Enhancing Phishing Website Detection via Feature Selection in URL-Based Analysis

Marwa A. Qasim, Nahla A. Flayh
E-mail: itpg.marwa.qasim@uobasrah.edu.iq, nahla.flayh@uobasrah.edu.iq
Department: Information Technology, Basrah University Basra, Iraq

Keywords: machine learning, detecting phishing, feature selection, regular expressions

Received: September 12, 2023

Detecting a phishing website accurately is crucial for ensuring the safety of online users, underscoring the importance of maintaining a secure digital environment. This research delves into the effectiveness of enhancing the detection of phishing websites by applying a new dataset generation method. The method involves the transformation of a pure dataset obtained from Mendeley, by the utilization of regular expressions to extract the important features so that a detection process can be performed correctly with high performance. Based on the proposed features, we selected the best machine-learning algorithm. We performed a rigorous evaluation using Three prominent machine learning algorithms: Decision Trees, Support Vector Machines (SVM), and Random Forests, achieving 0.96% for Decision Tree Accuracy, 0.97% for SVM Accuracy, and 0.98% for Random Forest Accuracy.One of the critical contributions of this research is the deliberate selection of features. We have leveraged regular expressions to create a feature set that captures salient aspects of URLs and optimizes the algorithms' detection capabilities. This research has examined how feature selection affects the performance of each algorithm, highlighting its strengths and uncovering its weaknesses.

Povzetek: Raziskava obravnava izboljšanje zaznavanja lažnih spletnih strani z novim načinom ustvarjanja podatkovnih zbirk in testiranjem različnih algoritmov strojnega učenja.

1 Introduction

Detecting phishing websites is a critical cybersecurity issue due to the sophistication of these attacks and their potential to compromise sensitive information. To protect individuals and organizations from financial loss, data breaches, and reputational damage, detecting and preventing phishing attacks is essential [1]. Consequently, it is essential to develop effective techniques for identifying, and mitigating phishing websites and combating this menace, With the advent of machine learning algorithms, security systems can recognize and prevent cyber-attacks more efficiently and effectively [2].

In the context of machine learning, it is a subset of artificial intelligence in which machines can learn from data, improve performance based on past experiences, and make predictions based on that data [3]. Machine learning can be divided into supervised, unsupervised, and reinforcement learning types. The focus of our research will be on supervised learning. In supervised learning, the training dataset consists of previous instances where both input and output values are known and labeled [4, 5]. One of the strategies that can enhance machine-learning performance is feature selection, so our research involved implementing a feature selection methodology, specifically a rule-based approach utilizing regular expressions. The approach is considered to be one of the most important methods for enhancing the efficiency of machine learning algorithms by refining the features to be used in the model.

2 Related work

Phishing detection is an essential component of cybersecurity, which aims to identify fraudulent attempts to deceive users into disclosing sensitive information. A variety of techniques, including machine learning algorithms and rule-based approaches, have been employed for effective detection. This section highlights essential machine learning techniques in the field of phishing detection and provides an overview of the state-of-the-art in phishing detection.

Bin B. Zhu, et al (2013) [6]: The authors evaluated the effectiveness of machine learning-based phishing detection methods using a secure website. 18 useful features were presented and tested for incorporation into

the detector based exclusively on the lexical and domain characteristics provided by the authors. Finding the appropriate combination of attributes has resulted in a detector with a detection rate higher than 98%. They employed support vector machines and Gaussian radial basis function algorithms. Phishing URLs were taken from the Taobao-phishing dataset, safe URLs were taken from the Yahoo! Directory, and well-known Chinese navigation sites were analyzed.

W. Fadheel, et al (2017) [7]: Datasets from the UCI machine learning repository were used in this research, including Domain, HTML, Address Bar, and URLs. The main contribution is represented by a comparative analysis of the impact of feature selection on the detection of phishing websites. A KMO test was applied in the research to evaluate the dataset using (LR) and (SVM) classification algorithms. A correlation matrix was used to analyze the performance of the test. LR with the KMO test achieved an accuracy of 91.68%, while SVM with the KMO test achieved an accuracy of 93.59%.

I. Tyagi, et al. (2018) [8]: The research uses machine learning algorithms to identify whether a website is legitimate or phishing. A URL is used to determine this. The most significant contribution is represented by the development of a new model, the Generalized Linear Model (GLM). Two different methods are combined in the model. Detecting phishing websites is most accurate when Random Forest and GLM are mixed with 98.4% accuracy.

Arun D. Kulkarni (2019) [9]: SVM, Nave Bayes, decision trees, and neural networks were evaluated in the research. It is used to detect phishing URLs. The research used a dataset containing 1353 URLs from the University of California, Irvine Machine Learning Repository. There are nine features associated with each URL. To evaluate the performance of the algorithms, two steps were taken. the process begins with the extraction of features from URLs. A model will be developed based on data from a training set in the second step. Based on the developed model, URLs will be classified. According to the results, the pruned decision tree produced the highest accuracy of 91.5%. It was followed by the Naive Bayes Classifier with 86.14 %, and the Neural Network with 84.87%.

S Premnath, et al. (2020) [10]: Using a sophisticated machine-learning framework, the research provides an in-depth analysis of phishing websites. the research used a dataset containing URLs from legitimate and phishing websites. Therefore, different machine learning algorithms could be evaluated to distinguish between phishing websites and legitimate websites. By combining the two phases of classification and phishing detection, the research contributed to the development of an efficient machine-learning framework. In the proposed system, five different machine learning classifiers are utilized to analyze URL features and detect phishing websites in an extremely accurate manner (Random Forest, Logistic Regression, Decision Tree, Nearest-Neighbor, and Support Vector Machine). Among machine learning classifiers, it has the highest accuracy according to the proposed system model. 91.4% accuracy was achieved by the Random Forest classifier.

A. Lakshmanarao, and P. Surya (2021) [11]: Using a dataset containing 11055 samples and 30 features of phishing websites from UCI's repository. Various machine learning techniques, including decision trees, AdaBoost, support vector machines (SVM), and random forests, were used to analyze specific features such as port, web traffic, URL length, URL_of_Anchor, and IP address. According to the research, the most effective method of detecting phishing websites was determined. PA1 and PA2 were introduced as part of the research. A 97% accuracy rate was achieved by these algorithms.

M Abutaha, et al. (2021) [12]: A method for detecting phishing attacks is presented using URL lexical analysis and machine learning classifiers. A variety of machine learning models were trained and tested on a variety of feature sets. It appears that the used approach is beneficial in phishing attacks. Web requests' headers contain URLs that are used for detection and prevention. Moreover, machine learning techniques have also proven effective in the area of security. The dataset used consisted of 1056937 labeled URLs (phishing and legitimate) and 14 different features. Different types of classifiers were evaluated, including gradient boosting, Random Forest, Support Vector Machine (SVM), and Neural Networks. Based on the results, SVM was the most accurate at 99.89% in detecting the URLs analyzed. Moreover, the neural network had the lowest accuracy score of all the classifiers, coming in at approximately 97%.

N. Choudhary b, S. Jain, K. Jain (2022) [13]: URL attributes are the focus of the research. The dataset used in the research was obtained from both Kaggle and Phishtank. The researchers used a hybrid approach that combined Principal Component Analysis (PCA), Support Vector Machines (SVM), and Random Forest algorithms to reduce the dataset's dimensionality while maintaining all relevant data. A higher accuracy rate of 96.8% was obtained with this method as compared to other techniques.

S. Arvind Anwekar, and V. Agrawal (2022) [14]: According to the authors, the research focused on extracting features from URLs. Several features were considered, including the SSL certificate's age, the

anchor's URI, the IFRAME, and the website's ranking. The total number of phishing URLs collected from Phish-Tank was 19653. The total number of benign URLs collected from Alexa was 17058. The authors developed a method for detecting phishing websites using randomly generated trees (RF), decision trees (DT), and support vector machines (SVM). The performance of the classifier also improved with the addition of more training data. As a result of splitting the dataset with 90 % training and 10% testing, it achieved a high detection accuracy of 97.14% and a low false positive rate of 3.14 percent.

A Prathap, et al (2023) [15]: This method could be used to automate systems that are highly effective in combating website phishing. Furthermore, as a result of its effectiveness and efficiency, this research performs well in literature comparisons. SVM and random forest algorithms were used to classify and predict phishing attacks. Data was collected from phishing websites. The UC Irvine Machine Learning Repository database contains approximately 11,000 data points containing 30 features derived from website features. Random Forest classifiers achieve an accuracy of 89.63%, while SVM classifiers achieve an accuracy of 89.84%.

UB Penta, et al (2023) [16]: The purpose of this research is to identify phishing websites using machine learning methods such as Support Vector Machines (SVM), K Nearest Neighbors (KNN), and Naive Bayes (NB). Feature Extraction (FE) techniques were used to extract essential attributes from the Phish-Tank website, 10,000 phishing URLs, and 10,000 benign URLs. An approach based on URLs and an approach based on hyperlinks was used. The results of both FE approaches are used as inputs for the ML model. SVM achieved the highest accuracy score of 98.05%, while KNN achieved the lowest accuracy score of 95.67%.

Table 1: A summary table of related works shows the results, methodologies, and performance metrics of the research studies reviewed

Study	**Year**	**Methodology**	**Performance Metrics**	**Results**	**SOTA Lacks in feature selection**
[6]	2013	ML-based on linguistic and domain characteristics; SVM and Gaussian radial basis function algorithms	Detection rate > 98%	Effective attribute combination	Limited feature selection from linguistic and domain characteristics
[7]	2017	Feature selection impact analysis; KMO test; LR and SVM classification	LR: 91.68%, SVM: 93.59%	Importance of feature selection	Limited exploration of feature selection impact
[8]	2018	Generalized Linear Model (GLM) combining Random Forest;	GLM: 98.4%	Novel GLM model	Lacks explanation of feature selection
[9]	2019	SVM, Naive Bayes, decision trees, neural networks; 1353 URL dataset	Pruned decision tree: 91.5%, Naive Bayes: 86.14%, Neural Network: 84.87%	Comparative algorithm performance	Limited feature extraction and selection
. [10]	2020	Multiple ML classifiers (Random Forest, LR, Decision Tree, KNN, SVM); URL features analysis	Random Forest: 91.4%	High accuracy using ensemble methods	It is lacking in the selection of features and the utilization of machine learning algorithms
[11]	2021	Various ML techniques; 11055 samples and 30 features dataset	PA1 and PA2 algorithms: 97%	Specific feature analysis	Select features using ANOVA F-value and Mutual Information. These feature selection methods are valid, but just a subset.
[12]	2021	Phishing URL detection using linguistic analysis; Various ML models	SVM: 99.89%, Neural Network: ~97%	High accuracy of SVM	URL length, hostname length, and keywords are frequently used. These features may miss advanced phishing methods that obfuscate or manipulate real URLs.
[13]	2022	PCA, SVM, Random Forest; Hybrid approach for dimensionality reduction	96.8%	Dimensionality reduction impact	focuses on URL-only feature extraction, which may miss some phishing website details.
[14]	2022	Feature extraction from URLs; RF, DT, SVM; Training data size impact	Accuracy: 97.14%, False positive rate: 3.14%	Performance with more training data	Feature extraction from URLs, does not clarify how these features were selected or whether relevance was determined using feature selection approaches.
. [15]	2023	SVM and Random Forest; 11000 data points, 30 features dataset	Random Forest: 89.63%, SVM: 89.84%	Accuracy of classification	The report lacks an explanation of the process of feature selection.
[16]	2023	SVM, KNN, NB; Feature extraction from Phish-Tank URLs	SVM: 98.05%, KNN: 95.67%	Comparison of ML methods	The used features are informative, but they may not accurately depict the specific attributes of more current phishing assaults.

Comparing Results

The methodology employed in our study involves the utilization of regular expressions for the extraction of significant features from URLs. These features are then used as input for Decision Trees, Support Vector Machines (SVM), and Random Forests. The results obtained from our experiments indicate noteworthy accuracies of 0.968, 0.973, and 0.976 for Decision Trees, SVM, and Random Forests, respectively. Upon comparing our findings with the research studies listed in the table, it becomes apparent that our approach demonstrates competitive or even greater levels of accuracy. The significance of this observation is particularly notable within the domain of detecting phishing websites, where achieving a high level of accuracy is of utmost importance in ensuring strong security measures. The uniqueness of our feature selection method utilizing regular expressions is in its capacity to accurately capture complex URL patterns and structural attributes, which are crucial for the detection of phishing endeavors. The utilization of this distinctive method for feature extraction boosts the detection capabilities of machine learning algorithms, hence enabling them to effectively distinguish subtle yet crucial distinctions between phishing and authentic URLs. Hence, utilizing the methodology as mentioned earlier is crucial as it provides a more precise and dependable method for identifying phishing websites, effectively tackling the difficulties related to selecting appropriate features, and eventually strengthening defenses in cybersecurity.

3 Structure of URL

First, the components of URLs should be known to understand the attackers' approach. A visual representation of the URL's basic structure is in Figure 1 [17].

URL (Uniform Resource Locator) is a web address that identifies the location of a particular website on the Internet [18]. In phishing attacks, attackers manipulate URLs in multiple ways, such as creating special URLs, manipulating URLs, and manipulating keywords [17]. URLs constitute different components, some required and others optional, URL basic structure consists of the following elements:

1. Protocol: The protocol describes how a browser connects to a website. The protocol could be HTTP (hypertext transfer protocol) or HTTPS (HTTP secure).
2. Domain name: A domain name is the name of a website, such as XYZ-company.com.
3. Sub-domain: Subdomains are prefixes used to identify a domain name, such as www.
4. Top-level domain: This refers to the suffix of the domain name, such as .com, .org, .net, etc.
5. The Path: The path refers to the location of the resource on the server, such as /info/.
6. The file name is a freely selectable portion of text appearing before the file extension. It should provide information about a particular file, such as index.html.

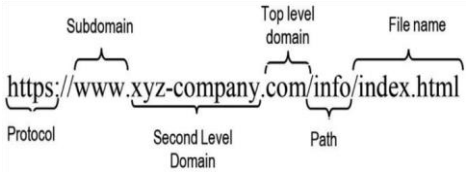

Figure 1: URL structure

4 Methodology

Machine learning-based systems depend strongly on the dataset and feature selection [19]. They have a direct impact on the system's effectiveness and efficiency. Therefore, these topics are discussed in detail in the following sections

4.1. Dataset

As part of Our research, we have enhanced phishing website detection through feature selection in URL-based analysis, by using Decision Trees, Support Vector Machines (SVM), and Random Forests to detect phishing websites. For this purpose, the Mendeley data [20] was utilized, which is a dataset that contains a collection of legitimate and phishing websites. The database contains 80,000 instances, including 50,000 legitimate websites and 30,000 phishing sites, each instance includes a URL, an HTML page, an index, and a result that has a binary value of either 0 or 1 (0 for legitimate, and 1 for phishing). This extensive database was reduced in size to expedite the process of feature extraction from URLs and optimize the use of computation resources. So, 8,000 URLs from the dataset were randomly selected, consisting of 4,000 legitimate URLs and 4,000 phishing URLs

4.2 Feature selection

The new strategy employed in creating datasets involved a careful rule-based methodology that effectively leveraged the capabilities of regular expressions to extract crucial attributes from URLs. In the context of this method, an extensive list of characteristics was initially considered, followed by rigorous testing to identify the 30 attributes that exhibited the highest relevance. The elements encompassed a diverse range of aspects, including address bar features, abnormal features, HTML and JavaScript features, and domain features. These features are listed in Table (2). A comprehensive rule-based analysis was conducted on each URL within the initial dataset to determine the presence or absence of these 30 specific attributes.

As an illustration, one of the requirements necessitated an assessment of the URL's length. Phishers often utilize extended Uniform Resource Locators (URLs) as a means to obscure their malicious intentions through the implementation of these strategies. In order to assess the dependability of the results, a mean URL length was calculated for each URL present in the dataset. URLs above the specified character limit of 65 were classified as possibly indicative of phishing

$$\text{Rule: IF } \begin{cases} URL\ length < 65 \rightarrow feature = Phishing(1) \\ Otherwise \rightarrow feature = legitimate(0) \end{cases}$$

websites, whilst URLs falling below this limit were deemed legitimate.

The utilization of regular expressions was imperative in this process as it facilitated the identification of patterns, substrings, and certain attributes present within the URLs. The aforementioned information was thereafter employed to develop informed assessments regarding the legality of the URLs under scrutiny. The utilization of regular expressions facilitated the meticulous rule-based methodology, enabling the creation of a novel binary dataset. In this dataset, instances denoting legitimate websites were assigned the label "0," while instances indicating potential phishing sites were assigned the label "1." This approach to dataset generation ensured the inclusion of only URLs that demonstrated predetermined features of significance, as identified through the utilization of regular expressions. This augmentation improved the accuracy and use of the dataset for subsequent modeling and analysis.

Table 2: List of features websites using the recently curated dataset

Criteria		**Phishing Features**
Address Bar-Features		IP Address
		Long URL
		"TinyURL"
		"@" Symbol
		using "//"
		Adding Prefix or Suffix Separated by (-) to the Domain
		Sub-Domain and Multi-subdomains
		HTTPS
		Domain Registration Length
		Favicon
		Non-Standard Port
		tilde_symbol
		The Existence of "HTTPS" Token in the Domain Part of the URL
Abnormal Features		Request URL
		URL of Anchor
		Links in <Meta>, <Script> and <Link> tags
		SFH
		Submitting Information to Email
		Abnormal URL
HTML and JavaScript-Features		Website Forwarding
		Status Bar Customization
		Disabling Right Click
		Using Pop-up Window
		IFrame Redirection
Domain-Features		Age of Domain
		DNS
		PageRank
		Google Index
		Number of Links Pointing to Page
		Statistical-Reports

Our machine-learning model Figure 2. was trained using the newly generated dataset, which included the URLs and feature availability results. Based on this dataset, the model was trained to learn patterns and associations between the features and the target labels (phishing or legitimate). The data set was divided into 80 % training and 20 % testing, as shown in Figure 3. As a result of leveraging decision trees, SVMs, and random forest algorithms, we were able to develop a powerful phishing website detection system that could effectively differentiate legitimate URLs from malicious URLs.

With this methodology, we were able to create a comprehensive detection system capable of detecting phishing attacks that might be conducted by malicious actors via URLs. In addition to improving the discrimination capabilities of the applied machine learning algorithms, the use of regular expression techniques for feature extraction further enhanced their accuracy, contributing to the exceptional level of detection.

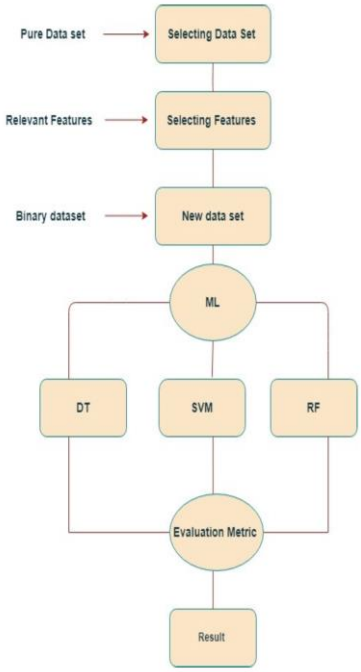

Figure 2: Machine-learning model

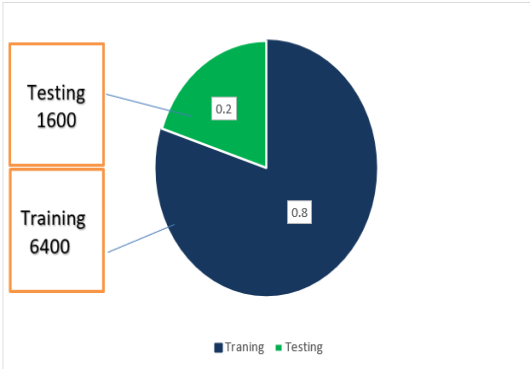

Figure 3: Data used for phishing URL

4.3 Used algorithms

The selection of the most appropriate machine learning algorithms is mostly dependent on their capacity to successfully handle rule-based feature sets and achieve high levels of classification accuracy. Decision Trees, Support Vector Machines (SVM), and Random Forests were chosen for evaluation based on their ability to capture complex decision boundaries, support rule-based features collected by regular expressions, and offer resilience against overfitting. Decision Trees demonstrate a notable proficiency in generating decision rules that are easily interpretable, while Support Vector Machines (SVMs) are particularly effective in maximizing the margin between different classes, thus aiding in their distinction. Random Forests, as an ensemble method, provide the advantages of feature selection and enhanced generalization. The selection of these algorithms was based on their compatibility with the rule-based feature selection procedure and their potential to achieve high accuracy in detecting phishing websites.

Decision Tree (DT)

The decision tree algorithm is a supervised learning algorithm used for classification and prediction tasks. Essentially, it resembles a tree, with each internal node representing a test on a certain attribute, each branch representing its outcome, and each leaf node representing a class label. A decision tree can be used in a variety of fields, including medicine, bioinformatics, and image classification [21], used splitting measures such as the Gini Index, Information Gain, etc., to determine how to split the tree.

The formula of the Gini Index is as follows [22]:

$$Gini = 1 - \sum_{i=1}^{n} (Pi)2$$

where,
'pi' is the probability of an object being classified to a particular class. Figure 4 shows the structure of DT.

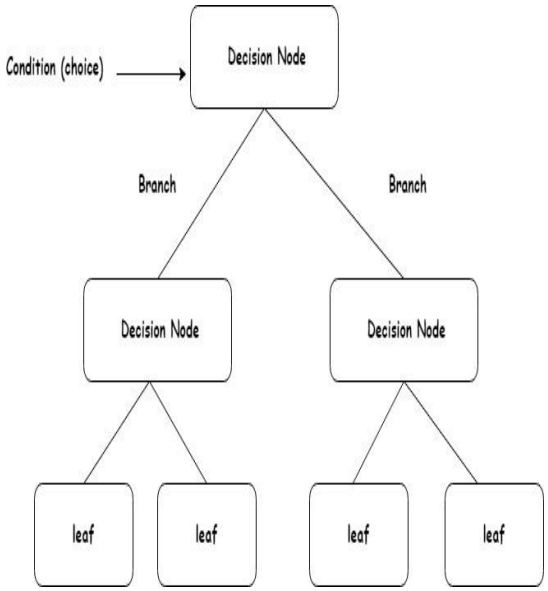

Figure 4: Structure of decision tree

Support vector machine (SVM)

Support Vector Machines (SVMs) are powerful and widely used supervised machine learning algorithms. This method is based on finding an optimal hyperplane that creates a clear boundary between data points of different classes. SVM has been extensively used in various fields, including recognition, image processing, natural language processing, and cybersecurity [23]. The general form of SVM [24]:

$$\beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots \beta pXp = 0$$

where p is the number of dimensions.
- For p=2 i.e. for a 2-D space it is a Line.
- the vector ((β1,β2,β3...βp)is just a Normal vector A vector in simple terms is just a 1-Dimentional Tensor or a 1-D array. Figure 5 shows the structure of SVM.

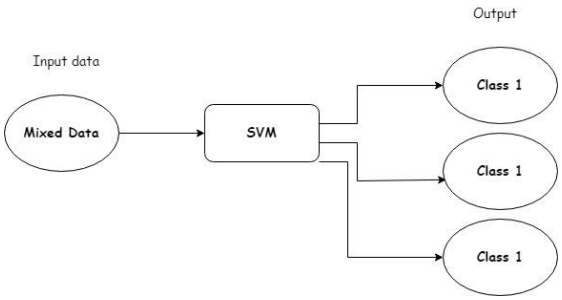

Figure 5: Structure of SVM

Random forest (RF)

This is a method of ensemble learning that is widely used for both classification and regression tasks in machine learning. During the training phase, multiple decision trees are constructed and their predictions are combined to produce more accurate and robust results. A random subset of data and features is used to train each decision tree, reducing the risk of overfitting and increasing the level of generalization [25].

$$\mathrm{E}\,\psi\,\theta(\mathrm{x}),\nu(\mathrm{x})(\mathrm{Oi})\mathrm{Xi} = \mathrm{x} = 0 \text{ for all } \mathrm{x} \in \mathrm{X}$$

The general form of Random forests [26]:

where ψ(·) is some scoring function and ν(x) is an optional nuisance parameter. Figure 5 shows the structure of the RF.

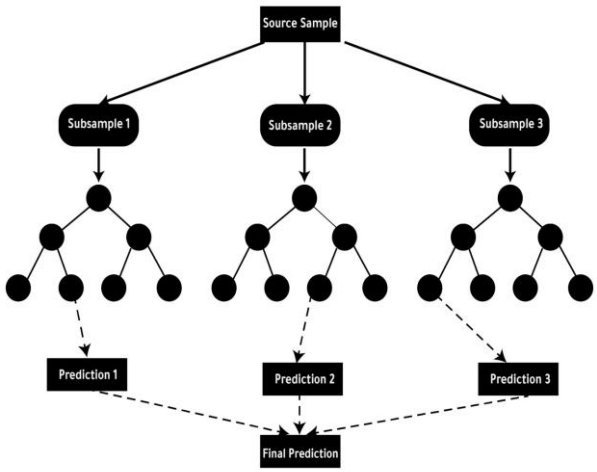

Figure 6: Structure of Random Forest

Evaluation metrics for machine learning are used to evaluate the effectiveness and performance of a model in solving a specific problem [27]. These metrics provide quantitative measures that assist researchers and practitioners in understanding how well a model performs. Additionally, they assist them in making informed decisions during the development of models [28]. As part of our research, we utilized the metrics Accuracy, Precision, Recall & F1-score

5 A comparative analysis

Our suggested feature selection method, which utilizes regular expressions, presents notable advantages when compared to other current methods for detecting phishing websites. While traditional techniques often face difficulties in selecting appropriate characteristics, our proposed solution employs a systematic extraction of essential features using rule-based regular expressions. This ensures that only the most relevant properties are taken into consideration. The thorough technique employed in this study serves to boost the performance of the model by effectively decreasing noise and reducing the dimension of the dataset. Furthermore, the flexibility of regular expressions enables us to effectively capture shifting phishing strategies, hence enhancing the adaptability and long-term reliability of our detection models. The effectiveness of our feature selection method is highlighted by the notable accuracy levels achieved using Decision Trees, Support Vector Machines (SVM), and Random Forests. This positions our method as a promising advancement in the field of phishing website detection, as it successfully overcomes the challenges associated with feature selection while retaining a high level of detection capabilities.

6 Ethical considerations

The ethical considerations of fraud website detection are of the highest priority, considering the sensitive nature of cybersecurity research. By ethical principles, our study ensures that all data utilized for model training and evaluation are acquired legitimately and with the appropriate authorizations. Ensuring the confidentiality and integrity of personally identifiable information is of the highest priority to us, and we take every precaution to safeguard it throughout the research process. Additionally, the dissemination of our findings emphasizes responsible disclosure of vulnerabilities to relevant authorities or organizations to facilitate mitigation and avoid any unintended misuse of our methods. It is essential to preserve the confidence and integrity of cybersecurity research that ethical standards be adhered to, and our work is in keeping with these principles.

7 Experimental Results and Discussion

In this section, we present the results of our model on phishing website detection using Decision Trees (DT), Support Vector Machines (SVM), and Random Forest (RF) algorithms. The implementation was performed using Python 3.10, and the evaluation metrics including accuracy, recall, precision, and F1-score were used to assess the performance of each algorithm, The results are organized in a tabular format, displaying the evaluation metrics for each algorithm as shown in Table 3 and figure 7.

Table 3: Performance Metrics of ML

MODEL	DT	SVM	RF
Accuracy	0.968	0.973	0.976
Recall	0.960	0.966	0.970
Precision	0.975	0.982	0.982
F1-Score	0.967	0.973	0.976

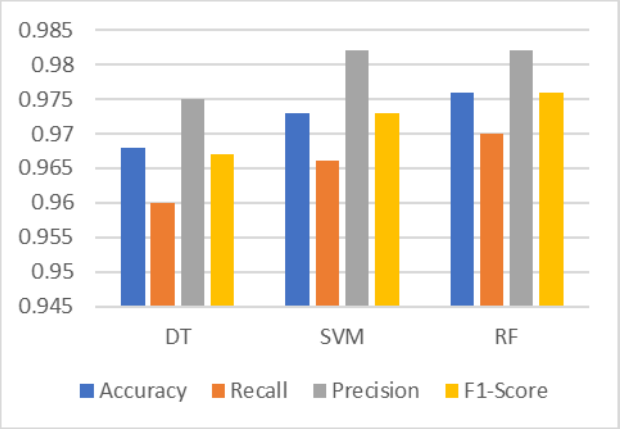

Figure 7: Performance metrics of ML

The mathematical equation is as follows:

Accuracy = (Number of correct predictions) / (The total number of predictions)

Furthermore, recall (also known as sensitivity) is a measure of the ability of algorithms to correctly identify phishing websites out of all the actual phishing attempts

[29]. In other words, a high recall indicates that the algorithm is effective at minimizing false negatives, that is, it correctly detects a greater number of phishing websites [28]. Recall can be expressed mathematically as follows:

Recall (TPR) = True Positives / True Positives + False Negatives

The precision metric measures the accuracy of positive predictions generated by the algorithms [29]. A higher precision indicates a lower rate of false positives, which means the algorithm does not misclassify legitimate websites as phishing sites. Precision can be calculated mathematically as follows:

Precision = True Positives / True Positives + False Positives [29]

Finally, the F1-score represents the harmonic mean of precision and recall, providing a balanced evaluation metric that accounts for both false positives and false negatives [28]. When there is an imbalance between the number of legitimate and phishing instances in the dataset, this method can be helpful [35]. F1-score can be calculated using the following mathematical equation:

F1-score =2×(Precision×Recall) / Precision + Recall.

The results obtained indicate that Random Forest achieved the highest accuracy of 0.976, followed by SVM with an accuracy of 0.973, and Decision Trees with an accuracy of 0.968. The Random Forest algorithm also performed well in terms of recall, precision, and F1-score, demonstrating its effectiveness in detecting both phishing and legitimate websites.

The experimental results demonstrate that SVM and Random Forest are effective in detecting phishing websites, indicating their potential as effective solutions for dealing with cybersecurity challenges. In the context of detecting phishing websites, the organized evaluation metrics provide valuable insight into the strengths of each algorithm and provide a basis for comparing their performance

When it comes to detecting phishing websites, Decision Trees, SVMs, and Random Forests each have their strengths and weaknesses. The decision tree offers simplicity and interpretability but may suffer from overfitting and limited expressiveness. An SVM is capable of handling complex data patterns, but requires careful hyperparameter tuning and may require a more extended training period. Since Random Forests are ensemble-based, they provide superior accuracy and feature importance analysis but may require more computational resources. By understanding these aspects, researchers and practitioners can choose the most suitable algorithm based on their specific requirements and constraints for phishing website detection.

8 Conclusion

By implementing a strategic approach to a feature selection in URL-based analysis, this research addressed the critical challenge of enhancing the detection of phishing websites. With regular expression techniques, we applied 30 pertinent features to a pure dataset sourced from Mendeley, to generate a new dataset which was used as input for machine learning. Our accuracy scores demonstrate the efficacy of our approach, with Decision Tree Accuracy reaching 0.968, SVM Accuracy exceeding 0.973, and Random Forest Accuracy exceeding 0.980.

Our findings emphasize the importance of feature selection when refining machine learning models for phishing detection. The extraction of relevant features from URLs using regular expressions has proven to be an effective strategy for optimizing algorithmic performance. Additionally, this contributes to the development of a broader range of efficient and accurate machine-learning applications, which are not limited to the cybersecurity domain.

There are several promising avenues for future research and enhancement. Integrate URL-based features with content-based attributes to establish an all-encompassing detection system. Furthermore, create real-time detection mechanisms so that phishing websites can be identified promptly, enhancing the safety of online transactions.

9 Future work and potential extensions

Since our methodology has exhibited significant efficacy in the detection of phishing websites, there exist other areas for further research and enhancement. The persistent advancement of phishing strategies requires constant modification and enhancement of the regular expressions employed for feature extraction. The placing of priority on enhancing the ability of our system to detect evolving phishing tactics will be of utmost importance.

1- The investigation of incorporating deep learning methodologies, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), together with our rule-based methodology, has the potential to enhance the accuracy of phishing attack detection. This is particularly relevant for complex and context-dependent phishing attacks.

2- The potential for expanding the utilization of our feature selection methodology to other cybersecurity fields, such as the identification of email phishing or the examination of malware, appears to be promising. Regular expressions have the potential to be utilized to

derive significant insights from textual data inside diverse cybersecurity situations. Exploration of the applicability of our methodology to these specific sectors has the potential to enhance its practicality and impact.

3-The engagement in interdisciplinary cooperation with linguistics and natural language processing specialists has the potential to generate new techniques for the detection of phishing efforts by using semantic details included in URLs and webpage content.

Through the examination of these prospective methods, our objective is to enhance the effectiveness of our approach in countering the ever-changing cyber threats and broaden its applicability to various domains within the field of cybersecurity.

References

[1] K. Ahmed and S. Naaz, "Detection of phishing websites using machine learning approach," in Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India, 2019. https://doi.org/10.2139/ssrn.3357736.

[2] M. Ahsan, K. E. Nygard, R. Gomes, M. M. Chowdhury, N. Rifat, and J. F. Connolly, "Cybersecurity threats and their mitigation approaches using Machine Learning—A Review," Journal of Cybersecurity and Privacy, vol. 2, no. 3, pp. 527-555, 2022. https://doi.org/10.3390/jcp2030027

[3] Y. Xu et al., "Artificial intelligence: A powerful paradigm for scientific research," The Innovation, vol. 2, no. 4, 2021.

https://doi.org/10.1016/j.xinn.2021.100179

[4] N. Kareem, "Afaster Training Algorithm and Genetic Algorithm to Recognize Some of Arabic Phonemes." https://doi.org/10.1109/isimp.2004.1434064

[5] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," in IOP Conference Series: Materials Science and Engineering, 2020, vol. 928, no. 3: IOP Publishing, p.032019. https://doi.org/10.1088/1757-899x/928/3/032019

[6] W. Chu, B. B. Zhu, F. Xue, X. Guan, and Z. Cai, "Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs," in 2013 IEEE international conference on communications (ICC), 2013: IEEE, pp.1990-1994. https://doi.org/10.1109/icc.2013.6654816

[7] W. Fadheel, M. Abusharkh, and I. Abdel-Qader, "On Feature selection for the prediction of phishing websites," in 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2017: IEEE, pp.871-876. https://doi.org/10.1109/dasc-picom-datacom-cyberscitec.2017.146

[8] I. Tyagi, J. Shad, S. Sharma, S. Gaur, and G. Kaur, "A novel machine learning approach to detect phishing websites," in 2018 5th International conference on signal processing and integrated networks (SPIN), 2018: IEEE, pp. 425-430. https://doi.org/10.1109/spin.2018.8474040

[9] A. D. Kulkarni and L. L. Brown III, "Phishing websites detection using machine learning," 2019. https://doi.org/10.14569/ijacsa.2019.0100702

[10] D. N. Kumar, N. S. R. Hemanth, S. Premnath, V. N. Kumar, and S. Uma, "Detection of phishing websites using an efficient machine learning framework," International Journal of Engineering Research and Technology, vol. 9, no. 5, 2020. https://doi.org/10.17577/ijertv9is050888

[11] A. Lakshmanarao, P. S. P. Rao, and M. B. Krishna, "Phishing website detection using novel machine learning fusion approach," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021: IEEE, pp. 1164-1169. https://doi.org/10.1109/icais50930.2021.9395810

[12] M. Abutaha, M. Ababneh, K. Mahmoud, and S. A.-H. Baddar, "URL phishing detection using machine learning techniques based on URLs lexical analysis," in 2021 12th International Conference on Information and Communication Systems (ICICS), 2021: IEEE, pp.147-152. https://doi.org/10.1109/icics52457.2021.9464539

[13] S. Jain, "Phishing Websites Detection Using Machine Learning," Available at SSRN 4121102. https://doi.org/10.2139/ssrn.4121102

[14] S. A. Anwekar and V. Agrawal, "PHISHING WEBSITE DETECTION USING MACHINE LEARNING ALGORITHMS." https://doi.org/10.5120/ijca2018918026

[15] A. Prathap, M. L. Mounika, M. Reethika, N. Navya, and R. S. Sahithi, "PHISHING WEBSITE DETECTION USING MACHINE LEARNING MODELS," Machine learning, vol. 52, no. 4, 2023. https://doi.org/10.17762/msea.v70i2.2447

[16] U. B. Penta, B. Panda, and S. S. Gantayat, "MACHINE LEARNING MODEL FOR IDENTIFYING PHISHING WEBSITES," Journal of Data Acquisition and Processing, vol. 38, no. 1, p. 2455, 2023. https:// DOI: 10.5281/zenodo.7764722

[17] 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. https://doi.org/10.1016/j.eswa.2018.09.029

[18] A. Aljofey et al., "An effective detection approach for phishing websites using URL and HTML features," Scientific Reports, vol. 12, no. 1, p. 8842, 2022. https://doi.org/10.1038/s41598-022-10841-5

[19] E. M. Karabulut, S. A. Özel, and T. Ibrikci, "A comparative study on the effect of feature selection on classification accuracy," Procedia Technology, vol. 1, pp. 323-327, 2012. https://doi.org/10.1016/j.protcy.2012.02.068

[20] S. F. Ariyadasa, Shantha; Fernando, Subha, "Phishing Websites Dataset," Mendeley Data, 2021, doi: http://doi.org/10.17632/n96ncsr5g4.1.

[21] G. Stiglic, S. Kocbek, I. Pernek, and P. Kokol, "Comprehensive decision tree models in bioinformatics," PloS one, vol. 7, no. 3, p. e33812, 2012. https://doi.org/10.1371/journal.pone.0033812

[22] S. V. Razavi-Termeh, A. Sadeghi-Niaraki, and S.-M. Choi, "Spatial modeling of asthma-prone areas using remote sensing and ensemble machine learning algorithms," Remote Sensing, vol. 13, no. 16,p.3222,2021. https://doi.org/10.3390/rs13163222

[23] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189-215, 2020. https://doi.org/10.1016/j.neucom.2019.10.118

[24] A. Mammone, M. Turchi, and N. Cristianini, "Support vector machines," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 1, no. 3, pp. 283-289, 2009. https://doi.org/10.1002/wics.49

[25] L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32., 2021, https://doi.org/10.1023/a:1010933404324

[26] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," 2019. https://doi.org/10.1214/18-aos1709

[27] Y. Liu, Y. Zhou, S. Wen, and C. Tang, "A Strategy on Selecting Performance Metrics for Classifier Evaluation," International Journal of Mobile Computing and Multimedia Communications, vol. 6, pp. 20-35, 10/01 2014, https://doi.org/10.4018/ijmcmc.2014100102.

[28] N. Japkowicz and M. Shah, Evaluating learning algorithms: a classification perspective. Cambridge University Press, 2011. https://doi.org/10.1017/cbo9780511921803.001

[29] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," arXiv preprint arXiv:2010.16061, 2020. https://doi.org/10.48550/arXiv.2010.16061