A genuine website usually includes the identity as part of its URL.
19	Change the route	Open redirects on websites can be exploited by phishers to link to their own malicious sites. Our dataset shows that phishing websites typically have a maximum of three redirect pages, while legitimate websites rarely utilize this feature. If a website has fewer than two redirect pages, we classify it as "Legitimate". If it has two or three redirect pages, we classify it as "Suspicious". Any website with more than three redirect pages is classified as "Phishing". We use this methodology to assign the appropriate label to each website in question.
20	Customize the status bar	Phishers could potentially utilize JavaScript to display a counterfeit URL in the status bar. In order to obtain this feature, it is necessary to inspect the web page's source code, specifically the "on Mouse Over" event, to determine if any modifications have been made to the status bar.
21	Disable right click	Phishers employ JavaScript to disable right-click functionality to prevent users from viewing and saving a webpage's source code. This tactic, also known as "use onMouseOver to hide the link," can be detected by searching for the event "event.button==2" within the source code to verify whether it has been right-clicked or not.
22	Use the pop-up window	It is uncommon to encounter a legitimate website requiring users to enter personal information via a pop-up window. Nonetheless, this functionality is in use by certain legitimate websites primarily for warning users against fraudulent activities or displaying pleasurable notifications, without any provision of personal data through these pop-ups.
23	IFrame redirection	An IFrame is an HTML element used to display an additional webpage within the current one. Phishers may use the "iframe" tag to make it invisible to the user. In this instance, phishers utilize the "frameBorder" attribute, which prompts the browser to create a visual border.
24	Domain age	This attribute can be obtained from the WHOIS database. Our dataset showcases domains with several phishing URLs at different time locations. By blacklisting the domain list rather than the URL address, users can be safeguarded. However, research indicates that 78% of phishing domains are compromised domains that also serve a legitimate website. Blacklisting these domains can lead to the blacklisting of legitimate websites. Even after removal of the phishing website from the domain, legitimate websites may remain blacklisted for a prolonged period, resulting in reputational harm to the affected website or organization. Certain blacklists such as "Google Blacklist" require an average of 7 hours to update. Upon analyzing the dataset, the minimum legal domain age was found to be 6 months. If the domain is created within the preceding six months, it will be classified as "phishing"; otherwise, it will be considered a "legitimate" website.
25	DNS record	If the claimed identity of a website is not established by the WHOIS database or the hosting brand name, it is categorized as a phishing website. Conversely, if the DNS record is present and not empty, the website is considered legitimate. This classification is based on a source.
26	Website traffic	This function gauges a website's popularity by tracking the quantity of visitors and pages viewed. However, as fraudulent websites often have a brief lifespan, Alexa's database may not acknowledge them. After analyzing our dataset, it was determined that even credible websites can rank within the top 150,000 at their lowest. Thus, if a domain receives no traffic or is not established in Alexa's database, it will be categorized as a "phishing" website. If a website is ranked within the top 150,000, it is classified as "legitimate"; otherwise, it is labeled as "suspicious".
27	page rank	PageRank is a metric with values ranging from 0 to 1, used to gauge web page importance. The higher the

		PageRank score, the more important the webpage is considered to be. Analysis of our dataset reveals that approximately 95% of phishing websites lack any PageRank score. Additionally, the remaining 5% of phishing websites tend to have a PageRank score of "0.2".
28	Google index	This feature verifies whether a website is indexed by Google. Indexed sites appear in search results, while phishing web pages are often short-lived and may not be included in Google's index[3].
29	The number of links pointing to the page	The quantity of links pointing to a webpage is indicative of its legitimacy, regardless of whether some of those links originate from the same domain. Our dataset, despite its brief existence, reveals that 98% of phishing datasets are unrelated.Conversely, authentic websites possess a minimum of 2 external links directing to them.
30	Statistical report based feature	Multiple sites, including PhishTank and UCI, regularly publish reports on phishing websites. These reports are issued on a monthly or quarterly basis.

5.1.2. The second step of the Principal Component Analysis algorithm

using the Principal Component Analysis algorithm, we reduce the number of features from 30 features to 14 features as shown in table (2).

Table 2. Principal Component Analysis of the number of features from 30 features to 14 features

Number	Features	Description
1	Use of IP address	If a user sees an IP address instead of a domain name in the URL, for example http://125.98.3.123/fake.html, it may indicate an attempt to steal their personal information. In some cases, the IP address may even be converted to a hexadecimal code like http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html.
2	Long address to hide suspicious part	Phishing attackers can mask the malicious portion of a URL by utilizing an elongated web address. For example, http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html.
3	Use URL shortening services	URL shortening is a technique employed on the World Wide Web to significantly reduce the length of a web address while still directing users to the intended web page. This is achieved through the use of an HTTP redirect on a short domain name, linking to a web page with a longer address. A prime example of this is the URL http://portal.hud.ac.uk/, which can be shortened to "bit.ly/22bcdx".
4	Web address with "@" symbol	The "@" symbol in a URL instructs the browser to disregard anything preceding the "@" symbol, with the URL commonly appearing after the "@" symbol.
5	Redirect using "//"	When a URL path contains "//", it signifies that the user is being redirected to another website. To determine the position of "//", we examine the URL closely. For instance, http://www.legitimate.com//http://www.phishing.com is an example of such URLs. If the URL begins with "HTTP", the "//" must appear in the sixth position. However, if the URL uses "HTTPS", "//" must appear in the seventh position.
6	Add a prefix or a separated suffix to the web address (-)	The dash symbol is seldom utilized in authentic URLs. Rather, cybercriminals usually affix prefixes or suffixes separated by (-) to domain names to present a façade of legitimacy to unsuspecting users. For instance: /http://www.Confirme-paypal.com.
7	Subdomains and multiple domains	Let's consider the following link: http://www.hud.ac.uk/students/. The domain name includes a country code top-level domain (ccTLD), which in this case is "UK". The "ac" component denotes "university", while "ac.uk" is referred to as a combined second-level domain (SLD), and "hud" is the actual domain name. It should be noted that a valid URL contains two dots due to the fact that we can reference it by including "www". URLs with three dots are considered "suspicious" as they include a subdomain. However, if there are more than three dots, the website is classified as "phishing" because it will have multiple subdomains.
8	Hyper Text Transfer Protocol with Secure Socket Layer	The Secure Socket Layer (SSL) is a secure system that encrypts data transmission. It is displayed with a lock symbol in browsers. Trusted certification authorities include GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster, and VeriSign. Having an SSL certificate is crucial for websites as phishers usually target sites without SSL certificates. Furthermore, after testing the dataset, it was discovered that a valid certificate must be at least two years old.
9	Length of domain registration	Given that phishing websites have a short lifespan, reliable domains are likely to pay for themselves several years beforehand. The longest fraudulent domains in our dataset were active for only a year.
10	Favicon	A favicon is a graphic icon linked to a particular webpage. Numerous user agents including graphical browsers and newsfeeds display the favicon in the address bar as a visual cue of the website's identity. However, downloading the favicon from a domain other than the one displayed in the address bar might suggest a phishing attempt.
11	Send information to email	The webpage's form permits the user to input personal data and sends it to the server for processing. A phisher can redirect the personal information to their own email by utilizing a server-side scripting language like the PHP "mail()" function or a client-side function such as the ":mailto" function.

12	Domain age	This attribute can be obtained from the WHOIS database. Our dataset showcases domains with several phishing URLs at different time locations. By blacklisting the domain list rather than the URL address, users can be safeguarded. However, research indicates that 78% of phishing domains are compromised domains that also serve a legitimate website. Blacklisting these domains can lead to the blacklisting of legitimate websites. Even after removal of the phishing website from the domain, legitimate websites may remain blacklisted for a prolonged period, resulting in reputational harm to the affected website or organization. Certain blacklists such as "Google Blacklist" require an average of 7 hours to update. Upon analyzing the dataset, the minimum legal domain age was found to be 6 months. If the domain is created within the preceding six months, it will be classified as "phishing"; otherwise, it will be considered a "legitimate" website.
13	Website traffic	This function gauges a website's popularity by tracking the quantity of visitors and pages viewed. However, as fraudulent websites often have a brief lifespan, Alexa's database may not acknowledge them. After analyzing our dataset, it was determined that even credible websites can rank within the top 150,000 at their lowest. Thus, if a domain receives no traffic or is not established in Alexa's database, it will be categorized as a "phishing" website. If a website is ranked within the top 150,000, it is classified as "legitimate"; otherwise, it is labeled as "suspicious".
14	Google index	This feature verifies whether a website is indexed by Google. Indexed sites appear in search results, while phishing web pages are often short-lived and may not be included in Google's index.

5.1.3. The third step of data sampling

In this step, we determine the quantity of training and test data sets to be used, with the inclusion of seed=1 and fold=10.

5.1.4. The fourth step of the multi-layer perceptron neural network algorithm

In this step, we select the control for the multi-layer perceptron neural network algorithm to implement it in the Veka tool.

5.1.5. The fifth stage of data classification

In this stage, we estimate the classification accuracy based on the percentage of test samples or data sets that have been successfully classified. Table 3 displays the implementation results.

Table 3. Results of the implementation of the Principal Component Analysis algorithm + multilayer perceptron neural network algorithm

Number	Evaluation criteria	PCA+CART
1	Correctly The number of correctly classified samples	91.64% 10131
2	Incorrectly Number of misclassified samples	3.36% 924
3	Kappa	0.83%
4	Mean absolute error	0.10%
5	Root mean squared error	0.25%
6	Relative absolute error	21.96%
7	Root relative squared error	51.01%
8	Total Number of Instances	11055

. Table 4. Comparison of the proposed method with previous methods

Number	algorithm	Evaluation criteria
1	PCA+MLP	91.64%
2	J48	91.05%
3	FT	88.91%
4	LMT	89.35%

5.2. Evaluation criteria

According to Table 3, 10131 samples were correctly classified, resulting in an accuracy rate of 91.64% for the obtained diagnosis. We used Equation (9) to calculate the detection accuracy.

$$Accuracy = \frac{TP^{\dagger}+TN^{\ddagger}}{TP+TN+FP^{\S}+FN^{**}} \quad (9)$$

According to Table 3, the number of samples that are incorrectly classified is 924 samples. The level of classification inaccuracy is equal to 8.36%, which is calculated based on equation (10):

$$Error = 100 - \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

6. CONCLUSION

Phishing refers to the act of stealing identity information from consumers of electronic stores, financial institutions, and other cyberspace users. In this study, we employed a hybrid approach combining a dimensionality reduction algorithm with a multilayer perceptron neural network to evaluate the accuracy of detecting phishing website attacks in the context of electronic banking. The integration of these two techniques achieved an accuracy rate of 91.64%. The primary objective of this research is to enhance the detection accuracy of phishing attacks in electronic banking by leveraging data mining algorithms.

Declaration

We acknowledge that we used ChatGPT to enhance the academic writing of our manuscript while ensuring the originality and integrity of our work.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Samad, S. R. A., Balasubaramanian, S., Al-Kaabi, A. S., Sharma, B., Chowdhury, S., Mehbodniya, A., Webber, J. L., & Bostani, A. (2023). Analysis of the performance impact of fine-tuned machine learning model

† - True Positive
‡ - True Negative
§ - False Positive
** - False Negative

for phishing URL detection. Electronics.

[2] Yun, X., Zhang, Y., Zhou, Y., Xiao, J., Wang, Y., & Li, S. (2013). Method and system for automatic detection of suspected counterfeit websites. China.

[3] Raychura, V. D., & Parekh, C. D. (2020). A new deceptive tactic aims better cyber-solutions for organizations during global pandemic. In Proceedings of the 7th International Conference on Cyber Security and Privacy (pp. 834–840).

[4] Raghupathi, V., & Raghupathi, W. (2020). The influence of education on health: An empirical assessment of OECD countries for the period 1995–2015. Archives of Public Health, 78. https://doi.org/10.1186/s13690-020-00402-5

[5] APWG. (2021). APWG Symposium on Electronic Crime Research, eCrime 2021, Boston, MA, USA, December 1–3, 2021.

[6] Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In European Conference on Computer Vision (pp. 740–755). https://doi.org/10.1007/978-3-319-10602-1_48

[7] Monson, K., Smith, E., & Bajic, S. (2022). Planning, design and logistics of a decision analysis study: The FBI/Ames study involving forensic firearms examiners. Forensic Science International: Synergy, 4, 100221. https://doi.org/10.1016/j.fsisyn.2022.100221

[8] Pinguelo, F. M., & Muller, B. W. (2012). Virtual crimes, real damages part II: What businesses can do today to protect themselves from cybercrime, and what public-private partnerships are attempting to achieve for the nation of tomorrow. eBusiness & eCommerce eJournal.

[9] Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S., Ramana, S., & Joga, K. (2023). Phishing detection system through hybrid machine learning based on URL. IEEE Access, 11, 36805–36822. https://doi.org/10.1109/ACCESS.2023.3252366

[10] Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., ... Zhao, G. (2020). Searching central difference convolutional networks for face anti-spoofing. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR42600.2020.00534

[11] Campajola, C., Cristodaro, R., De Collibus, F. M., Yan, T., Vallarano, N., & Tessone, C. (2022). The evolution of centralisation on cryptocurrency platforms. arXiv preprint arXiv:2206.05081.

[12] Distler, V. (2023). The influence of context on response to spear-phishing attacks: An in-situ deception study. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3544548.3581170

[13] Haddadnia, J., Seryasat, O. R., & Rabiee, H. (2013). Thyroid diseases diagnosis using probabilistic neural network and principal component analysis. Journal of Basic and Applied Science Research, 3(2), 593–598.

[14] Seryasat, O. R., Zadeh, H. G., Ghane, M., Abooalizadeh, Z., Taherkhani, A., & Maleki, F. (2013). Fault diagnosis of ball-bearings using principal component analysis and support-vector machine. Life Science Journal, 10(1s), 393–397.

[15] Sood, K., Nosouhi, M., Nguyen, D. D. N., Jiang, F., Chowdhury, M. U., & Doss, R. (2023). Intrusion detection scheme with dimensionality reduction in next generation networks. IEEE Transactions on Information Forensics and Security, 18(4), 965–979. https://doi.org/10.1109/TIFS.2022.3233777

[16] Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.

[17] Yu, H., Liu, Y., Zhou, G., & Peng, M. (2023). Multilayer perceptron algorithm-assisted flexible piezoresistive PDMS/Chitosan/cMWCNT sponge pressure sensor for sedentary healthcare monitoring. ACS Sensors. https://doi.org/10.1021/acssensors.3c01885

[18] Mosharzadeh, S., Hashemipour, S. Z., & Hekmatian Raz, M. (2022). Detection of phishing websites using a combination of dimensionality reduction and simple Bayesian algorithms. In 9th National Congress of Iranian Electrical and Computer Engineering News, Tehran. https://civilica.com/doc/1636775