ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 06 June 2024

Phishing Website Detection using Machine Learning and Deep Learning Techniques

[1] N.V. Naik, [2] Sangam Lasyapriya, [3] Cchanduh Bellamkonda, [4] Pulipaka Sravanthi

[1][2][3][4] Lakireddy BaliReddy College of Engineering (Autonomous), Mylavaram, Andhra Pradesh, India
Email: [1] nvnaikit@gmail.com, [2] lasyapriyasangam44@gmail.com, [3] cchanduh2002@gmail.com, [4] pulipakasravanthi2003@gmail.com

Abstract*— Our study methodology efficiently incorporates ma- chine learning and deep neural network technology to attack the widespread problem of phishing strategies used by fraudulent websites in the extensive digital environment. Our inquiry primarily centers on analyzing the correlation between several computational approaches, including XGBoost, SVM, Random Forest, Logistic Regression, CNN-LSTM, and CNN-BILSTM. Our technique differs from standard evaluations by concentrating on individual web pages rather than entire websites. We combine various characteristics and incorporate elements depending on the URL, domain, and content of the page. The study provides a comprehensive analysis of machine learning and deep learning methods, seeking to assess their comparative efficacy in several aspects of website functionality. This extensive evaluation encompasses a broad spectrum of prospective outcomes, guaranteeing flexibility and dependability in various circumstances. Specifically, the practical utility of our findings is enhanced by the fact that we sourced our dataset from the PhishTank website. Importantly, our work obtained a noteworthy milestone by obtaining the highest level of accuracy at 98.95% utilizing the SVM method. The exceptional precision exhibited here illustrates the efficacy of our approach in precisely detecting bogus websites. Our main aim is to help individuals and academics discover practical solutions to combat fraudulent online platforms by providing crucial comparative information. We want to contribute to the establishment of a safe online environment by exhaustively exploring the capabilities of machine learning and deep learning, particularly in the areas of fraud prevention and using various data attributes such as URL, domain, and content.

Index Terms*— Phishing, URL features, Machine learning, Deep learning

I. INTRODUCTION

The continually evolving nature of phishing assaults is a significant impediment in today's internet-dependent society. Phishing is a sophisticated cyber hazard that uses deceitful tactics to fool users into providing their personal information. As technology advances rapidly, deception efforts have grown deeper, exploiting people's vulnerabilities. This illicit action is ubiquitous and more detrimental, injuring individuals, businesses, and even nations.

As a consequence, phishing assaults have become more sophisticated and various, reaching victims via more channels than ever before (email, chat, websites, etc.). Fraudsters deploy innovative methods to give these fraudulent messages alluring identities, making it hard for people to discriminate between actual and phony information. The penalties for falling prey to phishing are severe, ranging from business losses and fraud to illicit utilization of data and information to security failures.

Social engineering is a major part of phishing, with tailored communications and employing important-soun- ding personalities to acquire confidence. The danger worsens when using cell phones and accessing unsecured Wi-Fi networks since tiny displays may disguise signals of deceit, and unsecured networks present possibilities for fraudsters to intercept information.

Additionally, the lack of two-factor authentication (also known as 2FA) increases risk, depending entirely on passwords and enhancing the likelihood of unlawful access if consumers fall for a fraudulent effort. Recognizing these hazards and comprehending the multidimensional nature of fraud is vital for remaining secure online. Regular cybersecurity education, prudence in online contacts, and adherence to security requirements are critical elements in limiting the hazards related to falling for phishing schemes.

II. LITERATURE SURVEY

The research done by Abutaha and other researchers in 2021 proposes a unique fraud detection system built upon URL linguistic analysis and machine learning classifiers. The authors processed a dataset, yielding 22 characteristics, which were further reduced using several strategies. The assessment comprised prominent algorithms such as random forest, gradient boosting, neural networks, and support vector machines (SVM). Results suggest that SVMs surpassed other classifiers, obtaining an outstanding success rate of 99.89% in recognizing examined URLs. [1] The suggested technique is positioned for practical deployment as an add-on or middleware function inside internet browsers, targeted at informing online users when attempting to visit a feasible phishing website only based on its URL.

The research done by Chu and other researchers in 2013 demonstrates the significant worldwide and Chinese-specific incidences of phishing as the third and top cybersecurity threat, respectively. The work focuses on defending recognized websites from phishing assaults, concentrating on machine learning-based detection employing lexical and domain at tributes, even when phishing web pages are

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 06 June 2024

unavailable. The researchers propose unique features, evaluating their efficacy using phishing attack data targeting major sites like Taobao and Tencent in China. [2] This research determines an ideal collection of characteristics for the phishing detector, attaining a detection rate above 98% while retaining a rate of false positives of 0.64% or below.

In their 2022 research, Almomani and other researchers im proved fraud website detection by extracting different semantic elements, including URL and domain identity, anomalous elements, HTML and JavaScript features, and domain features. These aspects increase the controllability and efficacy of the categorization process. [3] The researchers employ machine learning model algorithms to identify fraudulent websites, leveraging 16 distinct machine learning models with the ten semantic criteria regarded as most successful for phishing webpage identification. Mainly, GaussianNB and the stochastic gradient descent (SGD) classifier exhibit the lowest performance findings, with 84% and 81%, respectively, in contrast to other classifiers employed in the research.

In their 2017 research, Patil and other researchers concentrate on exploiting the fundamental visual elements of an online page's appearance as a foundation for recognizing similarities across sites, with a particular emphasis on rapidly identifying phishing web pages. The researchers suggest a unique strategy, recognizing how the site layouts and contents comprise essential characteristics of a web page's look. Given that website layouts are commonly set by Cascading Style Sheets (CSS), the researchers propose an algorithm to discover similarities in key CSS-related components.[4] Their suggested system incorporates the Support Vector Machine (SVM) approach and the map-reduce paradigm, proving its utility in obtaining improved accuracy in identifying spam emails.

In their research, Tyagi and other researchers emphasize the crucial need to safeguard information communicated online, specifically during growing phishing assaults. The primary emphasis of their work is on implementing several machine learning algorithms to determine the validity of websites. The researchers underscore the capabilities of machine learning technologies to efficiently identify zero-hour phishing assaults and exhibit greater flexibility to combat evolving forms of phishing threats. [5] The execution of their technique exhibits outstanding results, obtaining a precision rate of 98.4% in precisely identifying the degree to which a website is authentic or part of a fraud effort. This emphasizes the efficacy of machine learning in augmenting cybersecurity defenses, particularly in the context of rapidly developing phishing attack scenarios.

Overall, the content provides a comprehensive overview of the research and advancements in fraud detection using machine learning and deep learning.

III. PROPOSED METHODOLOGY

This paper presents an all-encompassing approach to fraud website identification, integrating machine and deep learning techniques. The core dataset, derived from PhishTank, undergoes comprehensive preparation processes where empty values, repeated anomalies, and probable outliers are methodically managed, and label encoding is utilized to assist in incorporating category variables. The suggested technique in- corporates several machine learning models, such as XGBoost, Random Forest, SVM, and Logistic Regression, with deep learning models that include CNN-LSTM and CNN-BILSTM. These models are trained on the preprocessed dataset, which incorporates critical features, including URL-based properties, allowlist and blocklist classifications, content-based characteristics, and domain-based features.

Performance evaluation employs several criteria, notably precision, recall, precision, accuracy, and F1 scores. The comparison of machine and deep learning models attempts to discover their comparative capabilities in detecting fraudulent websites. This exhaustive assessment procedure provides an extensive comprehension of the systems' capabilities and assists in selecting the most efficient techniques for phishing detection

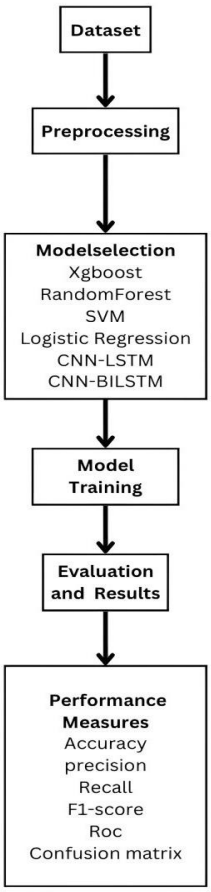

Fig. 1. Flowchart of Architecture.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 06 June 2024

A. Dataset Description

This dataset combines real-time information from Kaggle and Phishtank and is a thorough collection of all of that data. It features an extensive database of more than 25,470 items and covers a wide range of 47 properties. These characteristics are methodically divided into three primary types: URL-based, domain-based, and content-based. The various natures of these elements give a comprehensive picture, delivering significant insights into the complexity of phishing activities.

B. Preprocessing

Several critical procedures are involved in preprocessing the dataset to get the data clean and suitable for analysis. The first step in manipulating data is to load libraries such as Pandas. To make it easier to work with the dataset, it is then imported into a Pandas DataFrame. Potential mistakes or inconsistencies are addressed by identifying and removing duplicate rows. Following that, we deal with null values; depending on the data, we may delete rows with null entries or impute new values to fill them. By checking them, we handle repeated values, which may indicate redundancy or inaccuracies. After the dataset is cleaned, it is saved to a new file so it may be analyzed later. Depending on the dataset's qualities, such as encoding category variables or dataset's qualities, such as encoding category variables or scaling numerical features, additional considerations may be required. The particular implementation specifics rely on the dataset's unique properties and requirements.

C. Algorithms

1) *Logistic regression Algorithm:* Logistic regression to categorize URLs as phishing or genuine is an essential compo nent of phishing website detection. A dataset including various URL properties, represented as Xi , is used to train the logistic regression model. Here, i stands for each unique URL, and θ stands for the model parameters. The hypothesis function hθ(Xi) is defined by the sigmoid function:

$$h_\theta(X_i) = \frac{1}{1 + e^{-\theta^T X_i}}$$

This function predicts the likelihood that a URL belongs to the phishing class. The model parameters θ were iteratively updated to minimize the logistic loss function using optimization techniques such as gradient descent.

2) *Random Forest Algorithm:* Random forests use an ensemble learning technique that merges several decision trees to enhance classification accuracy. We train each decision tree using a different collection of characteristics and data samples to arrive at the final classification. Then, we combine their predictions. The average of each tree's projections is the random forest model's forecast:

$$\hat{y} = mode(y_i, y_2, \ldots, y_n)$$

Random forest models may effectively handle both high-dimensional data and overfitting.

3) *XGBoost algorithm:* One extreme gradient boosting technique is XGBoost, which optimizes a differentiable loss function progressively inside the gradient boosting framework. By definition, the goal function of XGBoost is defined as:

$$Objective(\theta) = \sum_{i=1}^{n} loss(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

This is where the loss function is defined: $(y_i, \hat{y}_i)$ denotes the loss function, $\Omega(f_k)$ is the regularization term, and K is the number of trees. XGBoost creates decision trees repeatedly, improving the objective function at each step to reduce loss and enhance model performance

4) *Support Vector Machine algorithm:* By projecting URLs into a three-dimensional space and locating the hyperplane that distinguishes between valid and malicious URLs, support vector machines are used for URL classification. One may find the SVM decision function $f(X_i)$ by:

$$f(X_i) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(X_i, X) + b\right)$$

The variable α_i represents the Lagrange multipliers, y_i stands for the class labels, $K(Xi, X)$ is the kernel function, and b denotes the bias term in this context. SVM accomplishes reliable classification by optimizing the space between the hyperplane and the support vectors.

5) *CNN-LSTM algorithm:* In phishing website identification, the CNN-LSTM architecture is applied to handle sequential data represented by URL characteristics. The CNN- LSTM model comprises convolutional neural network layers and long short-term memory layers. The CNN layers collect spatial features from the input sequences of URL features, while the LSTM layers capture temporal relationships in the sequences. The forward pass of a CNN-LSTM model comprises sending the input sequences through convolutional layers to extract features, followed by LSTM layers to capture sequential patterns. The output of the LSTM layers is then sent into a final entirely linked layer for classification. Mathematically, the forward pass of a CNN-LSTM model may be expressed as:

$$Z[l] = W[l]A[l-1] + b[l]$$

Here, $Z^{[l]}$ represents the output of layer l, $W^{[l]}$ and $b^{[l]}$ are the parameters of layer l, and $A^{[l-1]}$ is the activation of the preceding layer.

6) *CNN-BILSTM algorithm:* In Phishing website detection, the CNN-BILSTM architecture is implemented to

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 06 June 2024

manage sequential data represented by URL characteristics. The CNN BILSTM model incorporates convolutional neural network (CNN) layers and bidirectional long short-term memory layers. The CNN layers of the model gather spatial features from the input sequences of URL features, while the bidirectional LSTM layers capture bidirectional relationships in the sequences. Thus, the model learns from past and future information in the input sequences.

Mathematically, the forward pass of a CNN-BILSTM model comprises transmitting the input sequences through convo- lutional layers to extract features, followed by bidirectional LSTM layers to capture bidirectional relationships in the sequences. The output of the bidirectional LSTM layers is then sent into a final entirely linked layer for classification. The forward pass of a CNN-BILSTM model may be expressed as:

$$Z[l] = W[l]A[l-1] + b[l]$$

Here, $Z^{[l]}$ represents the output of layer l, $W^{[l]}$ and $b^{[l]}$ are the parameters of layer l, and $A^{[l-1]}$ is the activation of the preceding layer.

IV. RESULTS

A. Accuracy

The accuracy findings demonstrate differential efficacy across various algorithms for fraudulent website detection. SVM, Random Forest, and XGBoost demonstrated respectable performance, obtaining high accuracy scores of 98.11% , 98.9% , and 98.9% , respectively. Logistic Regression and CNN-BI-LSTM also exhibited strong results, with precisions of 97.7% and 97.95% , respectively. However, the CNN- LSTM model displayed a reduced accuracy of 57.7%. These results show that ensemble approaches like Random Forest and XGBoost and classic techniques like Logistic Regression are beneficial in this situation. Adjustments should be made based on the unique requirements of the application and the relevance of eliminating false positives and false negatives.

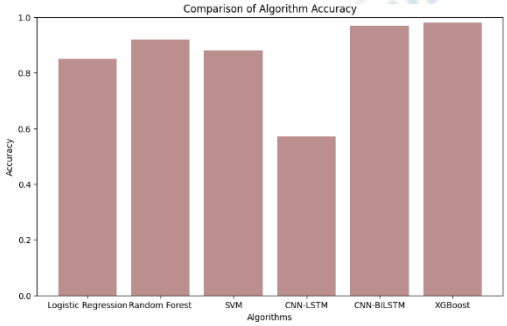

Fig. 2. Graphical representation of Model accuracy

B. ROC-AUC Curve

The region as part of the ROC curve (AUC) values gives a complete assessment of the discriminatory strength of classifi- cation algorithms. In this scenario, logistic regression exhibits excellent results with an AUC of 0.96, demonstrating its ability to discriminate between classes successfully. Random Forest obtains an ideal AUC of 1.0, signifying immaculate discrimination. The Support Vector Machine (SVM) performs well with an AUC of 0.98, indicating its tremendous dis- criminating power. XGBoost stood out with an ideal AUC of 1.0, exhibiting exceptional performance. The Convolutional Neural Network with Long Short-Term Memory (CNN- LSTM) exhibits robust discriminating powers with an AUC of0.97. In contrast, the CNN with Bidirectional LSTM (CNN- BILSTM) is out with a high AUC of 0.99, indicating its great discriminatory capacity.

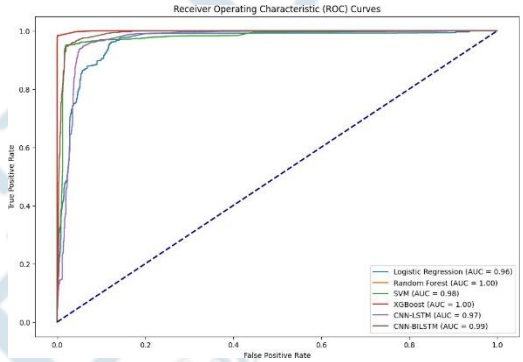

Fig. 3. ROC Curve representation of all Model

C. Confusion matrix

For binary classification, the XGBoost model performs strongly in the confusion matrix. The model effectively detected occurrences of the positive class, with 2142 genuine optimistic predictions. A low false positive count of 9 showed that it only misclassified a small number of negatives as positives. Regardless, there were 44 false negatives, which indicates that the model failed to account for some cases in the positive class. Positively, the model demonstrated its competence in recognizing occurrences of the hostile class by accurately predicting them 28,99 times. In conclusion, the XGBoost model generally demonstrated strong prediction skills, although it had a slight bias towards false negatives, which means it may be better at identifying positive cases.

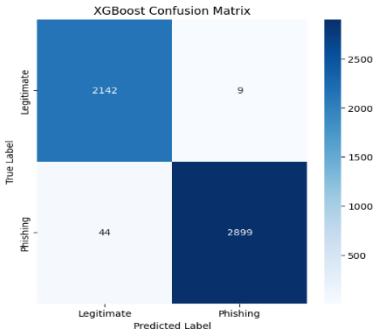

Fig. 4. Confusion matrix of XGBoost model

The SVM model's confusion matrix demonstrates an impressive level of effectiveness for binary classification. By

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 06 June 2024

correctly identifying 2107 positive class instances, the SVM model demonstrated its capacity to produce accurate positive classifications. Despite a low false positive count of 44, the model inaccurately categorized some negatives as positives. However, 52 occurrences of the positive class went missing (false negatives), indicating a potential area for development in collecting all positive cases. On the contrary, the SVM model exhibited vital accuracy by accurately foreseeing the negative class 2891 times (actual negatives). In conclusion, the SVM model demonstrated strong prediction skills when focused on true positives and negatives. However, if the model could increase its sensitivity to positive occurrences, it would be even better.

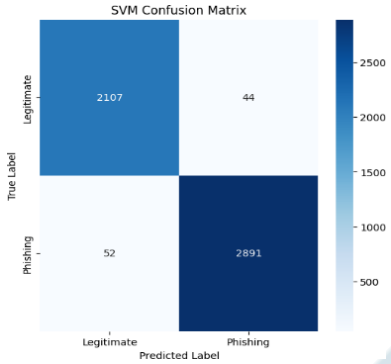

Fig. 5. Confusion matrix of SVM

D. Classification report

1) *Xgboost: The classification report for the XGBoost model* includes a comprehensive assessment of its binary classification performance, where labels ”0” and ”1” ostensibly denote independent classes. For class ”0,” which presumably designates one category, the model exhibited excellent accuracy (0.98), implying reliable predictions of positives among occurrences classified as positive. The recall for class ”0” is immaculate at 1.00, and the F1-score, reflecting an even distribution between the accuracy and recall, is exceptionally excellent at 0.99. The support column displays 2151 class ”0” instances in the dataset.

Similarly, for class ”1,” the model demonstrated immaculate accuracy (1.00), denoting accurate optimistic predictions and a recall of 0.99, coupled with an F1-score of 0.99. The support column states there were 2943 instances of class ”1.” Ultimately, the XGBoost model scored an exceptional accuracy of 0.99, highlighting its capacity to generate exact predictions for both classes in the complete dataset.

Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	2151
1	1.00	0.99	0.99	2943
accuracy			0.99	5094
macro avg	0.99	0.99	0.99	5094
weighted avg	0.99	0.99	0.99	5094

Fig. 6. Training and testing classification report of XGBoost Model

2) *Support Vector Machine*: The SVM model’s classification report exhaustively assesses its efficacy in a binary classification job where labels ”B” and ”M” are given, perhaps designating benign and malicious URLs, respectively. For the ”B” class, representing probable safe URLs, the model attained a high precision of 0.98, demonstrating reliable optimistic predictions among cases identified as positive. The recall, assessing the ability of the model to detect genuine positive events, is similarly 0.98. The F1-score, an even measure of accuracy and recall, is 0.98. The support column displays 2151 class ”B” cases in the dataset.

Similarly, for the ”M” class, signifying potentially malicious URLs, the model exhibited high performance with an accuracy of 0.99, recall of 0.98, and an F1-score of 0.98. The support column reveals 2943 class ”M” cases in the dataset. Overall, the SVM model acquired an accuracy of 0.98, demonstrating its ability to categorize website URLs, both benign and malicious, appropriately.

Classification Report:

	precision	recall	f1-score	support
B	0.98	0.98	0.98	2151
M	0.99	0.98	0.98	2943
accuracy			0.98	5094
macro avg	0.98	0.98	0.98	5094
weighted avg	0.98	0.98	0.98	5094

Fig. 7. Training and testing classification report of SVM

V. CONCLUSION

Our exploration of fraud website detection included multiple machine-learning techniques, which include Logistic Regression, Random Forest, SVM, XGBoost, CNN-LSTM, and CNN-BILSTM. ROC curves and AUC values demonstrated the algorithms’ ability to identify between phishing and genuine websites. Random Forest and XGBoost displayed notable accuracy, precision, recall, and F1 scores. The Random Forest model exhibited adaptability in effectively recognizing positive and negative examples, whereas XGBoost showcased excellent accuracy. Various algorithms obtained accuracy values surpassing 98%, indicating their efficacy in fraud detection. Trade-offs between accuracy and recall were investigated, emphasizing the necessity for application specific considerations. Our results contribute insights into cybersecurity, emphasizing the relevance of machine learning in phishing detection. Further study might investigate feature engineering and optimization strategies for increased model performance. Overall, our work emphasizes the potential benefits of machine learning in bolstering cybersecurity efforts and safeguarding people from online dangers.

REFERENCES

[1] Abutaha, M., Ababneh, M., Mahmoud, K. W., Baddar, S. W. A. (2021). URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis. International Conference on Information and Communication Systems (ICICS). https://doi.org/10.1109/icics52457.2021.9464539

[2] Chu, W., Zhu, B., Feng, X., Guan, X., Cai, Z. (2013). Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs. IEEE. https://doi.org/10.1109/icc.2013.6654816

[3] Almomani, A., Alauthman, M., Shatnawi, M. T., Alweshah, M., Alrosan, A., Alomoush, W., Gupta, B. B., Gupta, B. B., Gupta, B. B. (2022). Phishing website detection with semantic features based on machine learning classifiers.International Journal on Semantic Web and Information Systems, 18(1), 1–24. https://doi.org/10.4018/ijswis.297032.

[4] Patil, P., Rane, R., Bhalekar, M. (2017). Detecting spam and phishing mails using SVM and obfuscation URL detection algorithm. 2017 International Conference on Inventive Systems and Control (ICISC). https://doi.org/10.1109/icisc.2017.8068633.

[5] Tyagi, I., Shad, J., Sharma, S., Gaur, S., Kaur, G. (2018). A Novel Machine Learning Approach to Detect Phishing Websites. Eurasip Jour nal on Information Security. https://doi.org/10.1109/spin.2018.8474040.

[6] Yerima, S. Y., Alzaylaee, M. K. (2020). High Accuracy Phishing Detection Based on Convolutional Neural Networks.IEEE. https://doi.org/10.1109/iccais48893.2020.9096869.

[7] Jain, A. K., Gupta, B. B. (2016). A novel approach to protect against phishing attacks at client side using auto-updated white-list. Eurasip Journal on Information Security,2016(1). https://doi.org/10.1186/s13635-016-0034-3.

[8] Korkmaz, M., S ¸ahing¨ oz, ¨ O. K., of Phishing DiRi, B. (2020).Detection Websites by Using MachineLearning-Based URL Analysis.IEEE.https://doi.org/10.1109/icccnt49239.2020.9225561.

[9] Mahajan, R., Siddavatam, I. (2018). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications, 181(23), 45–47. https://doi.org/10.5120/ijca2018918026.

[10] Mandadi, A., Boppana, S., Ravella, V., Kavitha, R. (2022). Phishing website detection using machine learning. 2022 IEEE 7th International Conference for Convergence in Technology (I2CT). https://doi.org/10.1109/i2ct54291.2022.9824801.

[11] Singh, P., Maravi, Y. P., Sharma, S. (2015). Phishing websites detection through supervised learning networks. IEEE. https://doi.org/10.1109/iccct2.2015.7292720.

[12] Hawanna, V., Kulkarni, V., Rane, R. (n.d.). A novel algorithm to detect phishing URLs. IEEE. https://doi.org/10.1109/icacdot.2016.7877645.

[13] Aljabri, M., Altamimi, H. S., Albelali, S. A., Al-Harbi, M., Alhuraib, H. T., Alotaibi, N. K., Alahmadi, A. A., Alhaidari, F., Mohammad, R. M. A., Salah, K. (2022). Detecting malicious URLs using Machine Learning Techniques: Review and research directions. IEEE Access,10,121395–121417.https://doi.org/10.1109/access.2022.3222307.

[14] Tang, L., Mahmoud, Q. H. (2021). A survey of Machine Learning-Based Solutions for Phishing Website Detection. Machine Learning and Knowledge Extraction, 3(3), 672–694. https://doi.org/10.3390/make303003.