Improve the effectiveness of machine learning models in detecting website phishing using morphological features in URL analysis

DOI: https://doi.org/10.54654/isj.v2i22.1040

Dang Thi Mai, Nguyen Viet Hung*

***Abstract*— With the proliferation of the Internet, the emergence of various threats has become increasingly prevalent, particularly the danger posed by phishing websites. These websites are designed with malicious content aimed at exploiting users who inadvertently access them. This method of attack represents a significant potential risk for users in cyberspace. The problem of detecting and eliminating phishing websites has garnered significant interest and research within the community. In this study, we propose a set of morphological features in URL path analysis, combined with machine learning methods, to detect phishing website URLs. Experimental evaluation with the UCI Repository dataset results have demonstrated the effectiveness of the proposed feature set in terms of all metrics (Accuracy, Precision, Recall, and F1 Score) compared to previous methods.**

***Tóm tắt*— Cùng với việc Internet ngày càng phát triển nhanh chóng thì những mối đe dọa cũng xuất hiện ngày càng nhiều, nổi bật là mối đe dọa về các website lừa đảo (phishing). Các website phishing được tạo ra với những nội dung nguy hại với mục đích tấn công vào người sử dụng truy cập vào chúng. Thủ đoạn tấn công này tiềm ẩn nguy cơ lớn đối với người dùng trên không gian mạng. Bài toán phát hiện và bóc gỡ những website phishing đã được quan tâm và nghiên cứu rộng rãi trong cộng đồng. Trong nghiên cứu này, chúng tôi đề xuất bộ thuộc tính hình thái trong phân tích đường dẫn URL, kết hợp với các phương pháp học máy để phát hiện ra các URL của các website lừa đảo. Kết quả đánh giá thử nghiệm trên bộ dữ liệu UCI Repository đã chứng minh tính hiệu quả của bộ thuộc tính trong phát hiện website phishing so với các phương pháp trước đây cả về các độ đo đánh giá (Accuracy, Precision, Recall và F1 Score).**

***Keywords*— website phishing detection, phishing URL, morphological features.**

***Từ khóa*— phát hiện website lừa đảo, đường dẫn lừa đảo, đặc trưng hình thái.**

I. INTRODUCTION

Malicious websites form the backbone of numerous criminal activities on the Internet. These sites host a variety of harmful content, ranging from spam advertisements to phishing schemes and dangerous exploits that infect visiting machines with malicious code. The prevalence of malicious websites is increasing, with their variations becoming more diverse and their concealment methods more sophisticated [1]. Malicious URLs can be delivered to users through email, text messages, pop-up windows in web browsers, or embedded in electronic advertisements. When users inadvertently interact with these malicious URLs, they risk downloading malware onto their systems, thereby compromising information security. A single malicious website has the potential to infect thousands of computers within a very short period. Consequently, researching techniques to automatically detect and block malicious URLs before users access them is a crucial and effective measure for ensuring information security in cyberspace.

There has been widespread interest in developing systems to prevent end users from accessing such websites. The most prominent approach available to prevent phishing websites is to use blacklists of malicious URLs [2]. However, this method has many disadvantages, such as requiring regular and continuous updates

This manuscript was received on July 23, 2024. It was reviewed on August 10, 2024, revised on September 8, 2024 and accepted on September 23, 2024.

* Corresponding author

of blacklist. Otherwise, it will not be possible to detect new malicious domains. Furthermore, this method is complicated when the attacker adds malicious code executables to the valid URL or redirects valid links to another malicious website after the user accesses it. Therefore, the method of using blacklists cannot completely prevent the increasingly strong development of phishing websites. Currently, research based on machine learning models to classify and analyze malicious URLs is proving to be an effective method [3]. To improve the weaknesses of classic methods, this method has the ability to generalize to new URLs without the need for frequent data updates, promptly preventing new threats for users.

To effectively apply machine learning methods, URLs must be analyzed, and the most relevant features extracted to distinguish between legitimate and phishing URLs. Numerous studies have been conducted to identify and synthesize the characteristics of URLs to determine the most suitable features [4]. Currently, four main groups of features are widely used for this purpose [5]:

- Feature extraction based on character analysis of URLs.

- Feature extraction based on unusual information.

- Feature extraction based on HTML and JavaScript content.

- Attributes based on domain information.

Machine learning methods utilizing these features have demonstrated relatively good performance in classifying benign and phishing URLs, indicating that these feature groups adequately capture the differences between the two. However, the classification results still require further improvement. In this paper, we propose the addition of morphological features to the existing set of URL features to enhance the detection of phishing URLs using machine learning. Our main contributions are as follows:

- Proposing new morphological feature groups extracted from URLs.

- Enhancing the phishing URL dataset for training and evaluating machine learning models.

- Conducting the experimental evaluations of popular machine learning models using the new dataset and comparing results with those obtained using the original dataset.

II. Related work

Many studies have applied machine learning models to detect fraudulent URLs. The general steps for using machine learning to solve other problems are similar to those in Figure 1.

The study conducted by Jitendra Kumar et al. [6] involved training various classifiers such as Logistic Regression, Naive Bayes Classifier, Random Forest, Decision Tree, and K-Nearest Neighbor, based on features extracted from the lexical structure of URLs. They constructed the dataset of URLs to address issues of data imbalance, mistraining, variance, and overfitting. The dataset contains an equal number of labeled legitimate and phishing URLs, divided in a 7:3 ratio for training and testing purposes. All classifiers exhibited nearly identical AUC (area under the ROC curve) values, but the Naive Bayes Classifier emerged as the most suitable, achieving the highest AUC value. The Naive Bayes Classifier attained the highest accuracy of 98%, with a precision of 1, a recall of 0.95, and an F1 score of 0.97, underscoring the reliability of the results.

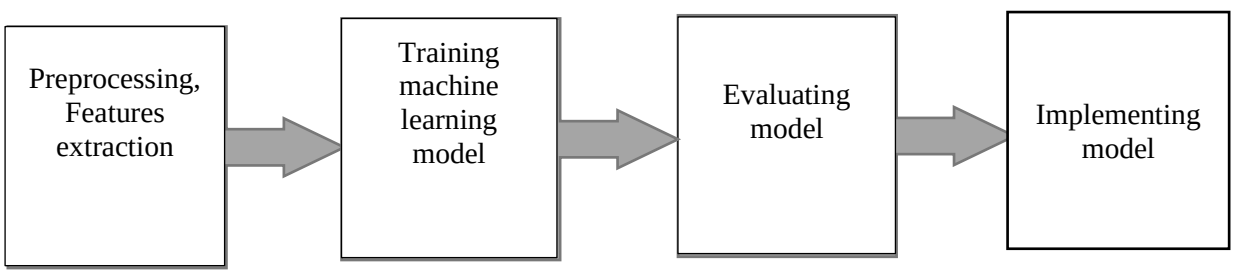

Figure 1. Process of applying machine learning models in detecting phishing URLs

Mehmet Korkmaz and colleagues proposed a machine learning-based phishing detection system using eight different algorithms on three different datasets represented by a set of 48 features, including UIC material [7]. The algorithms employed were Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost, Random Forest (RF), and Artificial Neural Network (ANN). It was observed that the models utilizing LR, SVM, and NB exhibited low accuracy. Conversely, the DT, RF, and ANN algorithms yielded better results regarding training time and accuracy. The study concluded that either the RF algorithm or ANN algorithm could be effectively used, as they require less training time and achieve a high accuracy rate.

Mohammad Nazmul Alam et al. [3] proposed a system for detecting phishing attacks using Random Forests and Decision Trees. Utilizing a Kaggle dataset with 32 features, they applied feature selection algorithms such as principal component analysis (PCA) to enhance model efficiency. Feature selection reduces redundancy by eliminating irrelevant data, thus making the model more practical for real-world applications. The proposed model employed REF, Relief-F, IG, and GR algorithms for feature selection prior to applying PCA. The Random Forest algorithm achieved 97% accuracy, exhibiting less variance and effectively addressing the overfitting problem.

Abdulhamit Subasi et al. [9] presented an intelligent phishing detection system using the UCI dataset. Various machine learning tools, including ANN, KNN, SVM, C4.5 DT, RF, and Rotated Forests, were utilized to classify phishing websites. The Random Forest classifier demonstrated superior performance in terms of accuracy, F-measure, and AUC. It was found to be faster, more powerful, and more accurate than the other classifiers.

In their article [10], Rishikesh Mahajan and Irfan Siddavatam selected three classification algorithms: Decision Tree, Random Forest, and Support Vector Machine. Their dataset comprised 17,058 benign URLs and 19,653 phishing URLs, collected from Alexa and PhishTank websites, each featuring 16 attributes. The dataset was divided into training and testing sets in ratios of 50:50, 70:30, and 90:10, respectively. Performance evaluation metrics included accuracy score, false negative rate, and false positive rate. They achieved 97.14% accuracy with the Random Forest algorithm, which also had the lowest false negative rate. The study concluded that accuracy increases with the amount of data used for training.

A study by Khan et al. [11] compared the performance of different machine learning methods (DT, SVM, RF, NB, KNN, and ANN) using three different datasets. Additionally, it evaluated the effectiveness of these models with datasets of reduced dimensions. The dataset was sourced from the UCI machine learning repository and other online sources. Experimental results indicated that Random Forest and artificial neural network methods achieved 97% accuracy.

In the article by Saleem Raja Abdul Samad et al. [12], eight classification machine learning algorithms were selected, including Logistic Regression (LogR), SVM, Gaussian Naive Bayes (GNB), KNN, DT, RF, Gradient Boosting (GB), and Extreme Gradient Boosting (XGB). They utilized two datasets for their research: the UIC dataset (30 features) and the Mendeley dataset (48 features). The study highlighted the importance of feature selection, demonstrating that a minimal number of features can achieve higher accuracy. This finding aids future research in selecting the optimal scoring function with the least number of features necessary. By combining fine-tuning factors with the machine learning model without additional procedural factors, the model's accuracy performance improved. The results indicated that factor tuning enhances the effectiveness of machine learning algorithms. For the UIC dataset, RF and XGB achieved accuracy rates of 97.44% and 97.47%, respectively. For the Mendeley dataset, GB and XGB achieved accuracy values of 98.27% and 98.21%, respectively.

In [13], the authors proposed using URL-based features alongside features related to transport layer security (such as length, number of slashes, number of points, and location). Utilizing these features, the Decision Tree method achieved an accurate detection rate of 93% with the apriori algorithm. In [14], a SVM and multinomial Naive Bayes model was used to determine whether a website is phishing or not. The authors in [15] employed Harmony Search and SVM, training on the UCI dataset containing 11,055 websites and 20 features. Features were selected from a pool of 30 using the Decision Tree method instead of the wrapper method. The proposed method achieved an accuracy rate of 92.80% using nonlinear regression based on led Harmony Search (HS), demonstrating the promising potential of these methods in the field of phishing detection. In [16], TF-IDF analysis is applied to identify phishing content, extracting key phrases that typify such websites. These extracted phrases are then used as queries across various search engines. The resulting web pages are aggregated, sorted, and prioritized based on their relevance and ranking. This process helps in identifying potential phishing websites, enabling further investigation into their deceptive practices and fraudulent intent. In [17], Do.X.C et.al. proposed using static and dynamic features of URLs and tested it with SVN and RF machine learning models. The experimental results on a self-collected dataset with 470,000 URLs from trusted sources such as Phishtank, URLhaus, and Alexa gave results above 90% in terms of accuracy, precision, and recall rate.

Most of the mentioned works used the 30 attributes of the UCI dataset, while some other authors added static and dynamic information on URLs to extract new features. However, a large number of URLs are automatically generated to avoid detection by blocklist filters; more analysis of the features of such URLs is required. The detection results of phishing URLs can most likely be further improved by adding more suitable features.

III. PROPOSAL FOR URL PHISHING DETECTION USING MORPHOLOGICAL FEATURES

The most widely used dataset in research on detecting phishing websites is the dataset shared on the Machine Learning Repository website provided by UCI, which includes 30 features extracted from URLs to identify fraudulent URLs [18]. These 30 features cover various aspects such as address bar-based properties, anomaly-based properties, and HTML & JavaScript-based properties. However, morphological features have not been extracted and utilized. Many fraudulent URLs are associated with domains generated using automatic domain generation algorithms (DGA - Domain Generated Algorithms) [19] to bypass the blacklist. Domains generated by DGA exhibit different morphological characteristics compared to regular domains. These morphological characteristics can significantly contribute to determining whether a URL is fraudulent. For exampe, DGA domains often consist of random or semi-random strings of characters, resulting in domains that appear nonsensical and have irregular lengths. In contrast, regular domains are usually shorter, meaningful, and easy to remember. The distribution of characters in DGA-generated domains is also typically more random, with a mix of letters and numbers that do not form recognizable words. Regular domains, on the other hand, often contain readable words or acronyms.

In this study, we propose to incorporate morphological features into the URL feature set to enhance the effectiveness of identifying fraudulent URLs. In linguistics, morphology is the study of word forms and their relationships with other words in the same language [20]. We propose to add a group of morphological features, including the following six features:

a. Ratio of vowels and consonants

Typically, in languages, the number of consonants exceeds the number of vowels. An uneven distribution of consonants relative to vowels can often indicate that domain names are being generated randomly by a DGA. The

calculation of the vowel-to-consonant ratio is performed using the following formula:

$$\text{VC\_}Ratio = \frac{|V|}{|C|} \quad (1)$$

where |V| is the number of vowels in URL, and |C| is the number of consonants.

b. Maximum number of consecutive consonants

This feature can indicate the unusual morphology of a domain. In standard English domains, it is rare to see more than seven consecutive consonants in a word. Domain Generation Algorithms (DGAs) often produce such results through random word generation or the concatenation of dictionary words. Therefore, examining the maximum number of consecutive consonants can be useful in detecting domain names generated by these algorithms.

c. Ratio between numbers and letters

Domain names typically use alphanumeric characters rather than numbers, with few exceptions. A high quantity of numerical characters in a domain name, particularly in longer domain names, is a notable characteristic of domain names generated by algorithms.

d. Repeated consonants

This feature represents the total number of consonant repetitions in the domain, calculated as the number of repetitions divided by the domain length. Domains randomly generated by algorithms exhibit a significantly higher number of repeated consonants compared to standard domains.

e. Vowels repeated

his feature calculates the total number of repeated vowels in the domain, normalized by its length (i.e., repetitions divided by domain length). Domains generated by algorithms from dictionaries typically exhibit a notably higher frequency of repeated vowels compared to standard domains.

f. Letter score

This is a quantifiable measure of the popularity of the alphabetic component of a domain name. The objective is to detect highly unusual or random letter combinations. Each of the 26 alphabetical characters in English has been assigned a score based on its frequency of occurrence in the current Alexa dataset of domains, as detailed in Table 1.

TABLE 1. FREQUENCY OF CHARACTERS USED

Letter	Score	Letter	Score	Letter	Score
a	0.08	j	0.00	s	0.06
b	0.02	k	0.01	t	0.09
c	0.03	l	0.04	u	0.03
d	0.04	m	0.02	v	0.01
e	0.13	n	0.07	w	0.02
f	0.02	o	0.08	x	0.00
g	0.02	p	0.02	y	0.02
h	0.06	q	0.00	z	0.00
i	0.07	r	0.06		

Letter score is determined by the formula:

$$L_{Score} = \frac{1}{n}\sum_{i=0}^{n}(S_{Ci}) \quad (2)$$

in which S_{Ci} is the frequency of letter C_i

IV. EXPERIMENTS AND EVALUATION

A. Dataset Description

The UCI Repository dataset comprises 11,054 links, with 6,157 labeled as legitimate and 4,897 as malicious. URLs are labeled 1 for legitimate and -1 for phishing. In our approach, we split the data into training and testing sets using an 80/20 ratio. Each URL in the dataset is associated with 30 attributes corresponding to the features mentioned earlier. We augment the dataset with 6 additional morphological features described in Section 3, resulting in a total of 36 attributes.

To comprehensively evaluate the new attribute set, we employ 8 diverse machine learning models: LogR, SVM, Bernoulli Naïve Bayes (BNB), KNN, DT, RF), GB, and XGB. The experiment is implemented in Python 3

using the Keras library, running on a computer with the following configuration:

- Intel Core i7-6700HQ CPU @2.60GHz
- RAM: 12GB.

B. Evaluation measure

To evaluate the effectiveness of the models, we use the following metrics: Accuracy, Precision, Recall, and F1 Score. These metrics are calculated based on the values True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In the context of URL classification:

- True Positive (TP) represents the number of phishing URLs correctly classified as phishing.

- True Negative (TN) represents the number of legitimate URLs correctly classified as legitimate.

- False Positive (FP) represents the number of legitimate URLs incorrectly classified as phishing.

- False Negative (FN) represents the number of phishing URLs incorrectly classified as legitimate.

These values are typically organized into a Confusion Matrix, as shown in Table 2, which allows for a clear assessment of the model's performance.

TABLE 2. CONFUSION MATRIX

Predict/Actual		**Actual**	
		Positive	**Negative**
Predict	Positive	TP-True Positive	FP-False Positive
	Negative	FN-False Negative	TN-True Negative

The Accuracy measure helps us evaluate the predictive effectiveness of a model on a set of data. The higher the accuracy, the more accurate our model is. Accuracy calculation formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (3)$$

Precision tells us that out of the cases predicted to be positive, how many cases are correct, and the higher the precision, the better our model is at classifying. The formula for precision is as follows:

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall measures the rate of correctly predicting positive cases across all samples belonging to the positive group. To calculate recall, we must know the data labels in advance. Therefore, recall can be used to evaluate the training and validation sets because we already know the labels. We will use precision when the data is considered entirely new and the labels are unknown on the test dataset. The formula for recall is as follows:

$$Recall = \frac{True\ positive}{Total\_actual\_positive} = \frac{TP}{TP+FN} \quad (5)$$

F1 Score, a significant metric, is the harmonic average between precision and recall. It serves as a representative measure of accuracy, taking into account both precision and recall. This makes it a comprehensive metric for evaluating a model's performance.

$$F_1 = 2\frac{Precision.Recall}{Precision+Recall} \quad (6)$$

C. Result of evaluation

The results of evaluating 8 machine learning models on the above UIC dataset with 30 features and with 36 features (6 proposed additional features) are shown in Table 3.

The results indicate that utilizing 36 features generally yields better performance across most algorithms when evaluated on the UCI dataset. However, the decision tree model shows lower performance with the 36-feature dataset compared to the 30-feature dataset. Notably, the XGBoost (XGB) model achieves the highest accuracy of 98.7% when utilizing the 36-feature dataset.

TABLE 3. EXPERIMENTAL RESULTS

AI model	Number of features	Accuracy	Precision	Recall	F_1
Logistic Regression (LogR)	30	0.929	0.935	0.943	0.939
	36	**0.938**	**0.940**	**0.950**	**0.945**
Support Vector Machine (SVM)	30	0.925	0.925	0.943	0.934
	36	**0.940**	**0.939**	**0.952**	**0.945**
Bernoulli Naïve Bayes (BNB)	30	0.907	0.916	**0.919**	**0.918**
	36	**0.909**	**0.921**	0.909	0.915
K-Neighbors Classifier (KNN)	30	0.950	0.950	**0.946**	0.948
	36	**0.953**	**0.958**	0.944	**0.951**
Decision Tree (DT)	30	**0.958**	**0.964**	**0.961**	**0.962**
	36	0.945	0.958	0.940	0.949
Random Forest (RF)	*30*	*0.968*	*0.967*	*0.977*	*0.972*
	36	**0.976**	**0.977**	**0.980**	**0.978**
Gradient Boosting (GB)	30	0.967	0.970	0.972	0.971
	36	**0.985**	**0.985**	**0.992**	**0.988**
Extreme Gradient Boosting (XGB)	30	0.972	0.971	0.980	0.976
	36	**0.987**	**0.989**	**0.985**	**0.987**

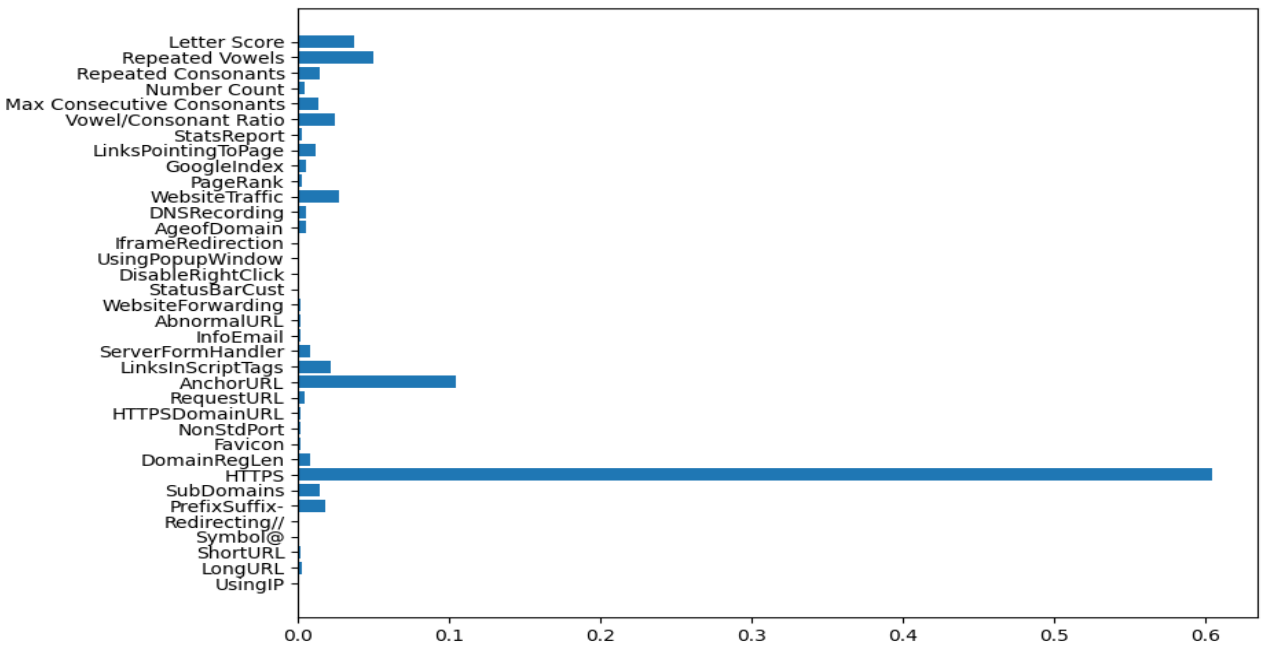

Figure 2. Features importance

In Figure 2, the feature importance analysis highlights that morphological features play a crucial role in distinguishing phishing URLs from legitimate ones. This experiment underscores the significant contribution of morphological features in enhancing the accuracy of phishing URL detection.

V. CONCLUSION

Phishing detection remains a critical focus in research, driven by its crucial role in safeguarding privacy and enhancing information security. Various methods, including machine learning-based classification of websites, are

employed for phishing detection. The efficacy of these methods heavily relies on the features extracted from URLs. In this study, we propose augmenting the original set of 30 features with morphological characteristics, resulting in a comprehensive set of 36 features. Experimental findings demonstrate that this enhanced feature set improves the detection of phishing URLs across several machine learning algorithms on the same dataset. Specifically, the Extreme Gradient Boosting (XGB) model achieves the highest accuracy of 98.7%, followed closely by Gradient Boosting (GB) at 98.5%. It's important to note that the performance of each algorithm can vary depending on factors such as dataset composition, the ratio of training to test data, and the specific feature selection techniques employed. Future research will focus on refining attribute selection to better capture the nuanced characteristics of phishing URLs, aiming to develop even more effective detection models.

Acknowledgements: This research is funded by University of Transport and Communications (UTC) under grant number T2024-CB-010.

REFERENCES

[1] Number of unique phishing sites detected worldwide from 3rd quarter 2013 to 1st quarter 2024 - https://www.statista.com/ - Last accessed: 6/2024.

[2] Narendra. M. Shekokar, Chaitali Shah, Mrunal Mahajan, Shruti Rachh, An Ideal Approach for Detection and Prevention of Phishing Attacks, Procedia Computer Science, Volume 49, 2015, Pages 82-91.

[3] S. A. Murad, N. Rahimi and A. J. Md Muzahid, "PhishGuard: Machine Learning-Powered Phishing URL Detection," 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE), Las Vegas, NV, USA, 2023, pp. 2279-2284.

[4] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," 2015 IEEE Conference on Communications and Network Security (CNS), Florence, Italy, 2015, pp. 769-770

[5] Rami M. Mohammad, Fadi Thabtah, and Lee McCluskey.Phishing websites features, 2015. Unpublished. Available via:http://eprints.hud.ac.uk/24330/6/RamiPhishing_Websites_Features.pdf.

[6] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran and B. S. Bindhumadhava, "Phishing Website Classification and Detection Using Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-6.

[7] Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri, Detection of phishing websites by using machine learning-based URL analysis, 11nth International Conference on Computing, Communication and Networking Technologies(ICCCNT), 2020.

[8] M. N. Alam, D. Sarma, F. F. Lima, I. Saha, R. -E. -. Ulfath and S. Hossain, "Phishing Attacks Detection using Machine Learning Approach," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1173-1179.

[9] Subasi, A.; Molah, E.; Almkallawi, F.; Chaudhery, T.J. Intelligent phishing website detection using random forest classifier. In Proceedings of the International Conference on Electrical and Computing Technologies and Applications (ICECTA), Phuket, Thailand, 12–13 October 2017; pp. 1–5.

[10] Rishikesh Mahajan, and Irfan Siddavatam, Phishing website detection using machine learning algorithms, International Journal of Computer Applications(0975-8887), vol. 181, no. 23, 2018.

[11] Khan, S.A.; Khan, W.; Hussain, A. Phishing Attacks and Websites Classification Using Machine Learning and Multiple Datasets (A Comparative Analysis). In Intelligent Computing Methodologies: 16th International Conference, ICIC 2020, Bari, Italy, 2–5 October 2020, Proceedings, Part III; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12465.

[12] Saleem Raja Abdul Samad, Sundarvadivazhagan Balasubaramanian, Amna Salim Al-Kaabi, Bhisham Sharma, Subrata Chowdhury, Abolfazl Mehbodniya, Julian L. Webber and Ali Bostani. Analysis of the Performance Impact of Fine-Tuned Machine Learning Model for Phishing URL Detection-Electronics 2023, 12, 1642.

[13] S. C. Jeeva and E. B. Rajsingh, “Intelligent phishing url detection using association rule

mining," Human-centric Computing and Information Sciences, vol. 6, no. 1, Oct. 2016.

[14] Thang, N. M., Anh, L. Q., Toan, H. S., & Trung, N. Q. (2023). A novel method for detecting URLs phishing using hybrid machine learning algorithm. Journal of Science and Technology on Information Security, 2(19), 15-28. https://doi.org/10.54654/isj.v2i19.978.

[15] M.Babagoli, M. P.Aghababa, and V.Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," Soft Computing, vol. 23, no. 12, pp. 4315–4327, 2018.

[16] Shaoming Chen, Yiyang Wang, and Xingkai Cheng. An approach for detecting phishing websites by using search engine. In Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application (ICMLCA '23).

[17] Cho Do Xuan, Hoa Dinh Nguyen and Tisenko Victor Nikolaevich, "Malicious URL Detection based on Machine Learning" International Journal of Advanced Computer Science and Applications(IJACSA), 11(1), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110119.

[18] Mohammad, R.; McCluskey, T.L.; Thabtah, F. UCI Machine Learning Repository: Phishing Websites Data Set. Available online: https://archive.ics.uci.edu/dataset/327/phishing+websites (accessed on 20 March, 2024).

[19] Dutta AK. Detecting phishing websites using machine learning technique. PLoS One. 2021 Oct 11;16(10):e0258361. doi: 10.1371/journal.pone.0258361. PMID: 34634081; PMCID: PMC8504731.

[20] S. Anderson, "A-Morphous Morphology" (2011). Cambridge University Press.

ABOUT THE AUTHORS

Dang Thi Mai

Workplace: Faculty of Basic science, University of Transport and Communications, Vietnam.

Email: dtmai@utc.edu.vn

Education: She received her BSc, MSc in Applied Mathematics and Informatics from Hanoi university of Natural Science, PhD degrees in Applied Mathematics from Moscow Institute of Physics and Technology in 2012.

Recent research detection: Applied Mathematics; Machine learning.

Tên tác giả: **Đặng Thị Mai**

Đơn vị công tác: Khoa Khoa học cơ bản, Trường Đại học Giao thông vận tải Hà Nội, Việt Nam

Email: dtmai@utc.edu.vn

Quá trình đào tạo: Nhận bằng Đại học năm 2004; Thạc sĩ năm 2006 tại Đại học Khoa học Tự nhiên Hà Nội; bằng Tiến sĩ năm 2012 tại Đại học Vật lý Kỹ thuật Matxcova.

Hướng nghiên cứu hiện nay: Toán ứng dụng, học máy.

Nguyen Viet Hung

Workplace: Institute of information and communication technology, Le Quy Don Technical University, Vietnam.

Email: hungnv@lqdtu.edu.vn

Education: He received his BSc, MSc and PhD degrees in Computer Science from Moscow Institute of Physics and Technology in 2006, 2008 and 2012 respectively.

Recent research detection: Information security; Malware detection; Intrusion detection.

Tên tác giả: **Nguyễn Việt Hùng**

Đơn vị công tác: Viện Công nghệ thông tin và truyền thông, Trường Đại học kỹ thuật Lê Quý Đôn, Việt Nam

Email: hungnv@lqdtu.edu.vn

Quá trình đào tạo: Nhận bằng Đại học năm 2006; Thạc sĩ năm 2008 và Tiến sĩ năm 2012 tại Đại học Vật lý Kỹ thuật Matxcova.

Hướng nghiên cứu hiện nay: An toàn thông tin; phát hiện mã độc; phát hiện xâm nhập.