International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Anti-Phishing: A Web Identifier for Spoofed Sites Using Neural Network

Akshatha A P[1], Chaithra R[1], Madhura S[1], Chandana C Sagar[1] , Hiriyanna G S[2]

* *1UG students, Dept. of CS&E , JNNCE, Shivamogga, India.*
* *2Asst. Prof ., Dept. of CS&E , JNNCE, Shivamogga, India.*

Abstract -***Phishing attacks persist as a significant threat to cybersecurity, exploiting deceptive websites to illicitly acquire sensitive user information. Conventional anti-phishing techniques often struggle to keep pace with the evolving sophistication of these attacks. An approach for detecting phishing websites using Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN) known for its ability to capture sequential dependencies in data.Method leverages the temporal dynamics inherent in website content and user interactions to discern between legitimate and phishing websites and constructed a comprehensive data-set comprising diverse samples of legitimate and phishing websites, ensuring the model's ability to generalize across various attack vectors.LSTM-based model exhibits resilience to adversarial evasion techniques commonly employed by attackers, demonstrating robustness in real-world scenarios and conducted extensive analysis to interpret the learned representations captured by the LSTM network, providing insights into the distinguishing features of phishing websites.***

Key Words:Phishing, Deep Learning,

1. INTRODUCTION

Phishing is a form of cyber attack where malicious actors attempt to trick individuals into disclosing sensitive information, such as login credentials, financial details, or personal data. The term "phishing" is a play on the word "fishing," as it involves luring victims with bait. These attacks typically take the form of deceptive emails, text messages, or websites that appear to be from legitimate sources, such as banks, social media platforms, or government agencies.

Phishing attacks often exploit psychological tactics, such as urgency or fear, to prompt victims to act quickly without questioning the authenticity of the communication.For instance, a phishing link might mimic a bank's website URL but feature a subtle alteration, redirecting unsuspecting users to a fraudulent site designed to harvest their login credentials. Unwary users who fall for these tricks may inadvertently provide their sensitive information to cybercriminals.

There are several common types of phishing attacks, including:

1. Email Phishing:This is the most prevalent form of phishing, where attackers send deceptive emails purporting to be from trusted organizations. These emails often contain links to fake websites or malicious attachments designed to steal information.

2. Spear Phishing: In spear phishing attacks, cybercriminals target specific individuals or organizations, often using personalized information to increase the likelihood of success. This could involve using the victim's name, job title, or other details obtained through research.

3. Whaling:Whaling attacks target high-profile individuals, such as executives or celebrities, with the aim of stealing sensitive information or perpetrating financial fraud. These attacks often involve sophisticated tactics and social engineering techniques.

4. Smishing:Smishing, or SMS phishing, involves sending fraudulent text messages to trick recipients into revealing personal information or clicking on malicious links. These messages often claim to be from banks, delivery services, or other trusted organizations.

5. Vishing:Vishing, or voice phishing, involves using phone calls to deceive victims into providing sensitive information or performing certain actions. Attackers may impersonate legitimate organizations or individuals to gain the victim's trust.

Phishing websites are fraudulent online platforms designed to imitate legitimate websites, often with the aim of deceiving users into divulging sensitive information such as passwords or financial details. These malicious sites typically employ various tactics to appear authentic, including mimicking the visual design and branding of reputable organizations or using deceptive URLs. Once users enter their information, cybercriminals exploit it for nefarious purposes, such as identity theft, financial fraud, or unauthorized access to accounts.

With the rapid development of machine learning, there are more and more applications in the field of cybersecurity, and we have proposed a deep learning-based framework to detect phishing links in a real-time web browsing environment. When the URL of the current tab of the browser is predicted to be a phishing link, the current page will receive an obvious warning prompt. The prediction result is obtained by the core prediction service calling a trained model.

2. IMPORTANCE OF PHISHING WEBSITE DETECTION

Detecting phishing websites is paramount in safeguarding personal and financial security in today's digital landscape. These deceptive websites are crafted to mimic legitimate platforms, aiming to trick unsuspecting users into divulging sensitive information like passwords, credit card details, and personal identifiers. By identifying and blocking phishing sites, individuals and organizations can prevent dire consequences such as identity theft, financial loss, and reputational damage. Moreover, swift detection helps curb the dissemination of malware, which often accompanies phishing

attempts, thus fortifying cybersecurity defenses and averting broader data breaches. Compliance with regulatory frameworks is also facilitated through proactive phishing website detection, ensuring adherence to stringent data protection standards. Furthermore, it presents an opportunity for educating users about recognizing and evading such threats, empowering them to navigate the digital realm with heightened vigilance. Ultimately, effective phishing website detection not only preserves trust and integrity in online interactions but also contributes to the broader mission of fortifying cybersecurity infrastructures globally.

3. OBJECTIVES

- The objective of this project is to train deep neural network on the dataset created to predict phishing websites.
- To recognize the legitimate websites and phished websites through URL.
- Implementation of Features Based Extraction of URL (Uniform Resource Locator) so as to decrease manual extraction.
- Deep learning training model that recognizes both the legitimate and phishing websites.
- Implementing the model to directly recognize the phishing website in the browser.

4. METHODOLOGY

1. **Data Collection**: In building a phishing detection system using deep learning, URL data collection is a crucial first step. The dataset, sourced from UCI, includes 11,055 website lists, containing both legitimate and phishing sites. A URL has five basic components: the Protocol, which sets rules for data transfer; the Domain, which includes the "www" prefix (World Wide Web Address), the website name, the organization type, and optionally the country code; the Path, detailing the specific section and page of the website; the Query, which addresses a part of a page; and the Query part, which sends additional information with the page request.
2. **Feature Extraction:**

Table -1: Address Based Feature

Feature name	Feature explanation
IP Address	If the domain contains an IP Address instead of domain name, it considers as phishing website.
URL length	Long URL name can contain malicious contents. If the length of the URL is more than the average length of an URL, it is considered as Suspicious or phishing.
TinyURL	TinyURL is used for shortening the URL length. By clicking on the shorter URL, it redirects to to main page. TinyURL links are considered as phishing website because it can redirect the user into fake website instead of legitimate website.
Having "@" symbol in URL	Browser generally skips the part attached with @ symbol so it is avoided in real addresses.
Using "//" symbol	"//" symbol is used for redirecting to another website. We consider it legit if the sign is used after HTTP or HTTPS. If the symbol is used after the initial protocol declaration, we consider it phishing.
Having "-" in domain name	Most of the real URLS don't contain "-" symbol . We considered an URL phishing if it contains "-" in its domain name.
Dots in domain	We need to add dot to append a sub-domain with the domain name. If more than 1 subdomain occurs, we consider it suspicious and greater than that will point it as a phishing site.
HTTPS	HTTPS protocol and the age of certificate is very important as most of the legit website uses HTTPS and has the trusted certificate.
Domain expiry date	A legit website generally have longer expiry date of their domain name.
Favicon	Favicon is a graphic image used in websites. If it is loaded from external domain, it can redirect user to suspicious sites.
Using unimportant ports	If a URL has some open ports which is unnecessary, phishers can take advantage.
"HTTPS" on domain	If an URL have "HTTPS" on it's domain name, it is considered as phishing website.

Table -2: Abnormal Based Features

Feature name	Feature explanation
Request URL	If a page contains higher amount of external URL or contents from another domain, we consider it suspicious or phishing based

	on the percentage.
Using <a> tags	Similar to the request URL features, the more we see <a> tags used in the website, the risk of phishing increases.
Links in <meta>,<script> and <Link> tag.	If <meta>,<script> and <Link> tag contains high amount of external links, it is considered as either suspicious or phishing based on the percentage.
Server Form Handler(SFH)	If SFHs is blank or empty, it is considered as phishing. It is marked as suspicious if the user is redirected to a different domain by SFHs.
Submitting Information to Email	If the information submitted by web form directed to a personal email instead of a server, it is considered as phishing.
Abnormal URL	If the identity is not included in the URL, it considered as phishing.

Table -3: HTML and JavaScript based features

Feature name	**Feature explanation**
Website forwarding	If redirecting is occurred multiple times, it can be alarming.
Status Bar Customization	"onMouseOver" event can be used to change the status bar of the URL. This can hide the fake URL and show the real URL to trick users. It is considered as phishing if it is applied on the website.
Disabling Right Click	Phishers generally disable right-click function so that users can't check the source code. If the function is disabled in the website, it can be taken as a phishing website.
Pop-up Window	If a web page consists of Pop-up window with a text field, it can be marked as phishing webpage.
IFrame usage	IFrame is used to attach external contents to show in a domain. Phishers might use IFrame tag by hiding them in the website.

Table -4: Domain based features

Feature name	**Feature extraction**
Age of Domain	If the age of domain is longer than 6 months, we can consider it as a legitimate website as phishing sites tend to live for shorter period of time.
DNS record	If a website doesn't contain any DNS record, it can highly considered as phishing site.
Website Traffic	If a website is visited by huge amount of people, it would have higher ranking. This ranking can help us to identify if a site is phishing or not. A higher ranked website tends to have lower chance of being a phishing website.
PageRank	A value is assigned to a webpage based on its importance. Most of the phishing sites have no pageRank value.
Google Index	If a site has name on the Google Index, we can assume that it is a legitimate website.
Links pointing to Page	A phishing website has shorter lifetime , so it doesn't have much links pointing towards it.
Statistical Reports	If the host of the webpage belongs in any Top phishing IP's or domains, we can count it as phishing webpage.

The phishing detection system uses four main features with a total of 30 sub-features to assess whether a website is legitimate, phishing, or suspicious. Address-based features (12

sub-features) as mentioned in the Table -1. Abnormal-based features (6 sub-features) as mentioned in the Table -2. HTML and JavaScript-based features (5 sub-features) as mentioned in the Table -3. Domain-based features (7 sub-features) as mentioned in the Table -4. Each feature provides critical information to classify the website's legitimacy.

3. **Data Preprocessing:** Data preprocessing is crucial for phishing website detection using deep learning. Typically, the dataset is split into training and testing sets to ensure proper model assessment. However, this reduces the number of training data points, potentially affecting accuracy. To mitigate this, k-fold cross-validation is employed. The dataset is shuffled to reduce variance and minimize overfitting, then divided into k equal folds. In each iteration, one fold is used as the test set, and the remaining k-1 folds are used for training. This process repeats k times, with each fold serving as the test set once. The evaluation scores from all iterations are averaged, providing a comprehensive assessment of the model's performance. This technique ensures every data point is used for training and testing, enhancing the model's robustness and reliability.
4. **Data Building and Training:** The next step involves building and training a deep learning model, specifically an LSTM neural network, to detect phishing websites. This model uses historical data from URLs of known phishing and legitimate sites. URL data is processed to extract features such as domain characteristics, URL structure, suspicious elements, and domain reputation. These features are inputs to the LSTM and dense layers. Training follows a supervised learning approach, where the LSTM learns to distinguish between phishing and legitimate sites based on these features. Various training parameters, including batch size, number of epochs, and model architecture, are meticulously optimized to enhance performance and generalization. This ensures the model is robust and capable of accurately identifying phishing threats in diverse real-world scenarios.
5. **Model Evaluation and Validation:** After training the LSTM model, its performance is rigorously evaluated to ensure effectiveness in detecting phishing websites. The dataset is split into training and testing sets, and the model is tested on unseen data. Key performance metrics, including accuracy, precision, recall, and F1 score, are calculated to assess its ability to identify phishing sites while minimizing false positives and negatives. Cross-validation techniques are employed to validate the model's generalization capability across different data subsets. The model's predictions are compared to actual labels to determine predictive accuracy, and any misclassification patterns are analyzed to further refine the model.
6. **Model Optimization and Fine-Tuning:** In this phase, the model's parameters and architecture are fine-tuned to enhance performance. This involves optimizing the number of layers and units, and tuning hyperparameters such as learning rate, batch size, and epochs using techniques like grid search. Regularization methods, such as dropout and early stopping, are employed to prevent overfitting. Additionally, activation functions and optimization algorithms are refined to improve accuracy and convergence speed. These adjustments ensure the model generalizes well and performs effectively in detecting phishing websites.
7. **Model Deployment:** After training the LSTM model, its performance is rigorously evaluated to ensure effectiveness in detecting phishing websites. The dataset is split into training and testing sets, and the model is tested on unseen data. Key performance metrics, including accuracy, precision, recall, and F1 score, are calculated to assess its ability to identify phishing sites while minimizing false positives and negatives. Cross-validation techniques are employed to validate the model's generalization capability across different data subsets. The model's predictions are compared to actual labels to determine predictive accuracy, and any misclassification patterns are analyzed to further refine the model.

5. SELENIUM MODULE

Selenium is a powerful and widely-used open-source tool for automating web browsers. It enables developers to write scripts in various programming languages, such as Python, Java, and C#, to interact with web elements, perform tasks like clicking buttons, filling out forms, and navigating pages. Selenium supports multiple browsers, including Chrome, Firefox, and Safari, making it versatile for cross-browser testing. It consists of components like WebDriver, which provides a programming interface to create and execute browser automation scripts. Selenium is essential for testing web applications, ensuring they function correctly across different browsers and environments.

Here selenium is utilized to automate web browsing for the purpose of evaluating URLs and detecting phishing websites. The Selenium WebDriver is instantiated, with Chrome WebDriver chosen as the automation tool. The URL of the website to be evaluated, in this case, "https://www.google.com/", is specified, and the WebDriver navigates to this URL using the get() function. A continuous loop is established to monitor the current URL, enabling dynamic evaluation as the WebDriver navigates through

different pages. Once a change in the URL is detected, indicating navigation to a new page, the process of phishing detection commences. By integrating Selenium with the phishing detection logic, the process of URL evaluation and facilitates automated detection of potentially malicious websites.

6. SYSTEM DESIGN

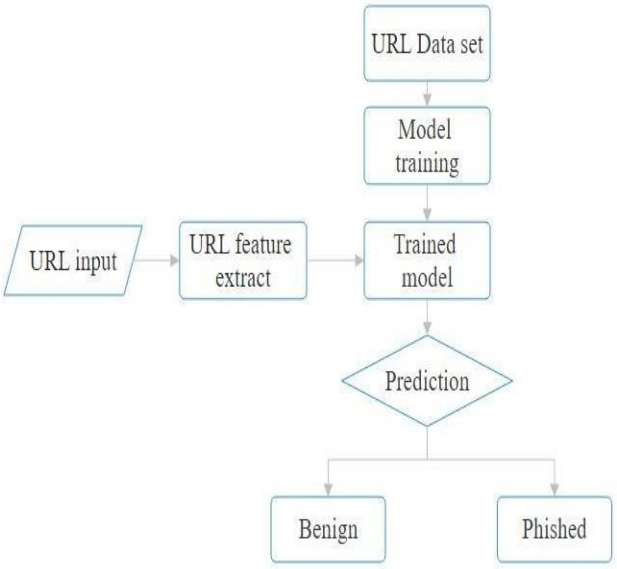

Fig -1: System Framework

This flowchart outlines a process for detecting phishing URLs using a machine learning model. Here's a step-by-step explanation:

1. **URL Input**: The process begins with the input of a URL that needs to be classified as either benign or phished.
2. **URL Feature Extraction**: The input URL undergoes a feature extraction process. During this stage, various features are derived from the URL, such as the address based features, abnormal based features, HTML and JavaScript based features and domain based features that can help in identifying phishing characteristics.
3. **URL Dataset**: Parallel to the URL input and feature extraction, there is an existing dataset of URLs that has already been classified as benign or phished. This dataset is crucial for training the deep learning model. It contains labeled examples that help the model learn the distinguishing features of benign and phishing URLs.
4. **Model Training**: Using the labeled URL dataset, a deep learning model is trained. During training, the model learns to recognize patterns and features that are indicative of phishing URLs versus benign URLs. LSTM (Long Short Term Memory) algorithm is used for training model.
5. **Trained Model**: After the training process is complete, the output is a trained model capable of making predictions. This model now has the knowledge encoded in its parameters to classify new, unseen URLs based on the features it has learned.
6. **Prediction**: The extracted features from the input URL are fed into the trained model. The model processes these features and outputs a prediction on whether the URL is benign or phished.
7. **Output (Benign or Phished)**: The final decision is made based on the model's prediction. The URL is classified into one of two categories:
    - **Benign**: The URL is considered safe and not associated with any phishing activity.
    - **Phished**: The URL is identified as potentially dangerous and likely used for phishing.

This process is crucial for improving cybersecurity measures by automating the detection of phishing attempts, thereby reducing the likelihood of successful phishing attacks. Machine learning models enhance the ability to handle vast numbers of URLs efficiently, providing a scalable solution to the growing threat of phishing in the digital age.

7. RESULTS

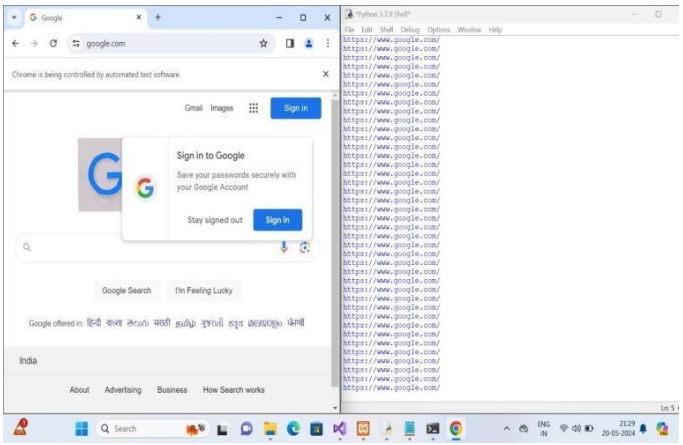

Fig -2: Snapshot of Default Chrome browser in both frontend and backend

In the Fig -2 the chrome is being controlled by automated testing software. Selenium, a web automation tool, is used to interact with a web browser (Google Chrome). The script opens a Chrome browser window and navigates to a predefined URL.

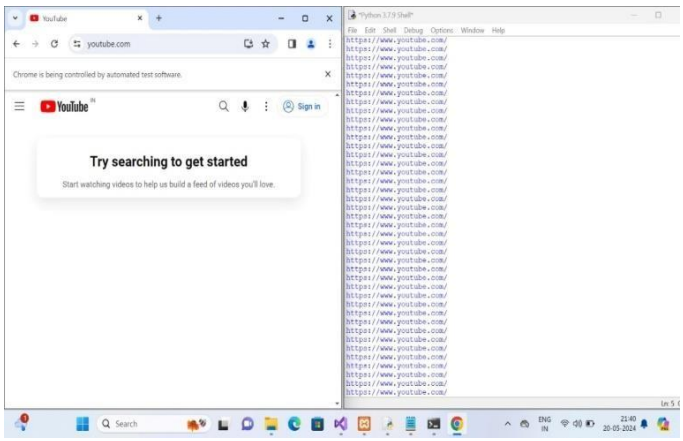

Fig -3: Snapshot of Genuine website (youtube.com)

In the Fig -3 the script enters a continuous loop (while (1)) where it continuously monitors the current URL in both frontend and backend.

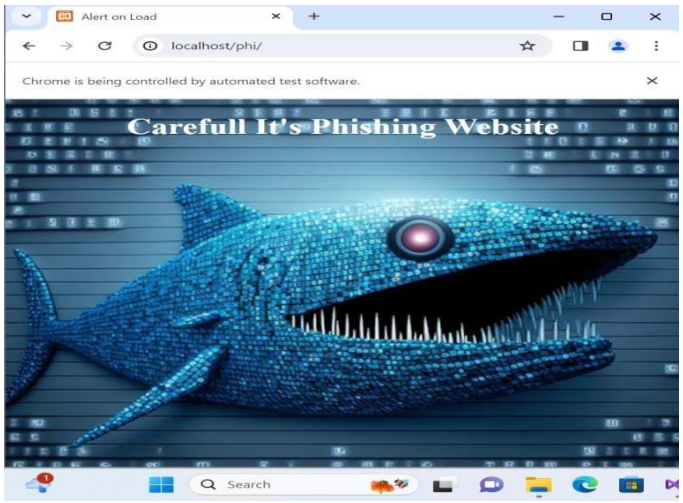

Fig -4: Snapshot of phished website.

In the Fig -4 if a change in the URL is detected, indicating a potential phishing attempt, the script initiates the phishing detection process.

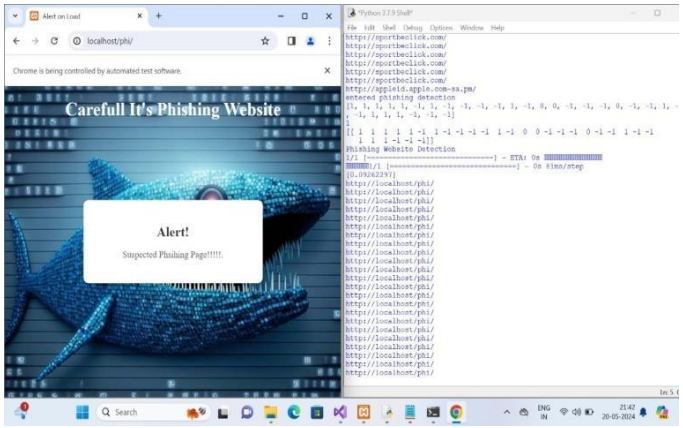

Fig -5: Snapshot of Popup message

In the Fig -5 if the predicted probability of phishing is below a certain threshold (0.96), it takes action by redirecting the user to a predefined phishing alert page (http://localhost/phi/). This action serves as a warning to the user about the potential phishing threat.

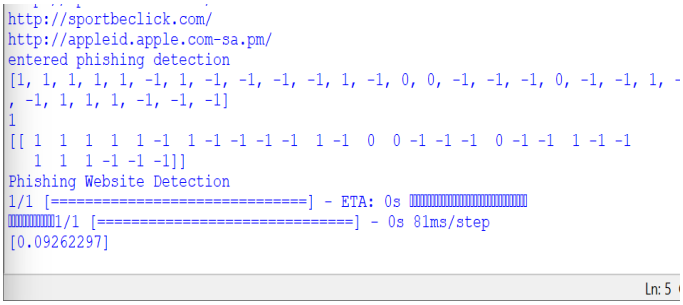

Fig -6: Snapshot of Feature analysis

In the Fig -6, it provides feature analysis of detected phished website.

8. CONCLUSION

Anti-phishing measures are critical in safeguarding users and organizations against malicious attempts to steal sensitive information. Through a combination of technological solutions, user education, and proactive detection methods, the risks posed by phishing attacks can be mitigated effectively. Technologies like deep learning models promising avenues for developing advanced anti-phishing systems capable of detecting and preventing sophisticated phishing attempts. Using LSTM got accuracy which effective than other two models. Deployment is done directly on browser using selenium module. By staying vigilant, adopting best practices, and continuously innovating anti-phishing measures, individuals and organizations can reduce the likelihood of falling victim to phishing attacks and protect their digital assets and privacy.

ACKNOWLEDGEMENT

We are grateful to the Department of Computer Science and Engineering and our institution Jawaharlal Nehru New College of Engineering for providing us the facilities to carry out the research and for imparting us the knowledge with which we can do our best.

Finally, we also would like to thank the whole teaching and non-teaching staff of Computer Science and Engineering Department.

REFERENCES

[1] OrunsoluAbioduna,SodiyaA.Sb, Kareem S.O, “Link Calculator – an efficient link-based phishing detection tool”,Acta Informatica Malaysia, Volume 4, Issue 2, September 2020, p. 37-44.

[2] Md.Faisal Khana, B L Rahab, “Detection of Phishing Websites Using Deep Learning Technique”, Turkish Journal, Volume 12, Issue 10, April 2021, p.3880-3892.

[4] Lizhen Tang, Qusay H. Mahmoud, “A Deep Learning-Based Framework for Phishing Website Detection”, IEEE, Volume 10, December 2021, p. 1509 – 1521.

[5] Rundong Yang, KangfengZheng, Bin Wu, Chunhua Wu and Xiujuan Wang, “Phishing Website Detection Based on Deep Convolutional Neural Network and Random Forest Ensemble Learning”,Sensors (Basel),doi: 10.3390/s21248281, December 2021.

[6] Tristan Bilot, Gregoire Geis and Badis Hammi, "PhishGNN: A Phishing Website Detection Framework using Graph Neural Networks", SECRYPT At Lisbo, DOI:10.5220/0011328600003283, July 2022.
[7] Aman Rangpur, Tarun Kanakam and Dhanvanthini P, "Phish-Defence, Phishing Detection Using Deep Recurrent Neural Networks", Cornell University,Volume 4, September 2022.
[8] Chenguang Wang, Yuanyuan Chen, "TCURL, Exploring hybrid transformer and convolutional neural network on phishing URL detection", Knowledge-Based Systems, Volume 258, Issue C, December 2022.
[9] ZainabAlshingiti ,Rabeah Alaqel ,Jalal Al-Muhtadi ,Qazi Emad UlHaq,Kashif Saleem and Muhammad Hamza Faheem, "A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN", Vol 23,Jan 2023.
[10]ShouqAlnemari, Majid Alshammari, "Detecting Phishing Domains Using Machine Learning", Applied Sciences (2076-3417), Vol. 13, Issue 8, April 2023, p4649. 16p.
[11] Eman Abdullah Aldakheel, Mohammed Zakariah, Ghada Abdalaziz Gashgari, Fahdah A. Almarshad and Abdullah I. A. Alzahrani, "A Deep Learning-Based Innovative Technique for Phishing Detection in Modern Security with Uniform Resource Locators", Sensors, Vol. 23, Issue 9, April 2023.
[12] Manoj Kumar Prabakaran, Parvathy Meenakshi Sundaram ,Abinaya Devi Chandrasekar, "An enhanced deep learning-based phishing detectionmechanism to effectively identify malicious URLs using variational autoencoder",IET Information Security, Volume 17, Issue 3, May 2023, p. 423 – 440.
[13] Wei Wei,qiaoke Jakub Nowak,Marcinkorytkowski, Rafat Scherer, Marcin wonxniak, "Accurate and fast URL phishing detector", Elsevier, Volume 198, September 2020.
[14] Mohamed A. El-Rashidy, "A Smart Model for Web Phishing Detection Based on New Proposed Feature Selection Technique", Menoufia J. of Electronic Engineering Research (MJEER), Vol. 30, No. 1, Jan.2023.
[15] Ahmet Selman Bozkir, Firat Coskun Dalgic, Murat Aydo, GramBeddings: "A New Neural Network for URL Based Identification of Phishing Web Pages Through N-gram Embeddings", Elsvier, Volume 124, January 2023.