International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 20 No. 2 (2024)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 20, No. 2, 2024

Research Paper

ENSEMBLE INTELLIGENCE: ADVANCED MACHINE LEARNING FOR PHISHING DETECTION AND URL-BASED THREAT CLASSIFICATION

B Mamatha[1], Thotakuri Sai Kiran[2], Sambeta Srinidhi[2], K Jagadeesh Goud[2], Tanguturi Mounika[2]
[1]Assistant Professor, [2]UG Student, [1,2] Department of Information Technology
[1,2]CMR Engineering College (UGC-Autonomous), Medchal Rd, Kandlakoya, Telangana 501401

Received: 07-3-2024 Accepted: 12-4-2024 Published: 22-4-2024

ABSTRACT

Phishing remains a significant cybersecurity threat, exploiting individuals and organizations by deceiving them into disclosing sensitive credentials such as usernames, passwords, and financial information. Attackers often use phishing URLs that mimic legitimate websites to lure victims into unknowingly submitting their private data. Given the increasing sophistication of these threats and the lack of cybersecurity awareness among users, it is crucial to develop robust detection mechanisms.

This research introduces an ensemble machine learning model for phishing intrusion detection and URL classification, leveraging multiple classification algorithms to enhance accuracy and reliability. The proposed model analyzes key URL-based features to differentiate between legitimate and phishing URLs, providing an adaptive and intelligent detection system. By integrating multiple machine learning techniques, the ensemble approach improves detection rates while reducing false positives, ensuring a more effective defense against evolving phishing threats.

The study aims to develop a client-side phishing detection tool that empowers internet users with real-time protection and assists cybersecurity professionals in identifying phishing patterns. The findings of this research contribute to advancing security measures for online users, mitigating the risks associated with phishing attacks, and fostering a safer digital ecosystem.

KEYWORDS : Phishing Detection, Machine Learning, Ensemble Learning, URL Classification, Cybersecurity

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.

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. These systems can be used either via a web browser on the client or through specific host-site software [8, 9].

3. PROPOSED SYSTEM

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

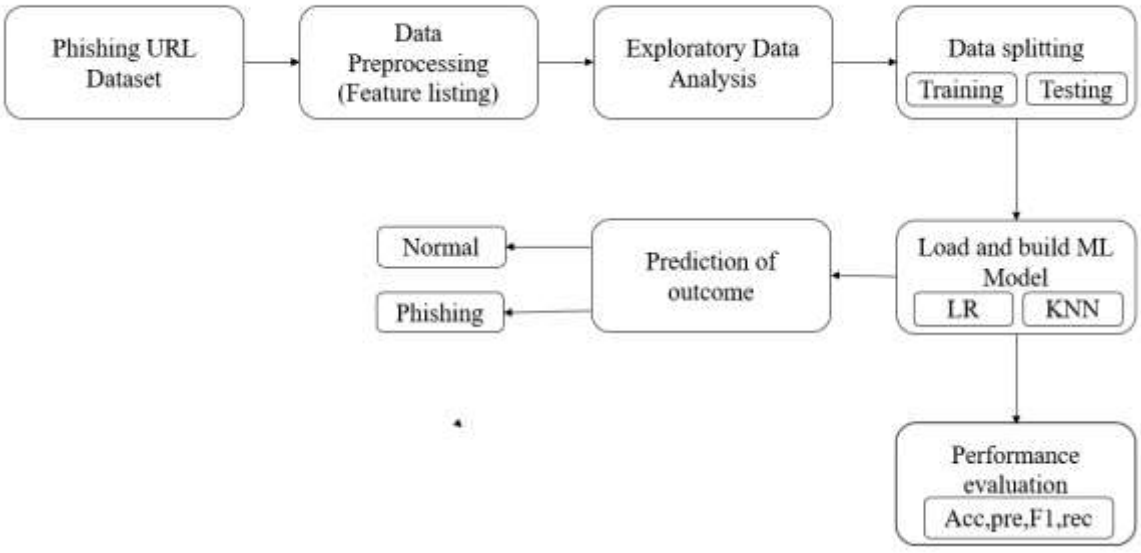

Figure 1: Block diagram of proposed system.

4. RESULTS AND DISCUSSION

The Python implementation focuses on phishing URL detection using machine learning, leveraging datasets containing website features to classify URLs as either phishing or legitimate. It begins with data loading and exploratory data analysis (EDA) to understand feature distributions, correlations, and summary statistics. Data visualization techniques, including heatmaps, pair plots, and pie charts, provide insights into feature relationships and class distributions. The dataset is then split into training (80%) and testing (20%) sets, and **two** classification

models Logistic Regression and K-Nearest Neighbors (KNN) are trained and evaluated using metrics such as accuracy, f1-score, recall, and precision. Model performance is further optimized through parameter tuning (e.g., selecting the best number of neighbors for KNN). The results are stored and visualized **using** confusion matrices and seaborn plots**,** enabling clear comparison between models. This study aims to develop an effective machine-learning-powered phishing detection system, enhancing cybersecurity by identifying malicious URLs before they deceive users.

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

11054 rows × 32 columns

	UsingPopupWindow	IframeRedirection	AgeofDomain	DNSRecording	WebsiteTraffic	PageRank	GoogleIndex	LinksPointingToPage	StatsReport	class
	1	1	-1	-1	0	-1	1	1	1	-1
	1	1	1	-1	1	-1	1	0	-1	-1
	1	1	-1	-1	1	-1	1	-1	1	-1
	-1	1	-1	-1	0	-1	1	1	1	1
	1	1	1	1	1	-1	1	-1	-1	1
	...	...	...	...	...	...	...	...	...	...
	-1	-1	1	1	-1	-1	1	1	1	1
	-1	1	1	1	1	1	1	-1	1	-1
	1	1	1	1	1	-1	1	0	1	-1
	-1	1	1	1	1	-1	1	1	1	-1
	1	1	-1	1	-1	-1	-1	1	-1	-1

Figure 2: Sample dataset used for URL phishing detection

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11054 entries, 0 to 11053
Data columns (total 32 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Index                11054 non-null  int64
 1   UsingIP              11054 non-null  int64
 2   LongURL              11054 non-null  int64
 3   ShortURL             11054 non-null  int64
 4   Symbol@              11054 non-null  int64
 5   Redirecting//        11054 non-null  int64
 6   PrefixSuffix-        11054 non-null  int64
 7   SubDomains           11054 non-null  int64
 8   HTTPS                11054 non-null  int64
 9   DomainRegLen         11054 non-null  int64
 10  Favicon              11054 non-null  int64
 11  NonStdPort           11054 non-null  int64
 12  HTTPSDomainURL       11054 non-null  int64
 13  RequestURL           11054 non-null  int64
 14  AnchorURL            11054 non-null  int64
 15  LinksInScriptTags    11054 non-null  int64
 16  ServerFormHandler    11054 non-null  int64
 17  InfoEmail            11054 non-null  int64
 18  AbnormalURL          11054 non-null  int64
 19  WebsiteForwarding    11054 non-null  int64
 20  StatusBarCust        11054 non-null  int64
 21  DisableRightClick    11054 non-null  int64
 22  UsingPopupWindow     11054 non-null  int64
 23  IframeRedirection    11054 non-null  int64
 24  AgeofDomain          11054 non-null  int64
 25  DNSRecording         11054 non-null  int64
 26  WebsiteTraffic       11054 non-null  int64
 27  PageRank             11054 non-null  int64
 28  GoogleIndex          11054 non-null  int64
 29  LinksPointingToPage  11054 non-null  int64
 30  StatsReport          11054 non-null  int64
 31  class                11054 non-null  int64
dtypes: int64(32)
memory usage: 2.7 MB
```

Figure 3: summary of the dataset used for URL Phishing detection

International Journal of Engineering Research and Science & Technology

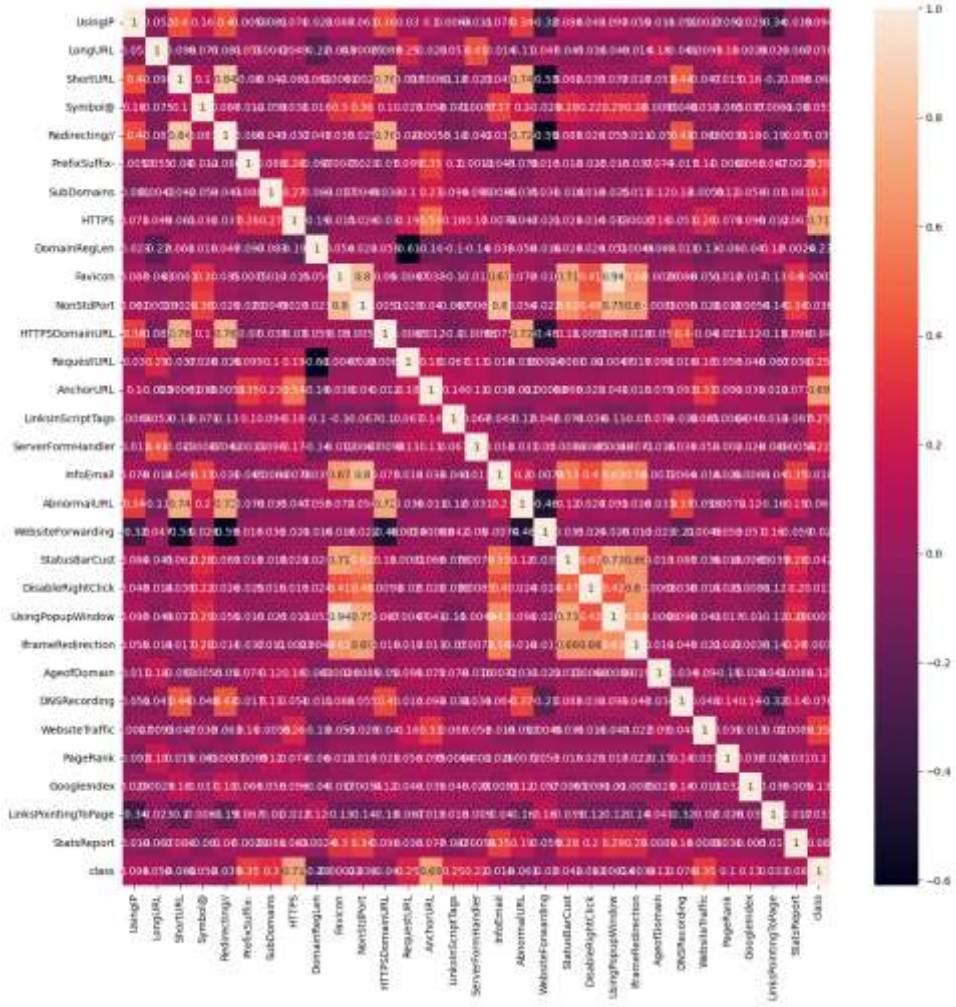

Figure 4: heatmap of correlation of variables in a dataset

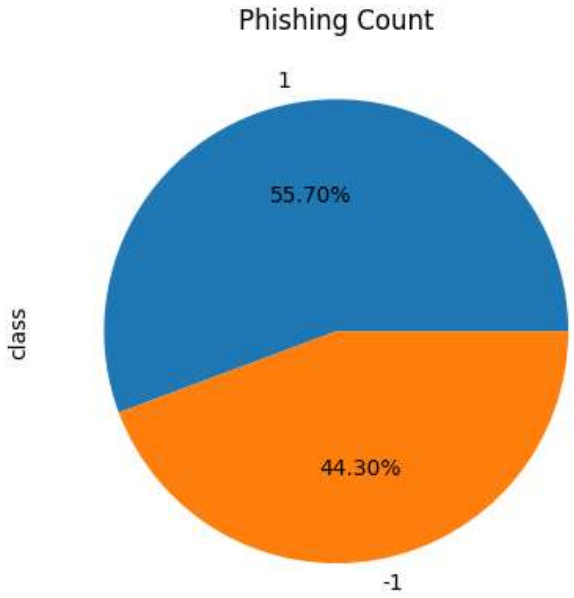

Figure 5: pie chart of target column of a dataset

International Journal of Engineering Research and Science & Technology

	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 6: 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 7: target column of a dataset used for URL Phishing detection

	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 8: classification report of Logistic regression

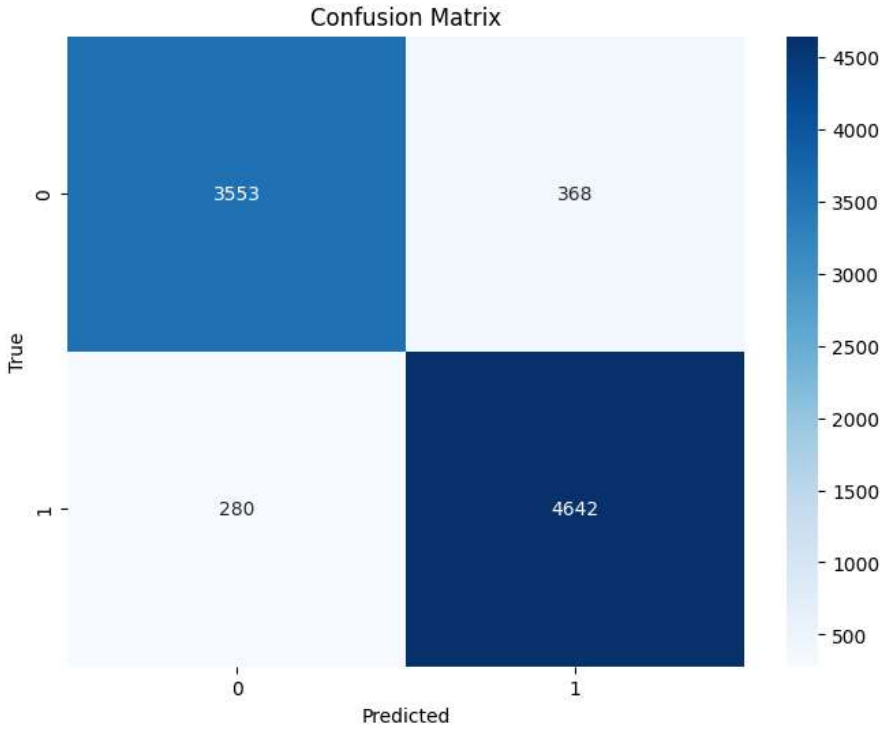

Figure 9: confusion matrix of logistic regression

	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 10: classification report of KNN Classifier

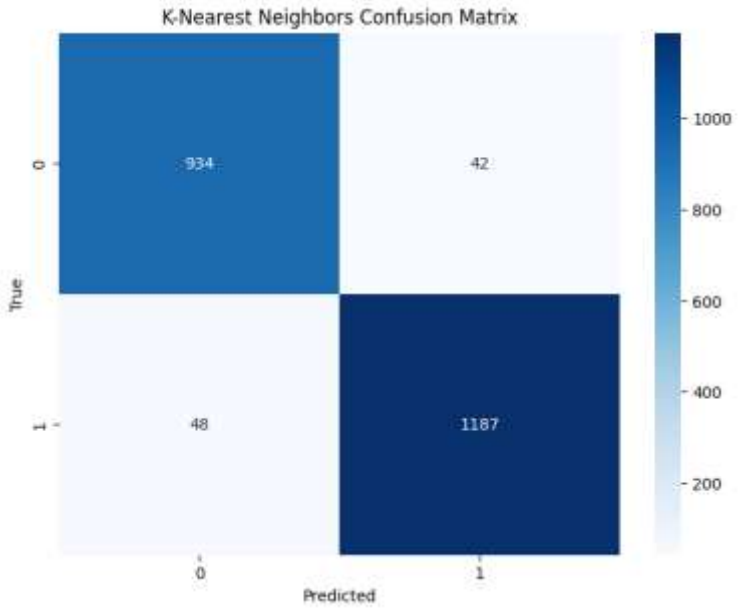

Figure 11: confusion matrix of KNN classifier

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.