International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Machine Learning-Based Detection of Phishing URLs: A Comprehensive Analysis of Features for Reliable Cybersecurity

H. Peda Sydulu[1], Pakala Siddardha Shyam[2], G. Pavan Kalyan[2], Boodida VIdyadhar Reddy[2], Akula Kranthi Sai[2], Akhil Suthrave[2], Bangalore Sri Gouri[2]

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of CSE (Cyber Security)
[1,2]Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal, 501401, Telangana.

ABSTRACT

Phishing is a fraudulent technique that uses social and technological tricks to steal customer identification and financial credentials. Social media systems use spoofed e-mails from legitimate companies and agencies to enable users to use fake websites to divulge financial details like usernames and passwords. Hackers install malicious software on computers to steal credentials, often using systems to intercept username and passwords of consumers' online accounts. Phishers use multiple methods, including email, Uniform Resource Locators (URL), instant messages, forum postings, telephone calls, and text messages to steal user information. Phishing attack is the simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password, and bank account details. Phishing assault is a most straightforward approach to get delicate data from honest clients. Point of the phishers is to obtain basic data like username, secret key, and ledger subtleties. Network safety people are currently searching for dependable and consistent location methods for phishing sites recognition. To overcome the drawbacks of blacklist and heuristics-based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of many algorithms which requires past data to decide or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This work deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. In addition, the main motive of this research is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy of each algorithm.

Keywords: Cyber security, Phishing URLs, Supervised learning, Logistic regression, K-nearest neighbour.

1. INTRODUCTION

Phishing is a fraudulent technique that uses social and technological tricks to steal customer identification and financial credentials. Social media systems use spoofed e-mails from legitimate companies and agencies to enable users to use fake websites to divulge financial details like usernames and passwords [1]. Hackers install malicious software on computers to steal credentials, often using systems to intercept username and passwords of consumers' online accounts. Phishers use multiple methods, including email, Uniform Resource Locators (URL), instant messages, forum postings, telephone calls, and text messages to steal user information. The structure of phishing content is similar to the original content and trick users to access the content in order to obtain their sensitive data. The primary objective of phishing is to gain certain personal information for financial gain or use of identity theft. Phishing attacks are causing severe economic damage around the world. Moreover, most phishing attacks target financial/payment institutions and webmail, according to the Anti-Phishing Working

Group (APWG) latest Phishing pattern studies [1]. In order to receive confidential data, criminals develop unauthorized replicas of a real website and email, typically from a financial institution or other organization dealing with financial data. This e-mail is rendered using a legitimate company's logos and slogans. The design and structure of HTML allow copying of images or an entire website. Also, it is one of the factors for the rapid growth of Internet as a communication medium, and enables the misuse of brands, trademarks, and other company identifiers that customers rely on as authentication mechanisms. To trap users, Phisher sends "spooled" mails to as many people as possible. When these e-mails are opened, the customers tend to be diverted from the legitimate entity to a spoofed website. There is a significant chance of exploitation of user information. For these reasons, phishing in modern society is highly urgent, challenging, and overly critical. There have been several recent studies against phishing based on the characteristics of a domain, such as website URLs, website content, incorporating both the website URLs and content, the source code of the website and the screenshot of the website. However, there is a lack of useful anti-phishing tools to detect malicious URL in an organization to protect its users. In the event of malicious code being implanted on the website, hackers may steal user information and install malware, which poses a serious risk to cybersecurity and user privacy. Phishing assault is a most straightforward approach to get delicate data from honest clients. Point of the phishers is to obtain basic data like username, secret key, and ledger subtleties. Network safety people are currently searching for dependable and consistent location methods for phishing sites recognition. To overcome the drawbacks of blacklist and heuristics-based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of many algorithms which requires past data to decide or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites.

2. LITERATURE SURVEY

Phishing attacks are categorized according to Phisher's mechanism for trapping alleged users. Several forms of these attacks are keyloggers, DNS toxicity, Etc., [2]. The initiation processes in social engineering include online blogs, short message services (SMS), social media platforms that use web 2.0 services, such as Facebook and Twitter, file-sharing services for peers, Voice over IP (VoIP) systems where the attackers use caller spoofing IDs [3, 4]. Each form of phishing has a little difference in how the process is carried out in order to defraud the unsuspecting consumer. E-mail phishing attacks occur when an attacker sends an e-mail with a link to potential users to direct them to phishing websites.

Phishing websites are challenging to an organization and individual due to its similarities with the legitimate websites [5]. There are multiple forms of phishing attacks. Technical subterfuge refers to the attacks include Keylogging, DNS poisoning, and Malwares. In these attacks, attacker intends to gain the access through a tool/technique. On the one hand, users believe the network and on the other hand, the network is compromised by the attackers. Social engineering attacks include Spear phishing, Whaling, SMS, Vishing, and mobile applications. In these attacks, attackers focus on the group of people or an organization and trick them to use the phishing URL [6, 7]. Apart from these attacks, many new attacks are emerging exponentially as the technology evolves constantly. Phishing detection schemes which detect phishing on the server side are better than phishing prevention strategies and user training systems.

3. PROPOSED SYSTEM

3.1 Overview

Throughout the research work, meticulous documentation of all steps, findings, and experimental setups is maintained to ensure transparency and reproducibility. Researchers may also consider further

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

iterations and explore additional machine learning techniques to enhance the accuracy of URL phishing detection. Figure 4.1 shows the proposed system model. The detailed operation illustrated as follows:

Data Preprocessing (Feature Listing): In this phase of the research, a dataset comprising both phishing and legitimate URLs is gathered. This dataset includes various attributes related to the URLs. The research begins by cleaning and organizing the data, addressing missing values, duplicates, and data transformations to ensure it's suitable for analysis. Feature engineering may also be conducted to derive new attributes or extract relevant information from the URLs.

Exploratory Data Analysis (EDA): EDA is a crucial research step that involves delving into the dataset's characteristics. Researchers create visualizations, summary statistics, and plots to gain insights into the distribution of phishing and legitimate URLs, discover potential patterns or correlations, and identify outliers or anomalies.

Data Splitting: After data preprocessing and EDA, the research progresses to data splitting. The dataset is divided into distinct subsets, typically including a training set for model training, a validation set for hyperparameter tuning, and a testing set to assess model performance.

Existing Logistic Regression Classifier (LRC): In this research phase, a logistic regression classifier, a common algorithm for binary classification tasks like phishing detection, is implemented as a baseline model. The logistic regression model is trained on the training data, and its performance is evaluated using appropriate metrics on both the validation and test sets.

Proposed K-Nearest Neighbors (KNN): The research introduces the K-Nearest Neighbors (KNN) algorithm as the proposed approach in this phase. KNN is implemented and trained on the training data. Researchers may perform hyperparameter tuning and assess the model's performance on the validation and test sets, employing the same evaluation metrics as used for the logistic regression model.

Prediction: In the final phase of this research, both the logistic regression and KNN models are employed to make predictions on new or unseen data, which could be the test set or real-world URLs. Researchers evaluate and compare the performance of these models to determine which one is more effective in detecting phishing URLs, based on the chosen evaluation metrics.

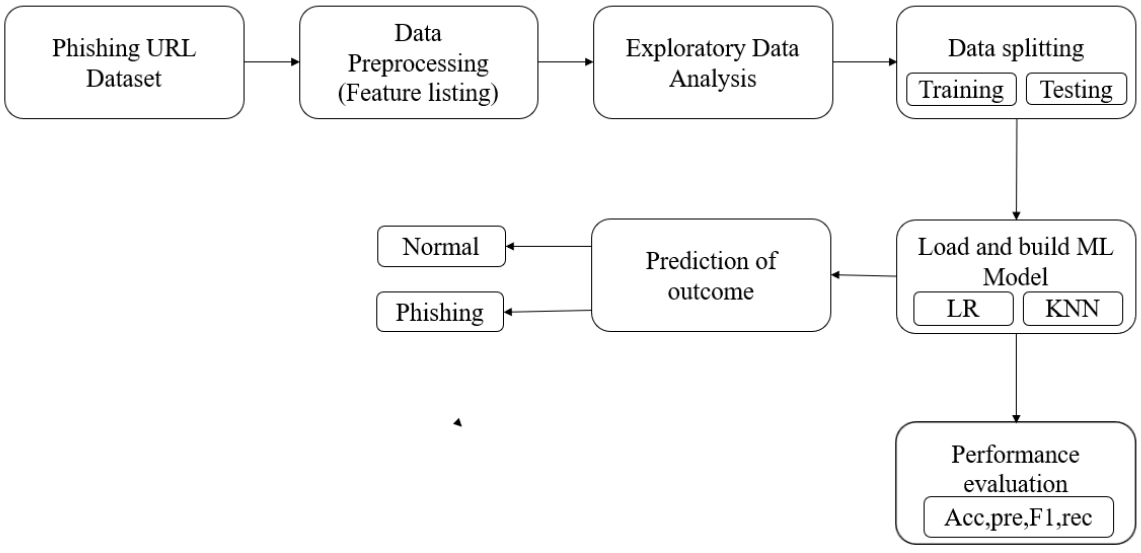

Figure 1. Block diagram of proposed system.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

3.2 KNN

K-Nearest Neighbors (KNN) is a simple yet powerful supervised machine learning algorithm used for classification and regression tasks. It's based on the idea that data points with similar features tend to belong to the same class or have similar values in the case of regression. KNN is a distance-based classification algorithm. It assigns a new data point to the majority class of its k-nearest neighbors. The choice of 'k' (the number of neighbors) is a crucial hyperparameter that impacts the model's performance. KNN is an instance-based learning method, meaning it doesn't build a model during training. Instead, it memorizes the entire training dataset and uses it for predictions.

Working Principle:

Step 1: Distance Metric:

- KNN uses a distance metric (typically Euclidean distance, but others like Manhattan, Minkowski, etc., are also possible) to measure the similarity between data points. The algorithm finds the 'k' nearest neighbors with the smallest distances to the new data point.

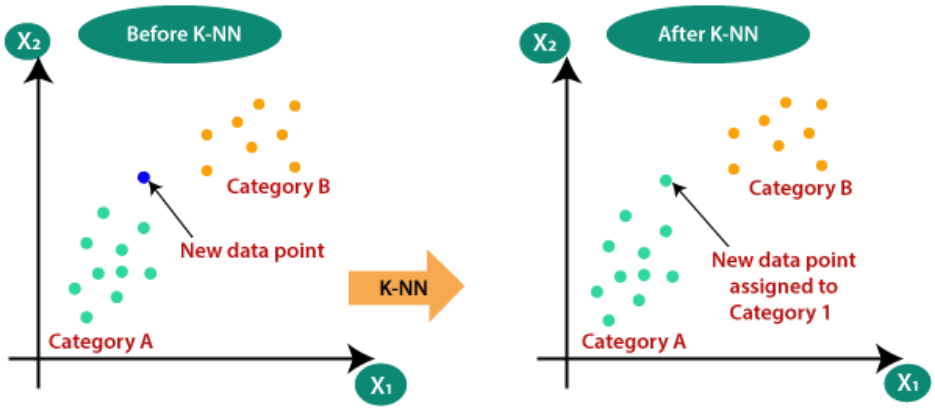

Figure 2. KNN initialization

- Voting Mechanism: For classification, KNN uses a majority voting mechanism among its neighbors. The class that occurs most frequently among the neighbors is assigned to the new data point. For regression, it takes the mean (or median) value of the 'k' nearest neighbors as the prediction.

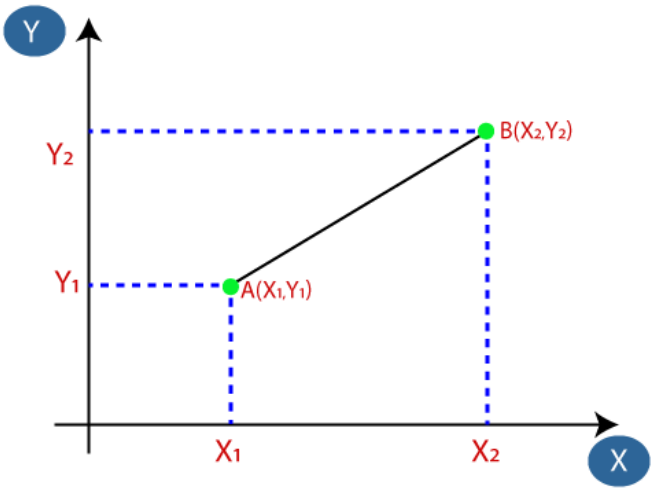

Figure 3. Distance measurement in KNN.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Step 2. Hyperparameter 'k':

- Choosing the Right 'k': The choice of 'k' is crucial. A small 'k' makes the model sensitive to noise and outliers but may capture local patterns well. A large 'k' smooths out local variations but can make the model less accurate.
- Methods for Choosing 'k': Cross-validation, grid search, and domain knowledge are common approaches to determine the optimal 'k' value.
- Simplicity: KNN is easy to understand and implement, making it a suitable choice for beginners.
- No Training Phase: It doesn't require a training phase since it memorizes the data, making it suitable for online learning and non-stationary data.
- Non-Parametric: KNN is non-parametric, meaning it makes no assumptions about the underlying data distribution.
- Works for Multiclass Problems: KNN naturally handles multi-class classification problems.

Step 3. Variants:

- Weighted KNN: Assigns different weights to neighbors based on their distance. Closer neighbors have a greater influence on the prediction.
- KNN with Feature Scaling: Feature scaling is essential when using KNN, as it's distance-based. Standardization (scaling features to have mean=0 and standard deviation=1) is often applied.
- KD-Tree and Ball-Tree: These data structures can be used to speed up KNN search for large datasets.

4. RESULTS AND DISCUSSION

Figure 4 displays a pie chart representing the distribution of the target variable in the dataset. It shows the proportion or percentage of each class (e.g., phishing and non-phishing) in the dataset. Figure 5 provides information about the features (independent variables) used for detecting phishing URLs. It may list or describe the various attributes or characteristics considered in the analysis.

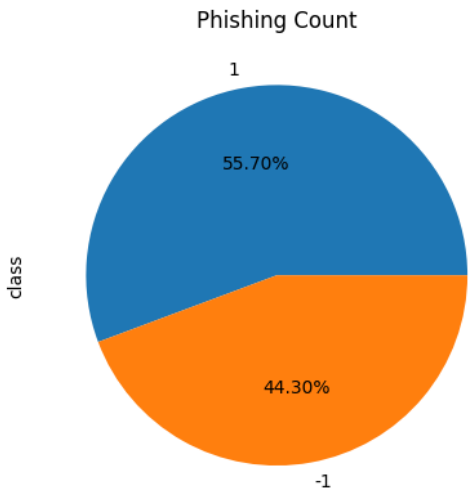

Figure 4: Pie chart of target column of a dataset

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	Favicon	...	DisableRightClick
0	1	1	1	1	1	-1	0	1	-1	1	...	1
1	1	0	1	1	1	-1	-1	-1	-1	1	...	1
2	1	0	1	1	1	-1	-1	-1	1	1	...	1
3	1	0	-1	1	1	-1	1	1	-1	1	...	1
4	-1	0	-1	1	-1	-1	1	1	-1	1	...	1
...	...	...	...	...	...	...	...	...	...	...	...	...
11049	1	-1	1	-1	1	1	1	1	-1	-1	...	-1
11050	-1	1	1	-1	-1	-1	1	-1	-1	-1	...	1
11051	1	-1	1	1	1	-1	1	-1	-1	1	...	1
11052	-1	-1	1	1	1	-1	-1	-1	1	-1	...	1
11053	-1	-1	1	1	1	-1	-1	-1	1	1	...	1

11054 rows × 30 columns

Figure 5: Features of a dataset used for URL phishing detection.

```
0       -1
1       -1
2       -1
3        1
4        1
        ..
11049    1
11050   -1
11051   -1
11052   -1
11053   -1
Name: class, Length: 11054, dtype: int64
```

Figure 6: Target column of a dataset used for URL Phishing detection.

Figure 6 focuses on the target variable in the dataset, which is the variable being predicted or classified. In the context of URL phishing detection, this could show the distribution of phishing and non-phishing labels. Figure 7 presents the classification report for a logistic regression model. A classification report provides metrics such as precision, recall, F1-score, and support for each class in the classification problem.

	precision	recall	f1-score	support
-1	0.94	0.91	0.92	976
1	0.93	0.95	0.94	1235
accuracy			0.93	2211
macro avg	0.93	0.93	0.93	2211
weighted avg	0.93	0.93	0.93	2211

Figure 7: Classification report of Logistic regression

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

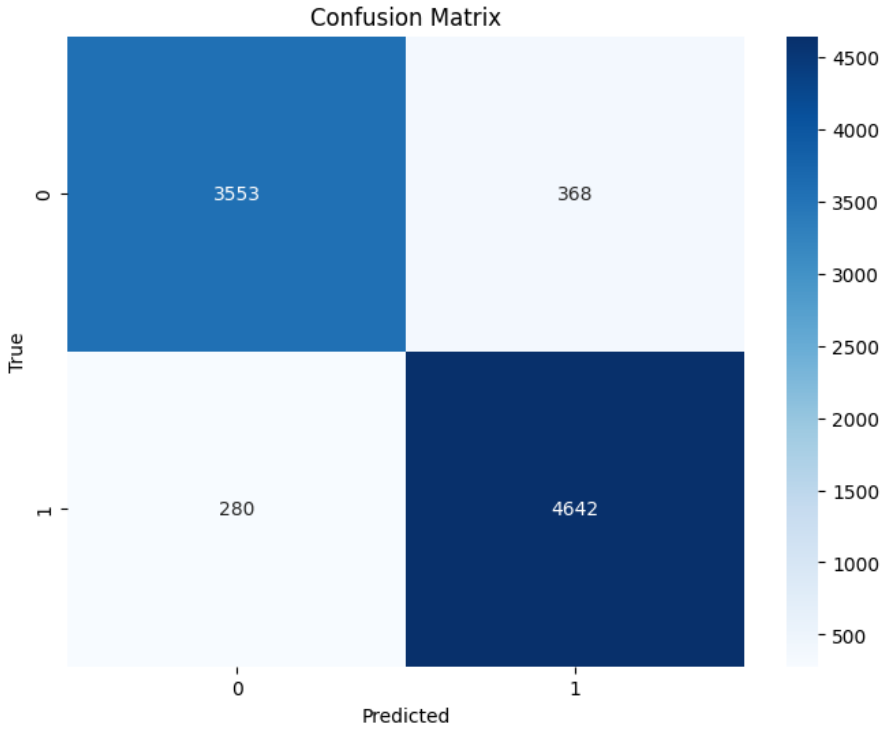

Figure 8: Confusion matrix of logistic regression

Figure 8 displays the confusion matrix for a logistic regression model. A confusion matrix is a table that shows the number of true positives, true negatives, false positives, and false negatives, which are essential for evaluating the performance of a classification model.

Figure 9 presents the classification report for a k-nearest neighbors (KNN) classifier. Similar to Figure 7, it provides metrics like precision, recall, F1-score, and support for each class. Figure 10 displays the confusion matrix for a k-nearest neighbors (KNN) classifier. It provides a detailed breakdown of the model's performance in terms of correct and incorrect classifications. These figures collectively provide a comprehensive overview of the dataset, its features, the relationships between variables, and the performance of the machine learning models used for URL phishing detection.

	precision	recall	f1-score	support
-1	0.95	0.96	0.95	976
1	0.97	0.96	0.96	1235
accuracy			0.96	2211
macro avg	0.96	0.96	0.96	2211
weighted avg	0.96	0.96	0.96	2211

Figure 9: Classification report of KNN Classifier

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

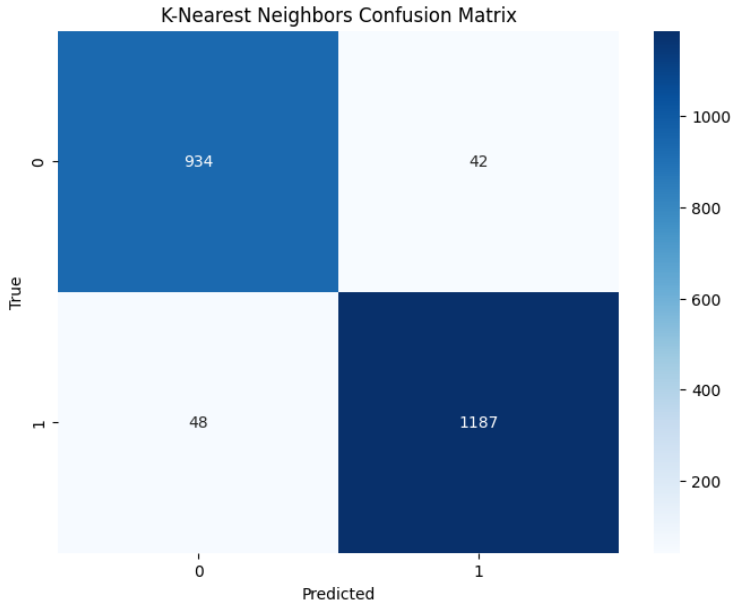

Figure 10: Confusion matrix of KNN classifier.

5. CONCLUSION

A machine learning (ML) based phishing URL was proposed in this work. The investigation utilizes many strategies to identify phishing intrusion detection. Standard datasets of phishing intrusion detection from kaggle.com were used as input for the ML algorithms. The machine learning algorithm KNN are implemented to analyze and select datasets for classification and detection. KNN was used to both classify the website and classification. Finally, the confusion matrix was drawn to evaluate the performance of KNN algorithms. The random KNN achieved a high accuracy.

REFERENCES

[1] Anti-Phishing Working Group (APWG), https://docs.apwg.org//reports/apwg_trends_report_q4_2019.pdf
[2] Jain A.K., Gupta B.B. "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning", Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018,
[3] Purbay M., Kumar D, "Split Behavior of Supervised Machine Learning Algorithms for Phishing URL Detection", Lecture Notes in Electrical Engineering, vol. 683, 2021,
[4] Gandotra E., Gupta D, "An Efficient Approach for Phishing Detection using Machine Learning", Algorithms for Intelligent Systems, Springer, Singapore, 2021, https://doi.org/10.1007/978-981-15-8711-5_12.
[5] Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, "URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection", Conference'17, Washington, DC, USA, arXiv:1802.03162, July 2017.
[6] Hong J., Kim T., Liu J., Park N., Kim SW, "Phishing URL Detection with Lexical Features and Blacklisted Domains", Autonomous Secure Cyber Systems. Springer, https://doi.org/10.1007/978-3-030-33432-1_12.
[7] 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, 10.1109/ICCCI48352.2020.9104161.

[8] Hassan Y.A. and Abdelfettah B, “Using case- based reasoning for phishing detection", Procedia Computer Science, vol. 109, 2017, pp. 281–288.
[9] Rao RS, Pais AR. Jail-Phish: An improved search engine-based phishing detection system. Computers & Security. 2019 Jun 1; 83:246–67.