*I.J. Modern Education and Computer Science*, **2023, 5, 42-59**
Published Online on October 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2023.05.04

# An Intelligent System for Detecting Fake Materials on the Internet

**Aya S. Noah***
Al-Azhar University, Faculty of Science (Girl's branch), Mathematics Department, Cairo, 11754, Egypt
Email: AyaNoah41167@azhar.edu.eg
ORCID iD: https://orcid.org/0009-0005-8333-1183
*Corresponding Author

**Naglaa E. Ghannam**
Al-Azhar University, Faculty of Science (Girl's branch), Mathematics Department, Cairo, 11754, Egypt
Email: naglaasaeed@azhar.edu.eg
ORCID iD: https://orcid.org/0000-0001-7145-5617

**Gaber A. Elsharawy**
Al-Azhar University, Faculty of Science (Girl's branch), Mathematics Department, Cairo, 11754, Egypt
Email: Gaber.Ahmed@buc.edu.eg
ORCID iD: https://orcid.org/0000-0001-6807-7600

**Abeer S. Desuky**
Al-Azhar University, Faculty of Science (Girl's branch), Mathematics Department, Cairo, 11754, Egypt
Email: abeerdesuky@azhar.edu.eg
ORCID iD: https://orcid.org/0000-0003-1661-9134

Received: 20 February, 2023; Revised: 27 March, 2023; Accepted: 09 June, 2023; Published: 08 October, 2023

**Abstract:** There has been a significant rise in internet usage in recent years, which has led to the presence of data theft and the diversity of counterfeit materials. This has resulted the proliferation of cybercrimes and the theft of personal data via social media, e-mail, and phishing websites that are similar to the websites commonly used to grab user data details like that of a credit card or login ID. Phishing, a prevalent form of cybercrime, poses a danger to online security through the theft of personal information, and with the emergence of the COVID-19 virus, which has led to people and organizations being drawn towards the Internet and many people and companies being forced to work remotely, it has led to an increase in the existing phishing threats. Previously, hackers took advantage of the situation to infiltrate the devices of people and companies in numerous ways, which caused huge financial losses and damage to organizations. Based on previous results and research, Machine Learning (ML) is selected by researchers as an efficient method for identifying malicious software web pages from original web pages. This paper presents 30 characteristics of websites, which are analyzed using a correlation matrix to determine the relationship between variables. Feature selection is performed through a wrapper method and Extra Tree Classifiers (ETC) to identify the top-ranked characteristics (Features) for website classification. To evaluate web pages, various machine learning techniques such as Random Forest Tree (RF), Multilayer Perceptron (MLP), Decision Tree (DT), and Support Vector Machine (SVM) are used. The results of monitoring indicate that MLP, a deep neural network, outperforms all other techniques in terms of performance.

**Index Terms:** Deep Neural Networks, MLP Classifier, Fake Websites, Fake materials, Feature Selection.

## 1. Introduction

As the number of internet users grows, so does the amount of data generated, which in turn leads to more attempts to access and steal this data through fraudulent means. Phishing is the most prevalent and severe form of attack in the digital world, and it is emerging as one of the techniques used in data theft and aims to obtain sensitive and confidential information about users, including usernames, bank account passwords, and credit card information [1]. An attacker exploits a replica of a legitimate website with the intent to deceive users. The phishing website's hyperlink is then distributed to a large number of Internet users via email and other means of communication [2].

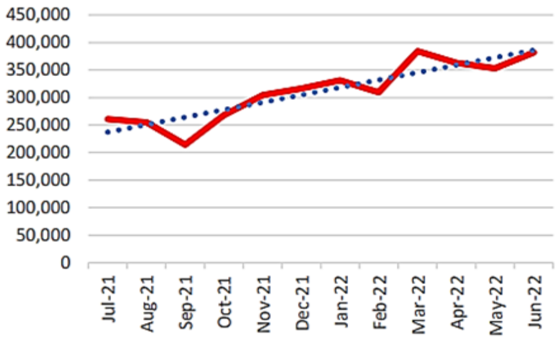


Fig. 1. Phishing Attacks increase from Q3 2021 to Q2 2022 [4].

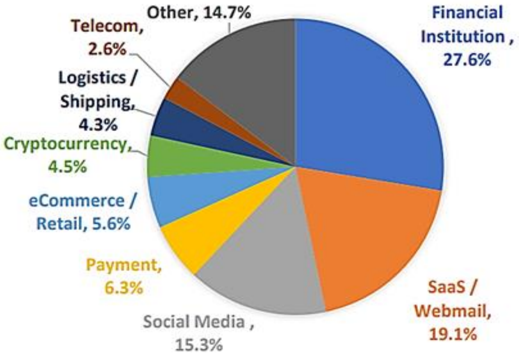


Fig. 2. Targeted manufacture, Q2 2022 [4].

In addition to using the usual methods such as email and SMS to direct to fake websites in phishing attacks, as the COVID-19 pandemic arose in March 2020, phishing attacks significantly increased everywhere, on social media, and especially for payment and e-commerce sites [3].

Phishing attacks continue to evolve and gain strength, as this increase is shown in figure 1 shows a significant increase in the number of phishing attacks from the third quarter of 2021 to the second quarter of 2022. The x-axis shows the time period, while the y-axis shows the number of confirmed phishing attacks, with the highest number of attacks recorded in Q2 2022. This trend highlights the growing threat of phishing attacks and the need for increased vigilance in protecting personal and sensitive information online. Anti-Phishing Working Group Research (APWG) In September 2022, the APWG recorded a record-breaking 1,097,811 as the sum of all confirmed cases of phishing, the worst quarter for phishing ever recorded by the APWG. Since the beginning of 2020, the APWG (Anti-Phishing Working Group) has recorded a monthly average of between 68,000 and 94,000 phishing attacks, the number of phishing attacks reported to the organization has quadrupled. The banking sector and attacks on webmail and software-as-a-service (SAAS) providers remained dominant, accounting for 27.6 and 19.1 percent, respectively, of the total attacks on phishing mentioned in figure 2 [4].

In a phishing process, a scammer sends emails to users that contain a hyperlink related to secure websites, such as banks, which are legitimate and then diverts them from the original website to the phishing website [5]. Once a consumer presses the link, they will be redirected to a phishing website that requests information, including personal banking information, credit card details, and changes in passwords, while the individual submits this information, the scammer selects the details from the server, which are used to transfer the individual's funds from the financial institution to their own account [6].

To combat the problem of detecting phishing attacks, blacklist-based, whitelisting-based, toolbar-based, and heuristic techniques are proposed. Some of the pre-defined blacklist URLs will be checked against those requested by the user, and a whitelist of good URLs is kept in rotation. The problem with this approach lies in the requirements for periodic updating of menus, the internet, its various functionalities, and a tool for searching and retrieving information [7, 8]. Additionally, the exertion required to maintain blacklists is substantial. This is due to the limited lifespan of phishing websites and the ease with which new ones can be made. Because of these hurdles with the blacklist-based detection method, heuristics-based procedures were created. Using methods of machine learning and artificial intelligence, phishing websites were identified. To distinguish phishing websites from legitimate websites, diverse classification techniques exist, a variety of methods including logistic regression, stochastic gradient descent, random forest, Support Vector Machine (SVM), nave bayes, k-nearest neighbors, and Decision Tree (DT) are utilized [9].

In this study, several machine learning techniques were employed, such as Random Forest Tree (RF), Multilayer Perceptron (MLP), Decision Tree (DT), and Support Vector Machine (SVM), to detect phishing URLs. Results indicate that deep neural networks utilizing Multi-Layer Perceptron Neural Networks (MLPNN) exhibited the best performance among all the techniques with 30 attributes of HTML page to distinguish between legitimate and phishing sites. Additionally, the study utilized feature selection to gain insights into website characteristics and the specific features that significantly impact the classification of a website as phishing or legitimate. The suggested supervised learning employs a classifier to evaluate a set of phishing frauds that can correctly categorize unseen data, thus effectively identifying new phishing sites.

The structure of the remaining sections is as follows: The next section presents a summary of current research on various phishing detection methods. Section explicitly describes the ML prediction technique used, all features that has been used in this study are briefly discussed, and feature selection techniques. After that, Section 4 examines the outcomes obtained through evaluating the execution of various ML classifiers utilizing FS methods. The top-performing classifier for identifying phishing attacks is presented based on the evaluation. Section 5 presents the conclusion.

## 2. Related Works

Internet business transactions pose a significant obstacle to implementing web security. URLs have a crucial role in both types of phishing schemes, whether they are web-based or email-based. So, the researchers suggested diverse ways to detect if a URL is phishing or legitimate, several features were identified based on the most relevant ones based on machine learning algorithms [1]. According to [11], the authors proposed the use of an updated feature selection algorithm to identify a set of capabilities that greatly improve the percentage of cases of phishing found in Internet of Things environments. To acquire the most fitting set of features, they employed a feature selection algorithm. When employing Random Forest with recommended data representation. In [12], the phishing defense system that is postulated utilizes seven various classification methods, and features based on Natural Language Processing (NLP) in addition to other distinct feature sets. The Random Forest technique with only NLP-based functionalities exhibits the best achievement, as shown by experimental and comparative results of the implemented classification methods for detecting phishing URLs, with an accuracy rate of 97.98%.

In recent times, several phishing detection models based on deep learning have been developed [13].

Ref. [14] proposed an efficient phishing detection system that is based on Hybrid Deep Learning (HDL) and the Modified crow search-based deep learning neural network (MCS-DNN) classifier. This system applies pre-processing, clustering, feature selection, classification, and KCV (k-fold cross-validation) as opposed to Artificial NN (ANN), K-Nearest Neighbors (KNN), and SVM techniques, which focus on metrics such as recall, accuracy, F-score, precision, False Negative Rate (FNR), False Positive Rate (FPR), Matthews Correlation Coefficient (MCC), True Negative Rate (TNR) and True Correct Classification (TCC). Compared to conventional methods, the one being proposed yields better results.

Authors in [15] proposed using a combination of deep neural network (DNN) and long short-term memory (LSTM) algorithms within Hybrid Deep Learning (HDL) features for identifying phishing URLs. These models are trained on two datasets, incorporating both character embedding and NLP features. The results of the study indicate the effectiveness of the proposed models is superior to other phishing detection models; however, the authors express concern that the datasets used may not accurately reflect real-world business attacks.

Ref. [16], intelligent phishing exposure in web pages was proposed using Adam's supervised deep learning classification and optimization technique. They trained the neural network using 11,000 websites and produced results that are more accurate.

Ref. [17], suggested feature engineering techniques that combine methods for linear and non-linear space transition to improve classifier performance in detecting phishing URLs. Five models of space transformation (Nystrom methods, NYS-DML, singular value decomposition, DML-NYS, and distance metric learning) were applied to 33,1622 URLs with 62 features. Each model concentrates on a specific space revision aspect, and the combined models leverage linear, nonlinear, supervised, unsupervised and models all have advantages. The research shows that the proposed methods improve the effectiveness and overall quality of specific classification models for detecting phishing URLs.

Table 1 summarizes and compares recent research papers that propose different solutions for phishing website detection using machine learning and deep learning. Specifically, we compare the authors' data sources, proposed solutions, limitations, and years of publication to gain insights into the state-of-the-art in phishing URL detection. By analyzing and comparing the approaches and limitations of these studies, we aim to provide a comprehensive overview of the current trends and challenges in the field of phishing website detection using machine learning and deep learning.

This study compares correlation methods based on Karl Pearson's correlation coefficient and an algorithm for machine learning. In addition, on specific features, the performance of four Random Forests (RF), (SVM), (DT), and MLP is estimated.

Table 1. A Comparative study of the Related works

| References | Data Source | Proposed Solution | Limitations | Year of Publication |
|---|---|---|---|---|
| Ref. [11] | IoT | Lightweight data representation for phishing URL detection | The proposed solution's effectiveness may be limited to specific IoT environments. | 2022 |
| Ref. [12] | URLs | Machine learning-based phishing detection from URLs | The proposed approach may not generalize well to new and unseen types of phishing URLs. | 2019 |
| Ref. [13] | URLs | Systematic literature review of applications of deep learning for phishing detection | The effectiveness of deep learning approaches for phishing detection in some contexts may be limited by the availability of labeled data. | 2022 |
| Ref. [14] | URLs | Hybrid deep learning-based phishing detection system using MCS-Dnn Classifier | The proposed approach may be limited to specific types of phishing URLs. | 2022 |
| Ref. [15] | URLs | Hybrid DNN-LSTM model for detecting phishing URLs | The proposed approach may not generalize well to new and unseen types of phishing URLs. | 2021 |
| Ref. [16] | Web pages | Smart phishing detection using supervised deep learning classification and optimization technique ADAM | The effectiveness of the proposed approach may be limited to specific types of phishing URLs. | 2021 |
| Ref. [17] | URLs | Feature engineering for improving malicious URL detection | The proposed approach may be limited by the availability of labeled data and the choice of feature engineering techniques. | 2020 |

Content analysis is less secure and less efficient than the detection of malicious uniform resource locators (URLs). Even so, the diagnosis of malicious URLs remains insufficient owing to inadequate characteristics and improper classification. This is due to the vulnerability of Web address features to hacker manipulation and the possibility of adaptive modification, which makes them inadequate for efficient illustration. Utilizing evasion technics, attackers could indeed circumvent security countermeasures. As a result, features gleaned from such URLs may be misleading, as attackers may have masked the site’s true nature and behavior or manipulated them. Therefore, attacker-uncontrollable functions can be advantageous for enhancing detection precision and reducing rates of false alarms [18].

## 3. Data and Methodology

This work aims to highlight the significant features of the websites, such as URL and domain identity, abnormal behavior-based, domain-based characteristics, JavaScript-based characteristics, and HTML. This increases the quantity of attributes to encompass basic and comprehensive web characteristics as compared to other research and improves the rate of classifying when using machine-learning technologies. By examining the relationship between features, results, and features, a subset of prominent features can be identified from the original set, decreasing the quantity of unnecessary datasets to enhance classification effectiveness, and understanding the difference between the quality of sites if they are legitimate or malicious. Choosing features helps understand data, reduces the "dimensional curse effect", reduces measurement requirements, enhances accuracy, and identifies features that can be applied to a specific problem [19].

First, a deep neural network-based phishing detection model is constructed utilizing MLP (Multi-Layer Perceptron) through the training dataset. Features are first rolled out, then the linkage between the qualities used and the outcomes is made to know the closest attributes that affect the outcome of the website, whether it is sound or deceptive, including to detect unknown websites. In this paper, the following sections describe the process of correlation between traits and each other, the attributes, and results of the site, the NN structure, the NN design, and the NN training process.

### *3.1 Methodology*

The technique utilized in this paper is to eliminate data characteristics. First, a cross-correlation between features using Karl Pearson's method, taking into account the correlation coefficient between the attributes, is used to identify which features affect the determination of whether a website is fraudulent or not. Second, removing the features with the least impact on the score through classification using an RFE; the second round of feature processing is carried out by using Recursive Feature Elimination (RFE) on the remaining features as a wrapper-style algorithm to find the features that can identify the type of website, whether it is phishing or legitimate, figure 3 depicts this. The set of data is then evaluated using ML and deep learning techniques such as RF, DT, and SVM, which produce ideal results when compared to deep learning algorithms like Multilayer Perceptron (MLP).

Recall that Multilayer Perceptron NN (MLPNN) is a deep learning technique that is a common representation of a Feed-Forward NN (FFNN) or Deep Feed Forward Network (DFFN). It is constituted of input, hidden, and output layers. [20] Among the various neural networks, MLPNN is considered one of the most popular, along with Convolutional NN (CNN), Recurrent NN (RNN), Autoencoder (AE), and Generative Adversarial Networks (GAN) [21]. The neurons in an MLPNN are trained with the backpropagation learning algorithm [22]. The proposed MLP model utilized the ReLU activation function, and the Adam optimizer provides a fitting model.

Random Forest offers two methods for selecting characteristics, together with average decrease in impurity and average decrease in, Random Forest is one of the most recent efficient research determinations for decision tree learning [23,24], RF is a ML algorithm that is highly regarded for its ease of use, high precision, and robustness. This algorithm utilizes a collection of decision trees to classify and predict outcomes from the training data. It is considered one of the most popular machine learning algorithms due to its versatility, as it is applicable to a variety of data sets and problems.

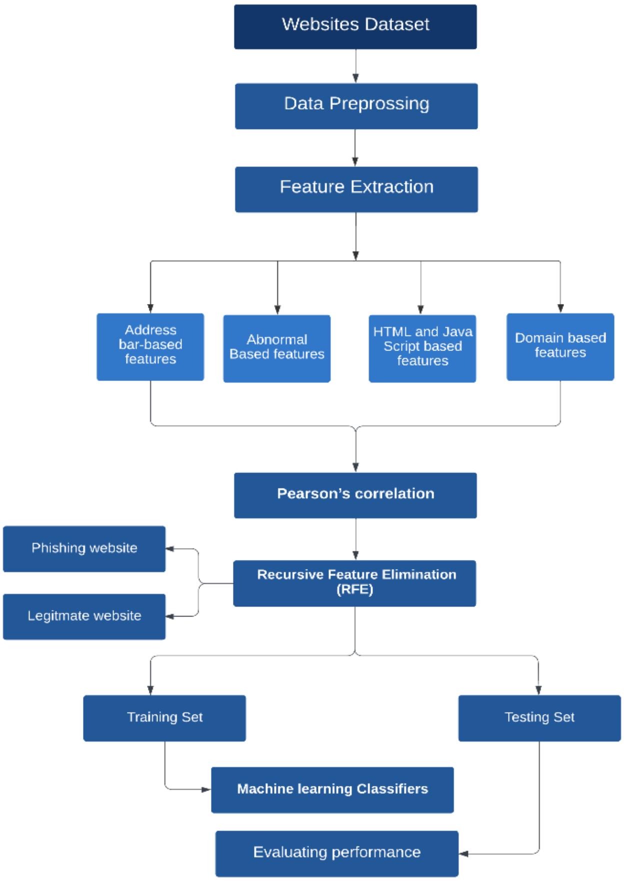


Fig. 3. System Architecture.

Decision tree algorithms are a popular choice for building classification models due to their similarity to human reasoning and ease of understanding. The structure of a DT model includes a root node, branches, internal nodes, and leaf nodes, with the root node at the top representing the input and the leaf nodes at the bottom representing the outcomes or decisions. This structure allows for easy interpretation and comprehension of the model's predictions and decision-making process [25]. Decision tree learning is a technique that utilizes a divide-and-conquer approach. It uses a greedy search algorithm to identify the best possible split points in a tree. This process is then repeated recursively until the majority, or all of the documents have been assigned specific class labels. This method enables the model to iteratively improve the classification of the data, resulting in a more accurate and efficient model [26].

The SVM is a model recognition technology that is developed as a part of the AI decision-making system. It is used for dual classification and requires statistical features as input. Compared to other classifiers, such as decision trees, SVM provides information that is more precise. The algorithm iteratively finds an optimal hyperplane that minimizes an error. The SVM aims to discover the Maximum Marginal Hyperplane (MMH), which ideally divides the dataset into classes. Because of this, SVM is also referred to as a discriminatory classifier [27].

### *3.2 Performance of ML algorithms*

In an effort to assess and compare the performance of various machine learning algorithms in identifying website types, researchers employed a range of metrics, including accuracy, precision, Sensitivity, F1-score, and weighted average. These metrics were used to measure the efficiency of the different models [28].

Table 2. Confusion Matrix

| | Positive Predicted Class | Anticipated Negative Class |
|---|---|---|
| Positive Original Class | Truly Positive (TP) | False Positive (FP) |
| Negative Original Class | False Negative (FN) | Truly Negative (TN) |

Equation (1), the metric of precision is used to measure the multitude of accurate forecasts phishing websites among all the websites that were predicted as phishing.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

Equation (2), the metric of recall, also referred to as sensibility, is used to measure the multitude of phishing websites that the model accurately identifies out of all the phishing websites present in the data.

$$TPR = Sensitivity = Recall = \frac{TP}{TP+FN} = 1 - FPR \tag{2}$$

Equation (3), the metric of accuracy measures the proportion of correct predictions.

$$Accuracy = \frac{\#Correct\ Predictions}{\#Total\ Predictions} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Equation (4), the metric of F1-score ranges from 0 to 1 and is the harmonic mean of the recall and precision parameters.

$$F1 - score = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \tag{4}$$

The efficiency of the selected classification model is assessed utilizing measures that are derived from the Confusion Matrix (CM) is shown in Table 2, which displays the general layout of the confusion matrix. All the measures used in this report are computed from the values in this matrix.

All the algorithms proposed in this report have their implementation in Python. Within the context of the CM, True Positive (TP) refers to a correct prediction of a phishing webpage, whereas True Negative (TN) is an accurate identification of a legitimate website. False Positive (FP) is when a webpage is incorrectly classified as phishing, and False Negative (FN) is whenever a website is incorrectly identified as legitimate.

### *3.3 Data Description and Pre-Processing*

Data preprocessing is a technique in data mining that transforms unstructured data into a format that is useful and efficient. It is used to enhance the cleansing, transformation, and structuring of data besides increasing the precision of a novel model while decreasing the amount of computing required. Preprocessing includes techniques such as cleaning, integration, modification, normalization, reduction, and feature selection to increase the accuracy of a new model while decreasing the amount of required computing [29]. Data is typically received in spreadsheet format and must be converted into a Python-compatible format, so spreadsheets are converted into CSV files.

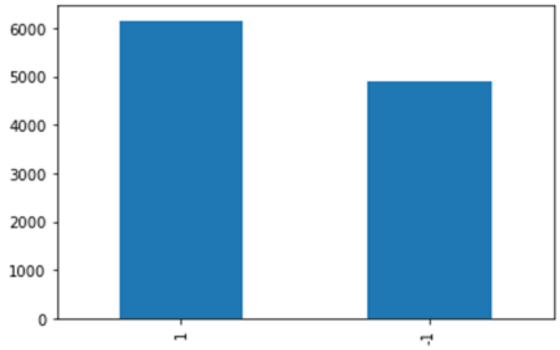


Fig. 4. Overview of Dataset.

To solve problems successfully in machine learning. Finding high-quality data is essential, so this research uses data from the UCI Machine Learning Repository [30]. It contains thirty attributes related to each website. The set of data utilized in this research includes instances (websites), with 6,157 legitimate and 4,898 phishing URLs, resulting in a ratio of 65% for legitimate websites and 44% for phishing websites, as shown in figure 4. As per the data description, the values in the dataset are associated with the following interpretations: (1) represents legitimate, (0) represents suspicious, and (-1) represents phishing. The result is output as a numeric value between -1 and 1, indicating whether or not the tested webpage is a phishing site.

Under this study, 80% of the dataset is used to train the model, and the other 20% is used to evaluate how well the model works after it has been trained. The distribution of the target variable values seems to be well balanced.

Table 3. Features Value Rang

| Features | Feature Type | Feature Explanation | Value Range |
|---|---|---|---|
| Address based features | Having Address | It is possible that a website is attempting to deceive users through phishing if the Internet Protocol (IP) address is used in place of the Domain Name System (DNS) in the URL.<br>IF conditional rule $\begin{cases} \text{DNS contains an IP address} \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |
| | Uniform Resource Locator (URL) Length | An attribute that is used to identify legitimate websites is the URL's length. If the URL length is less than 54 characters, the website is considered legitimate, and the attribute value is assigned as 1. If the length is between 54 and 75 characters, the attribute value is set to 0, indicating a suspicious website. The webpage is classified as phishing if the URL length exceeds 75 characters, and the attribute value is -1.<br>IF conditional rule $\begin{cases} \text{URL length is less than} < 54 \rightarrow \text{feature} = \text{Legitimate} \\ \text{else if URL length} \geq 54 \text{ and less than or equal } \leq 75 \rightarrow \text{feature} = \\ \text{Suspicious} \\ \text{ELSE} \rightarrow \text{feature} = \text{Phishing} \end{cases}$ | {1,0, -1} |
| | URL Shorte | On the World Wide Web, URL shortening is a technique whereby a URL can be significantly shortened while still leading to the requested web page. This is a method for inserting malicious links or logging user information.<br>IF conditional rule $\begin{cases} \text{TinyURL} \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | {1, -1} |
| | Having '@' symbol | The "@" symbol in a URL is an indicator of a phishing website. Browsers typically ignore the text preceding the "@" symbol and the actual address is often located after it. Therefore, if URL includes the "@" symbol, it is a phishing URL.<br>IF conditional rule $\begin{cases} \text{Url Containing @ Symbol} \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | {1, -1} |
| | Using '//' fo redirecting | In a valid URL, the protocol (HTTPS) is indicated by a colon and double forward slashes following it. These two slashes should only appear once in the URL. However, phishing sites may use double slashes in an attempt to redirect users to their fraudulent site. For example, a legitimate URL would be written as "https://github.com/features," while a phishing URL may appear as https://github.com//features.<br>IF conditional rule $\begin{cases} \text{Last "//" in the URL} > 7 \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |
| | Prefixing or Suffixing th Domain wit | Utilization of the dash emblem (-) in URLs is uncommon among reputable websites. However, scammers may incorporate this symbol by appending prefixes or suffixes to the web domain, making it difficult for users to distinguish between phishing and legitimate websites. A phishing Web page such as "http://www.Confirme-paypal.com/" is an example of this.<br>IF conditional rule $\begin{cases} \text{DNS Name Component Contains } (-) \text{ Symbol} \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |
| | Having Multiple subdomains | Subdomains are subsets of larger domains. These parameters indicate whether a website is phishing based on the number of remaining domain points expelling the domain's top-level, 2nd, and "www" sub - domains, have been removed. If it is less than 1 point, the site is not phishing and the parameter value is 1, if it is greater than 1 point, the site is suspicious and the parameter value is 0, and if it is less than -1, the parameter value is -1.<br>IF conditional rule $\begin{cases} \text{Dots In DNS Portion} = 1 \rightarrow \text{Legitimate} \\ \text{Dots In DNS Portion} = 2 \rightarrow \text{Suspicious} \\ \text{ELSE} \rightarrow \text{Phishing} \end{cases}$ | {1,0, -1} |
| | HTTPS - SSL final state | This parameter indicates that the presence of HTTPS is important for validating the site's legitimacy, but it is not sufficient. The certificate for HTTPS must be verified along with number of times the trust credential was issued and the certificate's age if it is one year old and was issued by a trusted authority. The parameter value is 1 if HTTPS is one year old and issued by a trusted authority. HTTPS issued by a suspect authority parameter value 0; otherwise, -1 for a non-HTTPS site.<br>IF conditional rule $\begin{cases} \text{HTTPS, Trusted Issuer, and Certificate Age} \geq 1 \text{ Year} \rightarrow \text{Legitimate} \\ \text{HTTPS Issuer Untrusted} \rightarrow \text{Suspicious} \\ \text{ELSE} \rightarrow \text{Phishing} \end{cases}$ | {1,0, -1} |
| | Term of Domain Registration | The duration of domain registration is measured by the amount of time for which a domain is registered. It is believed that reputable domains are registered for multiple years ahead of time, whereas phishing sites tend to be registered for shorter periods.<br>IF conditional rule $\begin{cases} \text{DNS Terminates on} \leq 1 \text{ years} \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |
| | Favicon | A favicon is a small website icon. It represents the website's identity in the browser's address bar. If the favicon is from a different domain than the address bar, the page may be phishing. The website may also be phishing if the favicon resembles the original.<br>IF conditional rule $\begin{cases} \text{Favicon Uploaded Externally to DNS} \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |
| | Utilizing an Unstandardi Port | This parameter is beneficial for determining whether a particular server is utilizing standard ports such as HTTPS, FTP, etc. If the port numbers are presented as legitimate. Other port numbers are commonly associated with phishing websites. To prevent intrusions, many security systems block a significant portion or all of the ports, only allowing access to specific ports. This is because unblocked open ports can provide phishers with access to sensitive user information.<br>IF conditional rule $\begin{cases} \text{Port \# is of the Preferential Recognition} \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |
| | Adding HTT to the domai | Scammers may use URLs like https://https-paypal.com in an effort to trick users by adding the "HTTPS" code to the domain portion of the URL.<br>IF conditional rule $\begin{cases} \text{Utilizing HTTP Symbol in the DNS Fraction of the URL.} \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |

| | | | |
|---|---|---|---|
| abnormal characteri-stics | URL reques | This parameter indicates regardless of request URL checks whether the images, videos, or any of them are from an external site. Therefore, if the percentage of copied content is less than 22% of the site, it is not considered phishing, and the parameter value is 1. If the site has more than 22 percent but less than 61 percent of sites classified as phishing, the parameter is set to 0, and if more than 61 percent of sites are classified as phishing, the parameter is set to -1.<br>IF conditional rule $\begin{cases} \text{\% of the Demand URL } < 22\% \rightarrow \text{ Legitimate} \\ \text{\% of the Demand URL} \geq 22\% \text{ and } 61\% \rightarrow \text{ Suspicious} \\ \text{ELSE} \rightarrow \text{ feature} = \text{Phishing} \end{cases}$ | { -1,1} |
| | URL Ancho | The <a> tag defines an anchor element, which serves the same function as the Request URL. This parameter indicates whether or not the website's anchor tags point to a different domain. If less than 31% of anchor tags point to different domains, the site is classified as legitimate and the parameter value is 1, The percentage of anchor tags pointing to different domains can be utilized to categorize a webpage as suspicious or phishing. If between 31% and 67% of anchor tags juncture to distinct domains, the website is deemed suspicious, and the parameter value is set to 0. If more than 67 percent of anchor tags point to different domains, the site is identified as phishing, and the parameter value is set to -1.<br>IF conditional rule $\begin{cases} \text{\% of Anchor URL} < 31\% \rightarrow \mathit{Legitimate} \\ \text{\% of Anchor URL} \geq 31\% \text{ And} \leq 67\% \rightarrow \text{ Suspicious} \\ \text{ELSE} \rightarrow \text{ Phishing} \end{cases}$ | {1,0, -1} |
| | Links cont in <M <Script>, and <Link> | Meta tags are frequently used by reputable websites to provide additional information about a Html page. These tags are typically associated with the same domain as the webpage, such as <Link> to retrieve additional web resources. This parameter indicates whether or not website tags link to the same domain as the website itself. Phishers also use it to direct users to additional phishing websites. If fewer than 17 percent of links point to a different page, the site will be ranked. As original and the parameter values are 1, if more than 17% but less than 81% of the website is classified as suspicious and the parameter value is 0, and if more than 81% of the webpage is regarded to be phishing, and the parameter value is assigned as -1.<br>IF conditional rule $\begin{cases} \text{\% of Links in " } < \text{Script} >, < \text{Meta} > \text{ and } < \text{Link} > \text{ " } < 17\% \rightarrow \text{ Legitimate} \\ \text{\% of Links in <Script>,<Meta> and <Link>} \geq 17\% \text{ And} \leq 81\% \rightarrow \\ \text{Suspicious} \\ \text{ELSE} \rightarrow \text{ Phishing} \end{cases}$ | {1,0, -1} |
| | (SFH) Server Form Handler | With a parameter value of -1, an SFH containing an empty string or "about: blank" is classified as a phishing site. If the domain name in the SFH differs from the domain name of the webpage, the domain name in the SFH will be used, it is considered suspicious, with a parameter value of 0, as external domains usually do not monitor the information provided. On the other hand, if a website's domain name is identical in SFH, it is classified as original, with a parameter value of 1.<br>IF conditional rule $\begin{cases} \text{"About: blank" or Empty SFH} \rightarrow \text{ Phishing} \\ \text{SFH denotes a different domain} \rightarrow \text{ Suspicious} \\ \text{ELSE} \rightarrow \text{ Legitimate} \end{cases}$ | {1,0, -1} |
| | Sending Information through Ema | This parameter indicates whether the mail () or mailto () function exists. If these functions exist, the parameter value is -1; otherwise, the parameter value is 1.<br>IF conditional rule $\begin{cases} \text{Submitting User Data with "mail()" or "mailto: "} \rightarrow \text{ Phishing} \\ \text{ELSE} \rightarrow \text{ Legitimate} \end{cases}$ | { -1,1} |
| | URL Abnor | This parameter indicates whether the identity field matches the domain in the URL; if not, the URL will be classified as phishing with a value of -1; otherwise, it will be classified as legitimate.<br>IF conditional rule $\begin{cases} \text{URL Has No Host Name} \rightarrow \text{ Phishing} \\ \text{ELSE} \rightarrow \text{ Legitimate} \end{cases}$ | { -1,1} |
| HTML JavaScript-based Features | Domain Redirection | This parameter tracks the number of redirections a website has and is an indication that the website has been redirected. It is observed that legitimate websites are redirected only once. However, phishing sites that have this feature tend to redirect visitors on at least four occasions.<br>IF conditional rule $\begin{cases} \text{Number of Page Redirects } \leq 1 \rightarrow \text{ Legitimate} \\ \text{Number of Page Redirects } \geq 2 \text{ And } < 4 \rightarrow \text{ Suspicious} \\ \text{ELSE} \rightarrow \text{ Phishing} \end{cases}$ | {1,0, -1} |
| | Customizing the Status B | This parameter checks whether a valid link is shown in the status bar when the mouse is moved over it. Scammers may use JavaScript to generate a phony URL for the status bar. This may be confirmed by inspecting the "On Mouseover" section of the webpage's code base and observing whether the status bar is modified. If it changes, phishing is given a rating of -1; otherwise, it is given a rating of 1.<br>IF conditional rule $\begin{cases} \text{OnMouseOver Transforms Status Bar} \rightarrow \text{ Phishing} \\ \text{Status Bar Remains} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |
| | Disabling Ri Click | This parameter, known as "Right-Click Disable," serves as an indicator of whether the ability to right-click has been blocked. Criminals often utilize JavaScript to prevent right clicking in order to conceal the source code of a website and make it easier to manipulate legitimate pages. It is similar in function to "Use On Mouseover to hide the link." The parameter is determined by searching the source code for the occurrence of "event. Button == 2" and evaluating whether right clicking has been disabled. If it has been disabled, the value of the parameter is -1, otherwise it is 1.<br>IF conditional rule $\begin{cases} \text{Right Click Disabled} \rightarrow \text{ Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |
| | Using Pop-u Window | This parameter, known as "Personal Information Requirement," serves as an indicator of whether a website prompts users to provide personal data collected via pop-up windows or form submissions. If such a prompt is present, it is considered a phishing attempt and the parameter value is set to -1. If no such prompt is present, the website is classified as legitimate, and the parameter value is 1.<br>IF conditional rule $\begin{cases} \text{Text Fields in Popup Window} \rightarrow \text{ Phishing} \\ \text{ELSE} \rightarrow \text{ Legitimate} \end{cases}$ | { -1,1} |
| | IFrame Redirection | The IFrame is an HTML element that allows for the display of another webpage within the current page. This parameter, known as "IFrame Usage," serves as an indicator of whether a website utilizes an IFrame without visible borders to deceive the viewer. Typically, criminals use the IFrame badge to conceal their websites. If the website is found to be using this tactic, It is classified as a phishing attempt, and the value of the parameter is set to -1. If no such IFrame usage is detected, the website is deemed valid, and the parametric value is changed to 1.<br>IF conditional rule $\begin{cases} \text{Using iframe} \rightarrow \text{ Phishing} \\ \text{ELSE} \rightarrow \text{ Legitimate} \end{cases}$ | { -1,1} |

| | | | |
|---|---|---|---|
| Domain-based Features | Age of Dom | This parameter indicates that the minimum legitimate domain age is 6 months, so if the site is less than 6 months old, it will be classified as phishing with a parameter value of -1; otherwise, it will be classified as legitimate.<br>IF conditional rule $\begin{cases} \text{DNS Age } \geq 6 \text{ months} \rightarrow \text{Legitimate} \\ \text{ELSE} \rightarrow \text{Phishing} \end{cases}$ | { -1,1} |
| | DNS Recor | This parameter, known as "DNS Record Verification," serves as an indicator of the validity of a website's DNS record. If a website's DNS record is empty or cannot be located, it is classified as a phishing attempt and the parametric value is changed to -1. The parametric value is set to 1 if a valid DNS record is present, indicating that the website is legitimate.<br>IF conditional rule $\begin{cases} \text{no Domain DNS Record} \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |
| | Webpage Tr | This parameter, known as "Website Popularity," serves as an indicator of a website's level of traffic by measuring the number of visitors and pages viewed. Due to the typically shorter lifespan of phishing sites, a low level of traffic may be an indication that a website should be considered suspect. However, it should be noted that this parameter should be considered in conjunction with other factors as it is not a conclusive indicator of a website's legitimacy, the Alexa database may not recognize them. If the Alexa database rank is less than 100,000, the site will be considered authentic with parameter value 1, and if it is greater than 100,000, the site will be considered suspicious with parameter value 0. In addition, if phishing is not mentioned, a parameter value of -1 is assigned.<br>IF conditional rule $\begin{cases} \text{Webpage Traffic} < 100{,}000 \rightarrow \text{Legitimate} \\ \text{Webpage Traffic} > 100{,}000 \rightarrow \text{Suspicious} \\ \text{ELSE} \rightarrow \text{Phish} \end{cases}$ | {1,0, -1} |
| | PageRank | "PageRank" indicates a web page's importance in search engines. Higher values indicate a more important website. High-quality links determine a website's importance. External and internal links determine a website's PageRank. 95% of phishing websites have no PageRank, while 5% have "0.2". Thus, the parameter value is set to -1 if a website's PageRank is below 0.2, indicating phishing. If the PageRank is above 0.2, the website is legitimate, and the parameter value is 1.<br>IF conditional rule $\begin{cases} \text{WebpageRank} < 0.2 \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |
| | Index on Go | This parameter, known as "Google Indexing," serves as an indicator of whether a website has been indexed by the search engine giant, Google. Many phishing web pages are really only available for a brief amount of time; consequently, Google may not index many of these pages. If Google has not indexed the website, it is considered a phishing attempt, and the parametric value is changed to -1. If Google has indexed the website, it is deemed legitimate, and the parametric value is changed to 1.<br>IF conditional rule $\begin{cases} \text{Google Indexed Website} \rightarrow \text{Legitimate} \\ \text{ELSE} \rightarrow \text{Phishing} \end{cases}$ | { -1,1} |
| | Webpage Li Count | This parameter, known as "Link Count," indicates the credibility of a website by measuring the number of hyperlinks that point to the page. The number of links pointing to a website may be indicative of its credibility. The parameter value is set to 1 because a greater number of links pointing to a website would indicate that the website is legitimate. If there are no or few links (less than 3) referring to a website, it is deemed suspicious with a parameter value of 0. With a parameter value of -1, a website with fewer than one link is considered to be a phishing attempt.<br>IF conditional rule $\begin{cases} \text{Webpage Link Count} = 0 \rightarrow \text{Phishing} \\ \text{Links to Website} > 0 \text{ and } \leq 2 \rightarrow \text{Suspicious} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | {1,0, -1} |
| | Statistical-Reports | Statistical reports offer valuable information about both phishing and legitimate websites, as well as the changes in the number of newly created web pages. Several organizations, such as Phish Tank, regularly generate a variety of statistical reports about phishing websites. These reports can be used to ascertain if a website is an attempt at phishing. The value of the parameter is set to -1 if the website is determined to be a phishing attempt. If the website is determined to be valid, the parametric value is changed to 1.<br>IF conditional rule $\begin{cases} \text{Top Phishing IPs/DNS Host} \rightarrow \text{Phishing} \\ \text{ELSE} \rightarrow \text{Legitimate} \end{cases}$ | { -1,1} |

The characteristics of the dataset can be split into four categories, as shown in Table 3. These categories include Address bar-based features (12 features), Domain-based features (7 features), Abnormal characteristics (6 features), and HTML and JavaScript-based features (5 features). These characteristics cover a variety of web elements and properties that can be used to identify potential indicators of phishing or other types of malicious websites. By examining these web elements and properties and use this information to assess the overall security and legitimacy of a website. While no single data point can definitively identify whether a website is legitimate or not, these characteristics can provide a more complete picture of a website's characteristics and help identify potential red flags.

## 4. Results and Discussion

### *4.1 Feature Selection*

Two techniques were used: the correlation matrix with heatmap and the feature selection by wrapper method:

### *4.1.1 Feature Correlation*

A correlation matrix is used to display the correlation between large numbers of variables present in the dataset. It is a tool for analyzing and summarizing data, allowing for the understanding of the relationship between variables, making informed decisions, and improving the predictability of the model by reducing noise. The matrix's rows and columns stand for independent variables, and each cell displays the correlation between the two variables in question. The strength of a link between two variables can be quantified by their correlation coefficient [31].

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} \tag{5}$$

Equation (5), Pearson's correlation coefficient is a statistical measure used to determine the strength and direction of the linear relationship between two variables. The range of the correlation coefficient varies from -1 to +1, where a value of -1 indicates a perfectly negative correlation, a value of +1 indicates a perfectly positive correlation, and a value of 0 indicates no correlation.

In the context of figure 5, a matrix of 31 x 31 is plotted to show the correlation between 31 different features. The color-fill for each cell is based on the value of Pearson's correlation coefficient between the corresponding two features. A lighter color in a cell indicates a strong positive correlation between the two features, with a value close to +1.0. On the other hand, a darker color in a cell represents a strong negative correlation between the two features, with a value close to -1.0. The plot provides insights into the relationship between different features and how they are related to each other. By examining the correlation matrix, one can identify which features are strongly related to each other and which ones have no correlation. This information is valuable for data analysis and modeling as it helps in selecting the most relevant features for the analysis, avoiding multicollinearity, and improving the accuracy of the model.

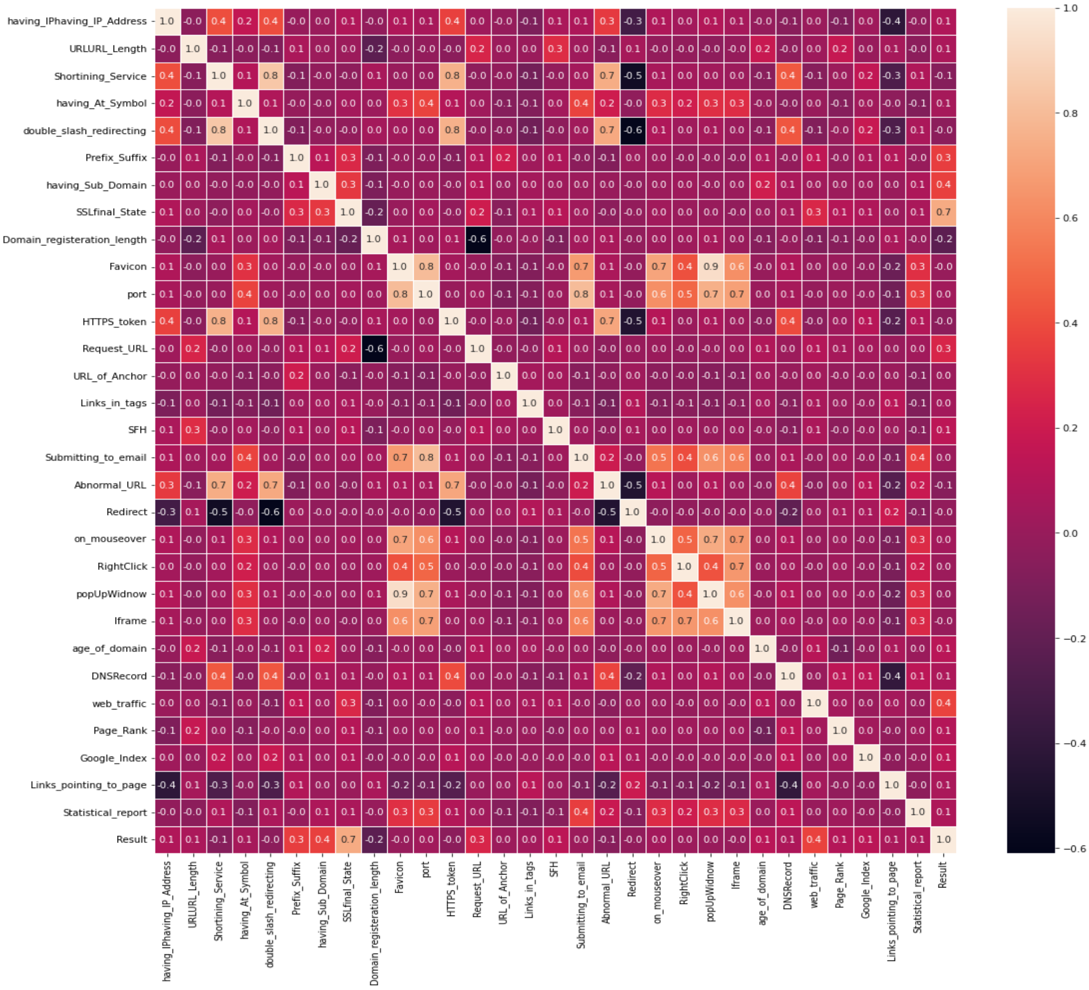


Fig. 5. Correlation between all the variables.

Table 4. Highly positively correlated features

| Features | | Corr. Coeff. |
|---|---|---|
| Popup Window | Favicon | 0.9 |
| HTTP Token | // redirecting | 0.8 |
| Favicon | Port | 0.8 |
| HTTP Token | Shortening Service | 0.8 |
| Shortening Service | // redirecting | 0.8 |
| Port | Submitting to email | 0.8 |

Table 4 shows that features are highly positively correlated, such as popup windows and favicons, with a correlation coefficient of (0.9). This means that for websites in which the favicon is packed from backlinks, the pop-up window refers to the structural a text field and HTTP Token with Double Slash and Shortening Service with correlation coefficient the majority of the time (0.8). I.e.

Table 5. Highly negatively correlated features

| Features | | Corr. Coeff. |
|---|---|---|
| // redirecting | Redirect | -0.6 |
| Domain Reg. length | Request URL | -0.6 |
| Shortening Service | Redirect | -0.5 |
| HTTP Token | Redirect | -0.5 |
| Abnormal URL | Redirect | -0.5 |
| DNS Record | Links pointing to the page | -0.4 |

Table 5 shows that features are highly negatively correlated, such as redirect and double slash redirecting with a correlation coefficient of (-0.6), and shortening service (-0.5), This is the case when one feature indicates phishing, and the other does not. Therefore, we eliminate features that have a negative impact on the result.

### *4.1.2 Feature Selection by Wrapper Method (FSWM).*

The process of identifying and selecting the most relevant features for training a ML model is known as FS, attribute selection, or variable selection [32]. It is an essential step in the modeling process because it aims to improve the model's performance and efficiency by removing unnecessary characteristics and preserving only the essential ones, while not impacting the learning performance [33,34].

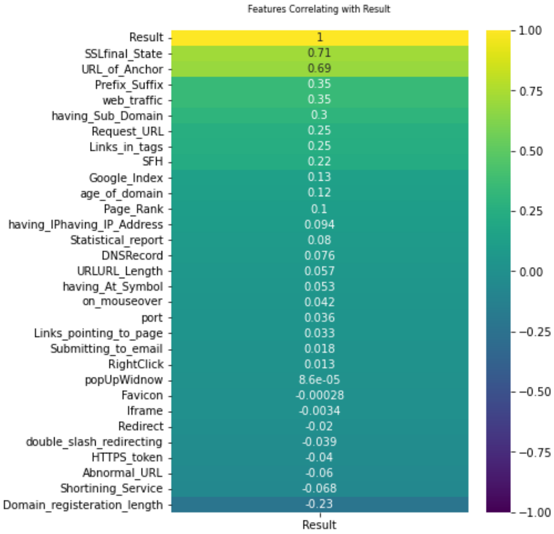


Fig. 6. Correlations with Result.

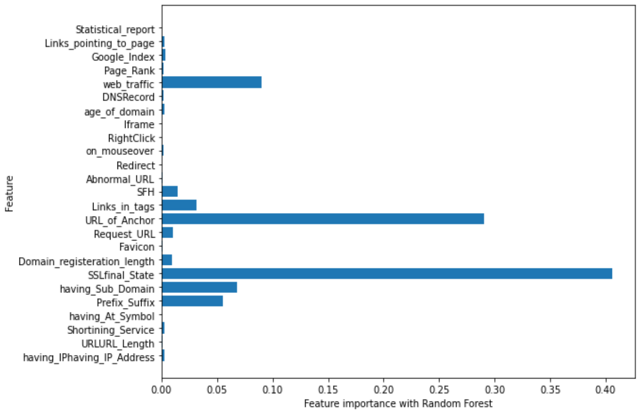


Fig. 7. Top Important Features with Random Forest.

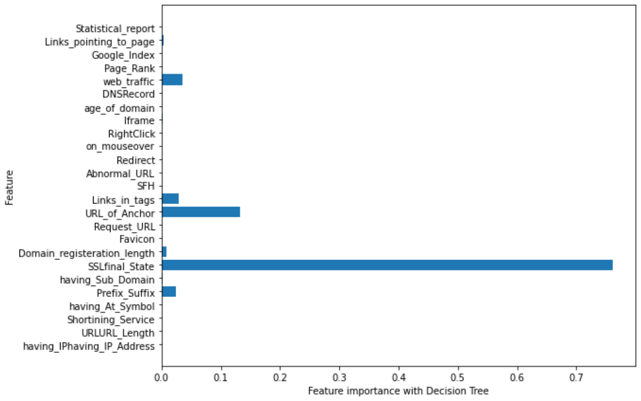


Fig. 8. Top Important with Decision Tree Features.

As we can see in figure 6, Pearson's correlation of the independent variables with the result. This study concentrates on wrapper-based feature selection methods and presents Recursive Feature Elimination (RFE) as an algorithm that follows a wrapper-style approach and uses filter-based feature selection internally. The wrapper model assumes classification algorithms and executes cross-validation to detect crucial features [35].

This process continues until the best features are selected as a subset, the features are trained using Extra Trees Classifier (ETC) [36], Logistic Regression, and Random Forest Classifier, these methods allowed us to construct a number of subgroups. Figure 7 and figure 8 show the top features of DT and RF, respectively.

To determine the optimal number of characteristics that are overfitted with an RFE to score subsets' unique features and choose the best classification combination of features [37]. We combine this with cross-validation using RFECV to automatically select the ideal number of features. Figure 9 uses RFECV to identify the minimum number of characteristics required to accurately classify a website as either phishing or legitimate [38]. Using the Extra Trees Classifier (ETC) to select the four best-ranked features for identifying the type of website. These features are SSL final State, URL of Anchor, Prefix Suffix, and Having Subdomain as they are highly correlated with the classification result, as shown in figure 10.

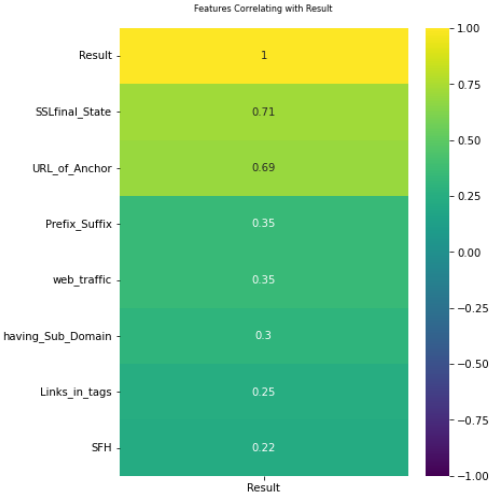


Fig. 9. Least optimal number of features with RFECV.

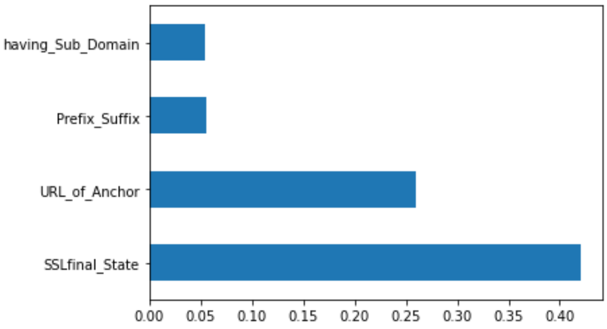


Fig. 10. The importance of features for better visualization.

Where the advantage (SSL_Final state) is superior. This means that the user must initially determine if the website has a secure connection or not. Then check the request URL If the <a> tags and the website have distinct domains, we categorize them as potentially fraudulent, or phishing based on the quantity and proportion of dashes and dots present within the domain. Due to the fact that an attribute's relative value rises the more frequently it is used to make crucial decisions, the relative value of an attribute increases as its frequency of use increases [39], The use of sub-domain registration services to create fake websites has become widespread, according to a report [40].

Based on the feature selection, it can be inferred that features pertinent to the address and features related to abnormalities are crucial for identifying and detecting phishing websites. Specifically, if either of these features are unique, the website is more likely to be legitimate. Conversely, if they are indicative of phishing, the website is more likely to be fraudulent.

### *4.2 Using Machine Learning Algorithms*

The performance was evaluated by applying several classifiers for training and testing: DT, RF, SVM, and MLP. The effectiveness of machine learning algorithms by displaying various metrics, such as accuracy, precision, recall, and F1-score, for each one of the 30 features.

Table 6. Performance results using 30 features.

| Classification model | Accuracy | Precision | Recall | F1- score |
|---|---|---|---|---|
| Decision Tree | 96.2% | 96.45 | 96.91 | 96.69 |
| Support Vector Machine | 93.4% | 92.65 | 95 | 93.8 |
| Random Forest | 96.56% | 96.24 | 97.51 | 96.87 |
| Multilayer Perceptron | 98.8% | 96.74 | 97.46 | 97.15 |

The results of the phishing detection dataset are summarized in Table 6 for all classifiers on the same data and under the same circumstances, including DT, SVM, RF, and MLP. The performance of the MLP classifier is determined to be superior to that of conventional ML models.

The model's effectiveness was evaluated by implementing batch normalization to enhance the precision and efficiency of the classification process. Additionally, it was found that MLP classifiers, which are components of neural networks, exhibit superior performance compared to other techniques for classifying, including DT, RF, and SVM.

Table 7. Analogy to prior studies using the same data

| References | Model | Accuracy |
|---|---|---|
| Ref. [41] | Random Forest (RF)<br>Decision Tree (DT)<br>Naive Bayes<br>Logistic Regression | 83%<br>80%<br>78%<br>78% |
| Ref. [42] | Multilayer Perceptron (MLP) | 97.4% |
| Ref. [43] | Multilayer Perceptron (MLP)<br>SVM<br>DT<br>Random Forest (RF) | 96.65%<br>91.81%<br>95.44%<br>95.60% |
| Ref. [44] | Multilayer Perceptron (MLP)<br>SVM<br>Random Forest (RF)<br>DT | 97%<br>95%<br>93%<br>91% |
| The Proposed Method | Multilayer Perceptron (MLP)<br>Decision Tree (DT)<br>Support Vector Machine (SVM)<br>Random Forest (RF) | 98.8%<br>96.2%<br>95.4%<br>96.56% |

Table 7 provided depicts a comparison between the proposed method and previous research utilizing the same dataset and same circumstances [45]. The proposed model has produced the highest accuracy using MLP, which is 98.8%, and also the highest accuracy using other classifiers, which are 96.2%, 95.4% and 96.56% for DT, SVM, and Random Forest (RF), respectively as shown in figure 11. The previous research includes four different models: Random Forest (RF), Decision Tree (DT), Naive Bayes, and Logistic Regression, with accuracies ranging from 78% to 83% [41]. Additionally, another study compared several classifiers, including Multilayer Perceptron (MLP), SVM, DT, and RF,

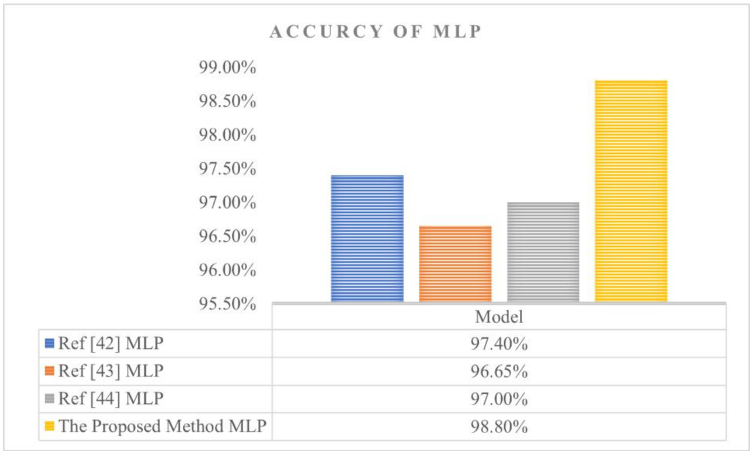


Fig. 11. Comparison of ML classifier for the proposed method with other research.

with accuracies ranging from 91.81% to 96.65% [43]. Finally, another comparative analysis considered MLP, SVM, RF, and DT, with accuracies ranging from 91% to 97% [44], where the proposed model has produced the highest accuracy using MLP, which is 98.8%, and also the highest accuracy using other classifiers, which are 96.2%, 95.4% and 96.56% for DT, SVM, and Random Forest (RF), respectively, as shown in figure 12.

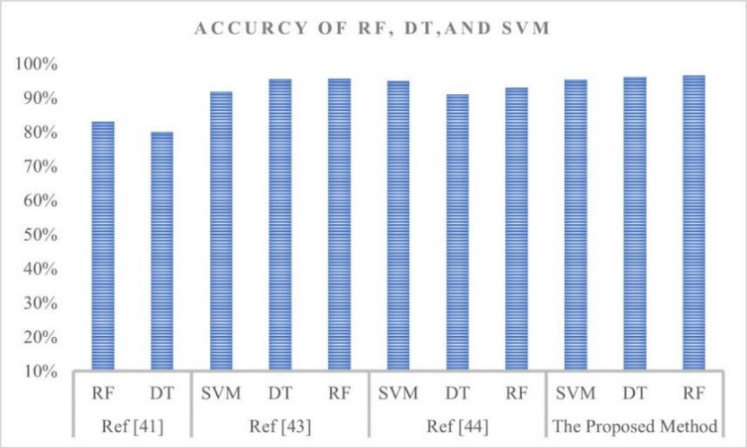


Fig. 12. Comparison of MLP for the proposed method with other research.

Table 8. Analogy to prior studies using different data and models

| References | Model | Accuracy | Features |
|---|---|---|---|
| Ref. [46] | DT<br>Na ïve Bayesian (NB)<br>SVM<br>NN | 91.5 %<br>86.69 %<br>86.14 %<br>84.87 % | A dataset of 1,353 real-world URLs that could be classified as either legitimate, suspicious, or phishing was used to evaluate the classifiers [47]. |
| Ref. [48] | XGBoost | 96.6% | A dataset of 11,430 total URLs from Kaggle includes 86 features, trained their data with DT, RF, XGBoost, MLP, K-Nearest Neighbors, Naive Bayes, AdaBoost, and Gradient Boosting, with XGBoost producing the highest accuracy. |
| Ref. [49] | Generative Adversarial Network (GAN) | 94% | A dataset of 24,084 and 11,267 URLs from Amazon, Phish Tank, Basic and workstation. Utilizes a generator network that produces both authentic and manufactured phishing characteristics to train a discriminator network. |
| Ref. [50] | Convolutional Neural Network (CNN)<br>Long Short-Term Memory (LSTM) | 93.28% | A dataset of 1 Million URLs and 10,000 images, using LSTM and CNN to construct a classification model |
| The Proposed Method | MLP<br>DT<br>SVM<br>Random Forest (RF) | 98.8%<br>96.2%<br>95.4%<br>96.56% | A dataset of 11055 websites using four of the most widely used ML, and DL techniques: DT, RF, SVM, and MLP. |

Table 8 provided a comparison of the performance of different machine learning (ML) and deep learning (DL) models in detecting phishing websites in various studies with different datasets [51].

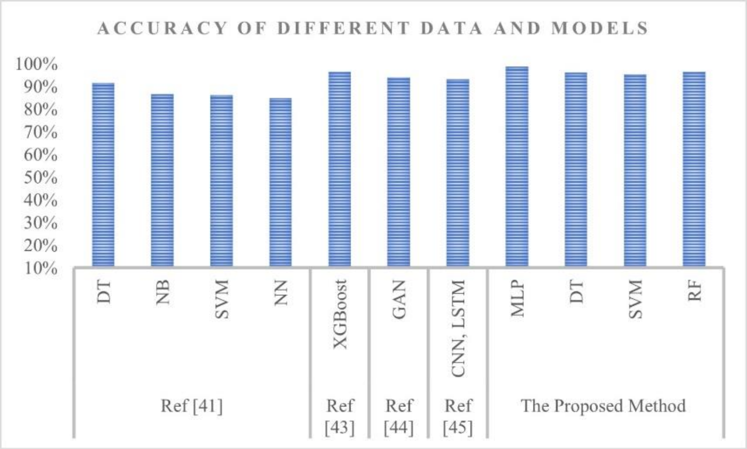


Fig. 13. Comparison of the proposed method with other research.

The first row presents the accuracy results of DT, Naïve Bayesian (NB), SVM, and NN models, which were evaluated on a dataset of 1,353 real-world URLs categorized as legitimate, suspicious, or phishing. The highest accuracy of 91.5% was achieved by the DT model. The second row shows the performance of different models, including DT, RF, XGBoost, MLP, K-Nearest Neighbors, Naive Bayes, AdaBoost, and Gradient Boosting, on a dataset of 11,430 URLs with 86 features. The XGBoost model achieved the highest accuracy of 96.6%. The third row describes the performance of a Generative Adversarial Network (GAN) model that was trained on a dataset of 24,084 and 11,267 URLs from Amazon, Phish Tank, Basic, and workstation. The GAN model utilized a generator network that produced both authentic and manufactured phishing characteristics to train a discriminator network and achieved an accuracy of 94%. The fourth row shows the results of using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models on a dataset of 1 million URLs and 10,000 images. The highest accuracy achieved was 93.28%.

Finally, the proposed method achieved the highest accuracy of 98.8% using MLP, and 96.2%, 95.4%, and 96.56% using DT, SVM, and RF models, respectively. The proposed method was evaluated on a dataset of 11,055 websites using four widely used ML and DL techniques, namely DT, RF, SVM, and MLP as shown in figure 13.

## 5. Conclusion

This investigation's main objective was to separate legitimate web pages from phishing software web pages using a dataset of 11055 with four of the most widely used ML techniques: DT, RF, SVM, and MLP. The effectiveness of these algorithms is evaluated using four key parameters: precision, recall, F1-score, and accuracy. In accordance with the results, deep neural networks using Multi-Layer Perceptron Neural Networks (MLPNN) outperformed the rest of all the methods. The results of this analysis demonstrate that the phishing detection accuracy of the proposed method is 98.8%. Additionally, the study utilized feature selection to gain insights into website characteristics and the specific features that significantly impact the classification of a website as phishing or legitimate. The wrapper method is used to calculate the significance of features, and a correlation matrix with a heatmap is generated to identify the most critical features, then using the ExtraTreesClassifier (ETC) algorithm is employed to decrease the number of characteristics in a set of data, and they are SSL final State, URL of Anchor, Prefix Suffix, and Having Subdomain for the purpose of identifying phishing websites for individuals with little to no technical expertise.

The suggested supervised learning employs a classifier to evaluate a set of phishing frauds that can correctly categorize unseen data, thus effectively identifying new phishing sites.

In future studies, we will expand the dataset used for training and testing the models to include more diverse types of phishing websites, as well as legitimate websites with similar features. This could help to improve the generalization ability of the models and make them more robust to new types of phishing attacks, comparing MLP to other deep learning techniques like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to see which approach works best for this problem, and exploring the use of ensemble learning techniques to combine the predictions of multiple models trained with different algorithms or parameters. This could potentially lead to further improvements in the accuracy and robustness of phishing detection.

## References

[1] A. Abuzuraiq, M. Alkasassbeh, and M. Almseidin, “Intelligent methods for accurately detecting phishing websites,” 2020 11th International Conference on Information and Communication Systems (ICICS), 2020. doi:10.1109/icics49469.2020.239509.

[2] A. K. Jain and B. B. Gupta, “A machine learning based approach for phishing detection using hyperlinks information,” Journal of Ambient Intelligence and Humanized Computing, vol. 10, no. 5, pp. 2015–2028, 2018. Doi: 10.1007/s12652-018-0798-z.

[3] APWG GA, Manning R (2020) APWG Phishing Reports.

[4] "2nd quarter 2022 - docs.apwg.org." [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q2_2022.pdf. [Accessed: 26-Dec-2022].

[5] A. K. Dutta, "Detecting phishing websites using Machine Learning Technique," PLOS ONE, vol. 16, no. 10, 2021. doi: 10.1371/journal.pone.0258361.

[6] S. A. Khan, W. Khan, and A. Hussain, "Phishing attacks and websites classification using machine learning and multiple datasets (a comparative analysis)," Intelligent Computing Methodologies, pp. 301–313, 2020. doi:10.1007/978-3-030-60796 - 8_26.

[7] P. Saravanan and S. Subramanian, "A framework for detecting phishing websites using GA based feature selection and ARTMAP based website classification," Procedia Computer Science, vol. 171, pp. 1083–1092, 2020. doi: 10.1016/j.procs.2020.04.116.

[8] M. G. HR, A. MV, G. P. S, and V. S, "Development of anti-phishing browser based on Random Forest and rule of extraction framework," Cybersecurity, vol. 3, no. 1, 2020. doi:10.1186/s42400-020-00059-1.

[9] Shantanu, B. Janet, and R. Joshua Arul Kumar, "Malicious URL detection: A comparative study," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021. doi:10.1109/icais50930.2021.9396014.

[10] A. Mandadi, S. Boppana, V. Ravella, and R. Kavitha, "Phishing website detection using Machine Learning," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), 2022.

[11] L. Bustio-Mart ́nez, M. A. Álvarez-Carmona, V. Herrera-Semenets, C. Feregrino-Uribe, and R. Cumplido, "A lightweight data representation for phishing urls detection in iot environments," Information Sciences, vol. 603, pp. 42–59, 2022. doi: 10.1016/j.ins.2022.04.059.

[12] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from urls," Expert Systems with Applications, vol. 117, pp. 345–357, 2019. doi: 10.1016/j.eswa.2018.09.029.

[13] C. Catal, G. Giray, B. Tekinerdogan, S. Kumar, and S. Shukla, "Applications of deep learning for phishing detection: A systematic literature review," Knowledge and Information Systems, vol. 64, no. 6, pp. 1457–1500, 2022. doi:10.1007/s10115-022-01672-x.

[14] J. Anitha and M. Kalaiarasu, "A new hybrid deep learning-based phishing detection system using MCS-Dnn Classifier," Neural Computing and Applications, vol. 34, no. 8, pp. 5867–5882, 2022. doi:10.1007/s00521-021-06717-w.

[15] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, "A hybrid DNN–LSTM model for detecting phishing urls," Neural Computing and Applications, 2021. doi:10.1007/s00521-021-06401-z.

[16] L. Lakshmi, M. P. Reddy, C. Santhaiah, and U. J. Reddy, "Smart phishing detection in web pages using supervised deep learning classification and Optimization Technique adam," Wireless Personal Communications, vol. 118, no. 4, pp. 3549–3564, 2021. doi:10.1007/s11277-021-08196-7.

[17] T. Li, G. Kou, and Y. Peng, "Improving malicious urls detection via feature engineering: Linear and nonlinear space transformation methods," Information Systems, vol. 91, p. 101494, 2020. doi:10.1016/j.is.2020.101494.

[18] M. Alsaedi, F. Ghaleb, F. Saeed, J. Ahmad, and M. Alasli, "Cyber threat intelligence-based malicious URL detection model using ensemble learning," Sensors, vol. 22, no. 9, p. 3373, 2022. doi:10.3390/s22093373.

[19] A. Almomani, M. Alauthman, M. T. Shatnawi, M. Alweshah, A. Alrosan, W. Alomoush, B. B. Gupta, B. B. Gupta, and B. B. Gupta, "Phishing website detection with semantic features based on machine learning classifiers," International Journal on Semantic Web and Information Systems, vol. 18, no. 1, pp. 1–24, 2022.

[20] L. Gajic, D. Cvetnic, M. Zivkovic, T. Bezdan, N. Bacanin, and S. Milosevic, "Multi-layer perceptron training using hybridized bat algorithm," Computational Vision and Bio-Inspired Computing, pp. 689–705, 2021. doi:10.1007/978-981-33-6862-0_54.

[21] J. Zhang, C. Li, Y. Yin, J. Zhang, and M. Grzegorzek, "Applications of artificial neural networks in Microorganism Image Analysis: A comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer," Artificial Intelligence Review, 2022. doi:10.1007/s10462-022-10192-7.

[22] S. Abirami and P. Chitra, "Energy-efficient edge based real-time healthcare support system," Advances in Computers, pp. 339–368, 2020. doi: 10.1016/bs.adcom.2019.09.007.

[23] Y. Xiao, W. Huang, and J. Wang, "A random forest classification algorithm based on dichotomy rule fusion," 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2020. doi:10.1109/iceiec49280.2020.9152236.

[24] L. Xie, Z. Li, Y. Zhou, Y. He, and J. Zhu, "Computational diagnostic techniques for electrocardiogram signal analysis," Sensors, vol. 20, no. 21, p. 6318, 2020. doi:10.3390/s20216318.

[25] R. M. Panda and B. S. Daya Sagar, "Decision tree," Encyclopedia of Mathematical Geosciences, pp. 1–7, 2022. doi:10.1007/978-3-030-26050-7_81-2.

[26] C. El Morr, M. Jammal, H. Ali-Hassan, and W. El-Hallak, "Decision trees," International Series in Operations Research &amp; Management Science, pp. 251–278, 2022. doi:10.1007/978-3-031-16990-8_8.

[27] B. Kamiel, T. Akbar, Sudarisman, and Krisdiyanto, "Cavitation detection of centrifugal pumps using SVM and statistical features," Lecture Notes in Mechanical Engineering, pp. 1–9, 2022. doi:10.1007/978-981-19-0867-5_1.

[28] M.-T. Wu, "Confusion matrix and minimum cross-entropy metrics-based motion recognition system in the classroom," Scientific Reports, vol. 12, no. 1, 2022. doi:10.1038/s41598-022-07137-z.

[29] C. V. Gonzalez Zelaya, "Towards explaining the effects of data preprocessing on machine learning," 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019.

[30] UCI Machine Learning Repository: Phishing Websites Data Set [Online]. Available: https://archive.ics.uci.edu/ml/datasets/phishing+websites. [Accessed: 30-Dec-2022].

[31] J. Moedjahedy, A. Setyanto, F. K. Alarfaj, and M. Alreshoodi, "CCrFS: Combine correlation features selection for detecting phishing websites using machine learning," Future Internet, vol. 14, no. 8, p. 229, 2022. doi:10.3390/fi14080229.

[32] E. O. Abiodun, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, and R. S. Alkhawaldeh, "A systematic review of emerging feature selection optimization methods for optimal text classification: The Present State and Prospective Opportunities," Neural Computing and Applications, vol. 33, no. 22, pp. 15091–15118, 2021.

[33] E. O. Abiodun, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, and R. S. Alkhawaldeh, “A systematic review of emerging feature selection optimization methods for optimal text classification: The Present State and Prospective Opportunities,” Neural Computing and Applications, vol. 33, no. 22, pp. 15091–15118, 2021.

[34] J. Brank, D. Mladenić, M. Grobelnik, H. Liu, D. Mladenić, P. A. Flach, G. C. Garriga, H. Toivonen, and H. Toivonen, “Feature selection,” Encyclopedia of Machine Learning, pp. 402–406, 2011.

[35] I. Attoui, B. Oudjani, N. Boutasseta, N. Fergani, M.-S. Bouakkaz, and A. Bouraiou, “Novel predictive features using a wrapper model for rolling bearing fault diagnosis based on vibration signal analysis,” The International Journal of Advanced Manufacturing Technology, vol. 106, no. 7-8, pp. 3409–3435, 2020.

[36] M. G. Lanjewar, J. S. Parab, A. Y. Shaikh, and M. Sequeira, “CNN with machine learning approaches using extratreesclassifier and MRMR feature selection techniques to detect liver diseases on cloud,” Cluster Computing, 2022.

[37] S. O. Abdulsalam, A. A. Mohammed, J. F. Ajao, R. S. Babatunde, R. O. Ogundokun, C. T. Nnodim, and M. O. Arowolo, “Performance evaluation of ANOVA and RFE algorithms for classifying microarray dataset using SVM,” Information Systems, pp. 480–492, 2020.

[38] H. H. Luong, N. T. Phan, T. T. Duong, T. M. Dang, T. D. Nguyen, and H. T. Nguyen, “Dimensionality reduction on metagenomic data with recursive feature elimination,” Complex, Intelligent and Software Intensive Systems, pp. 68–79, 2021.

[39] N. Tabassum, F. F. Neha, M. S. Hossain, and H. S. Narman, “A hybrid machine learning based phishing website detection technique through dimensionality reduction,” 2021 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), 2021. doi: 10.1109/BlackSeaCom52164.2021.9527806.

[40] “1st Quarter - APWG.” [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q1_2021.pdf. [Accessed: 15-Jan-2023].

[41] B. Geyik, K. Erensoy, and E. Kocyigit, “Detection of phishing websites from urls by using classification techniques on Weka,” 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021. doi: 10.1109/ICICT50816.2021.9358642.

[42] S. S. Motiur Rahman, T. Islam, and M. I. Jabiullah, “PhishStack: Evaluation of stacked generalization in phishing URLs detection,” Procedia Computer Science, vol. 167, pp. 2410–2418, 2020. doi: 10.1016/j.procs.2020.03.294.

[43] S. Al-Ahmadi and T. Lasloum, “PDMLP: Phishing detection using multilayer perceptron,” International Journal of Network Security &amp; Its Applications, vol. 12, no. 3, pp. 59–72, 2020.

[44] S. Wassan, C. Xi, N. Jhanjhi, and H. Raza, “A smart comparative analysis for secure electronic websites,” Intelligent Automation &amp; Soft Computing, vol. 29, no. 3, pp. 187–199, 2021. doi:10.32604/iasc.2021.015859.

[45] A. Niranjan, D. K. Haripriya, R. Pooja, S. Sarah, P. Deepa Shenoy, and K. R. Venugopal, “EKRV: Ensemble of KNN and Random Committee using voting for efficient classification of phishing,” Advances in Intelligent Systems and Computing, pp. 403–414, 2018. doi: 10.1007/978-981-13-1708-8_37.

[46] A. Kulkarni and L. L., “Phishing websites detection using machine learning,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, 2019. doi: 10.14569/IJACSA.2019.0100702.

[47] UCI Machine Learning Repository: Website Phishing Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Website+Phishing. [Accessed: 06-Feb-2023].

[48] S. Alrefaai, G. Ozdemir, and A. Mohamed, “Detecting phishing websites using machine learning,” 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2022. doi: 10.1109/HORA55278.2022.9799917.

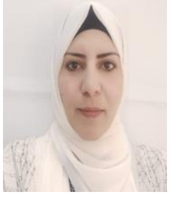
[49] P. Robic-Butez and T. Y. Win, “Detection of phishing websites using generative Adversarial Network,” 2019 IEEE International Conference on Big Data (Big Data), 2019. doi: 10.1109/BigData47090.2019.9006352.

[50] M. A. Adebowale, K. T. Lwin, and M. A. Hossain, “Deep learning with convolutional neural network and long short-term memory for phishing detection,” 2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2019. doi: 10.1109/SKIMA47702.2019.8982427.

[51] N. Tabassum, F. F. Neha, M. S. Hossain, and H. S. Narman, “A hybrid machine learning based phishing website detection technique through dimensionality reduction,” 2021 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), 2021. doi: 10.1109/ACCESS.2022.3151903.

## Authors’ Profiles

**Aya S. Noah** has completed her Bachelor of Science (BSc) in Computer Science & Pure Mathematics at Faculty of Science, Al-Azhar University, Cairo, Egypt, in 2019 with V. Good grade. She can be contacted at email: AyaNoah41167@azhar.edu.eg.



**Naglaa E. Ghannam** received the B.Sc. degree in science, in 2002, and the M.Sc. and Ph.D. degrees in computer science, in 2017 and 2020, respectively. She is currently a lecturer in computer science with the Mathematics Department, Faculty of Science, Al-Azhar University, Cairo, Egypt. She has published several research papers in the field of Machine learning, software engineering, software quality, and security. She is also a supervisor of some master’s theses. She can be contacted at email: naglaasaeed@azhar.edu.eg.

**Gaber A. Elsharawy,** Professor of computer science at Faculty of science, Al Azhar University. Ph.D. in Computer and Systems Engineering, Faculty of Engineering, Al Azhar University.

M.Sc.in computer system U.S. Air Force University, Air Force Institute of Technology (AFIT), Dayton, Ohio, USA. Author of many publications in the fields of Database management systems, artificial intelligent, modeling & simulation, and programming languages. He can be contacted ae email: Gaber.Ahmed@buc.edu.eg.

**Abeer S. Desuky** received the B.Sc. degree in science, in 2003, and the M.Sc. and Ph.D. degrees in computer science, in 2008 and 2012, respectively.

She is currently a Professor in computer science with the Mathematics Department, Faculty of Science, Al-Azhar University, Cairo, Egypt. She has published several research papers in the field of AI, machine learning, meta-heuristic optimization, and data mining and analysis. She is also a supervisor of some master’s and Ph.D. theses. She is a reviewer in many Scopus-indexed journals, such as IEEE Access and Peerj. She can be contacted at email: abeerdesuky@azhar.edu.eg.

**How to cite this paper:** Aya S. Noah, Naglaa E. Ghannam, Gaber A. Elsharawy, Abeer S. Desuky, "An Intelligent System for Detecting Fake Materials on the Internet", International Journal of Modern Education and Computer Science(IJMECS), Vol.15, No.5, pp. 42-59, 2023. DOI:10.5815/ijmecs.2023.05.04