Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 44 (2023), No. 3, pp. 315–330
DOI : 10.47974/JIOS-1343

Deep learning based phishing website identification system using CNN-LSTM classifier

Vinod Sapkal *
Department of Computer Science and Engineering
CSMU, Navi Mumbai
Panvel
Maharashtra
India

Praveen Gupta †
Department of Computer Science and Information Technology
CSMU, Navi Mumbai
Panvel
Maharashtra
India

Aboo Bakar Khan §
Department of Electrical and Electronics Engineering
CSMU, Navi Mumbai
Panvel
Maharashtra
India

Abstract

The term phishing refers to an attack that pretends to be the website of a large corporation, typically one dealing with money, such as a bank or other financial institution or an online retailer. Its primary objective is to acquire personally identifiable information from users, such as their social security numbers, credit card information, and passwords. Due to the rise of phishing attacks, various techniques have been developed in order to combat these threats. One of these is deep learning algorithms, which are capable of learning and analyzing massive datasets. Due to their capabilities, these algorithms are very useful in identifying and preventing phishing attacks. Due to the complexity of the phishing

* *E-mail:* **vinod180129@csmu.ac.in (Corresponding Author)**
† *E-mail:* **praveengupta@csmu.ac.in**
§ *E-mail:* **aboobakarkhan@csmu.ac.in**

© Taru Publications

websites, many development systems have been created to detect them. Unfortunately, the output that was desired cannot be achieved by these systems, and they have a number of other flaws as well. The purpose of this paper is to propose a hybrid deep learning-based phishing detection system that is easy to put into practice. The quality of the input dataset is improved through the process of preprocessing the dataset. After that, the procedures of clustering and feature selection are carried out in order to improve the accuracy and decrease the amount of time required for the processing. The resulting features are then fed into the CNN_LSTM, which is a classification system that classifies websites that are phishing and legitimate. Proposed Hybrid deep learning models are proposed to combine the features of natural language processing (NLP) and character embedding. They can then reveal high-level connections between characters. In terms of the metric that is being used for the evaluation, the performance of the models that have been proposed is better than that of the other models.

Subject Classification: *68T07 Artificial neural networks and deep learning, 68T27 Logic in artificial intelligence, 68T45 Machine vision and scene understanding.*

Keywords: *Phishing, CNN, LSTM, NLP.*

1. Introduction

Due to the rapid emergence and evolution of internet technologies, phishing has become one of the most dangerous online threats. The rise of social networking sites, e-commerce, and banking services has increased the number of people who use these platforms. According to a report released by "Wearesocial" in 2021, "the number of internet users has increased by over 300 million in just six years. The penetration rate of the internet is currently at 59.5 percent, which provides an opportunity for criminal organizations to make money from phishing attacks". They are able to con their victims into giving them their personal information by creating a fake website and leading them to believe it is legitimate. When a user inadvertently gives cybercriminals access to their information, such as their personal and financial details, by clicking on a link or updating their sensitive credentials, they are giving the attackers permission to access their information[17]. This data is then used by cybercriminals for various illegal activities. Major reason behind user get affected URLs are usually beyond the comprehension of users and they are unsure of which websites provide reliable sources of information. Redirection or a secret URL may be responsible for concealing the address of the website. People are so busy that they do not have time to accidentally visit websites or look up URLs and unable to tell the difference between legitimate and phishing websites.

Regardless of the type of business it is, phishing attack can affect any organization. It occurs when an individual or group creates a duplicate or malicious website that imitates the look and feel of a legitimate website. Phishing attacks usually target individuals through various channels, such as social networks, email, text messages, and other communication mediums. The objective of these attacks is to obtain sensitive personal information about the victims, such as bank account information, credit card numbers, and passwords. [4].

Phishing attacks usually involve creating forms that try to trick users into entering their personal information. They then allocate the chance of stealthily accessing their data by making them put their real details into a deceptive trap. Attackers can use various methods to carry out their operations, such as email, web pages, and malware.

A new type of phishing attack is called hijacked email reply chains. It involves a hacker gaining access to an individual's email and taking over a legitimate conversation. Then, they forward the message to a friend or colleague with a malicious payload. Since the details of the conversation are real, the recipient is likely to open it. If anyone open the file, it could be infected with Emotet, or another banking Trojan such as Gozi. As the number of phishing attacks continues to grow, it is important that take the necessary steps to prevent your sensitive data from being stolen.

The number of known phishing sites has significantly increased during 2019's first six months. The biggest difference is that the number of sites that use HTTPS to trick their users into thinking they're safe has increased significantly. In 2018, around 15% of these sites used this security technique. By 2019, this figure had risen to 27%. Top phishing affected industries are crypto exchange, Gaming, web-mail, financial institute[5].

There exist two different ways of anti-phishing techniques; One of the most common ways to prevent people from getting phishing is by using a list-based technique. This method involves identifying and monitoring the approved and illegitimate websites. The other is a whitelist-based technique that compares the status of the visited site with the database of whitelisted targets. One of the main limitations of these techniques is that they can't detect legitimate sites that are out of the list. For instance, a blacklist-based technique compares the list of blocked URLs with the suspicious URL to identify phishing sites. There are still websites that are similar to blacklisted sites but are hosted on different domains. These sites can be easily bypassed by users.

Another technique is based on heuristics, which are features that can be used to classify a site based on its abnormal behavior. For instance, the

characteristics of the site such as its links, page rank, and SSL certificates are considered when it comes to identifying phishing sites[6,12]. The attackers can also easily bypass these techniques once they know the details of the phishing scheme. For instance, it's not guaranteed that the phishing features will work in all the sites. This means that the users might end up getting a false positive result if the features are present in the legitimate sites. However, these third-party techniques can still fail to identify new legitimate sites. For instance, they can't detect the original text content of a site when it's embedded in an image or in a web page. The main contribution of our approaches are described as:

1. In this study, hybrid phishing detection models combine the strengths of character embedding and selected features. hybrid phishing detection models are becoming increasingly popular, are proposed and tested. These models can extract deep features by combining and distilling the data from the given elements.
2. The general model in the proposed framework is composed of two different parts: a standard CNN model and a LSTM-based model. After these two components have been connected, the model will transform into a conventional CNN model. The results of this model are then passed into a final node, which is responsible for the categorization of URLs.

2. Related Work

In order to identify non-blacklisted phishing sites, Routhu Srinivasa Rao et. al.[10] utilised a combination of heuristic features, such as source code and URL based detection. The blacklist filter detected over 50% of these malicious phishing sites, which were copies or exact replicas. In order to investigate the role that heuristic and blacklist filters play in the overall performance of an ensemble, the proposed ensemble model incorporated XGBoost, Extra-Tree, and Random Forest into its structure. It functioned effectively and came close to achieving an accuracy of one hundred percent.

In this paper, Ankit Kumar Jain et. al.[1] present a method that can detect phishing attacks by analysing the source code of a website. The proposed method incorporates new web surfing fraud detection features. This method trains machine learning algorithms using hyperlink-specific feature categories. The proposed method is very efficient in identifying

phishing websites. It has a 98.42% accuracy and a 98.39% true positive rate.

M. Somesha et. al.[9] proposed three new phishing site detection models: CNN, DNN, and LSTM (CNN). They assert that these models are capable of achieving an accuracy of 99.52 percent, 99.57 percent, and 99.43 percent, respectively. The fact that the proposed techniques are only implemented with one feature of a third-party service each makes them more secure and accelerates the rate at which they can be discovered.

The authors L. Lakshmi et. al.[7] proposed a feature vector that contains 30 parameters that can be used to identify malignant web pages. They then trained the model with Adam Optimizer using these features. The proposed model can differentiate genuine websites from fraudulent ones. The researchers made the observation that the proposed model was able to achieve an accuracy of 98% when compared to the various machine learning techniques that were used in the detection of web surfing fraud.

The authors Routhu Srinivasa Rao et. al.[10] proposed a new method that can be used to detect phishing sites that are composed of plain text and domain specific content. Multi-modal and ensemble methods tested the model. The domain-specific text model had 99.34% accuracy after evaluation.

In this paper, Alper Ozcan et. al.[2] combine the two feature vector extraction methods. They then proposed a hybrid model that can be used to detect phishing sites. The paper also analysed the performance of the models against different phishing datasets. The authors present a hybrid model that combines the two techniques. They claim that it can achieve an accuracy of 98.79% by exploiting the deep connections between characters. NLP-based model showed high-level connections.

S. Priya et. al.[15] proposed Weighted Fuzzy condense K-modes for clustering. The RBF centres and the kernels of the base classifiers can be found using this method, which was designed specifically for that purpose.. In experiments, the proposed system detected phishing sites better than other ensemble methods.

EPDB, a secure browser architecture by Mohith Gowda HR et. al. [8], improves security and user experience. The new browser detects phishing instantly. This method detected 99.36% of real-time phishing websites. Umer Ahmed Butt et. al.[16] created a feature vector with separated values. They detected phishing sites using Naive Bayes, Long Short-term Memory, and SVM. Their experiment focused on classifying phished emails. The SVM, LSTM, and NB models outperformed the other ensemble methods

in email phishing attack detection. They achieved 99.62%, 97%, and 98% accuracy.

Sumitra Das et. al.[13] developed a hybrid model to detect client-side-only phishing sites. They tested the model on various phishing datasets. The XG Boost model outperformed the other ensemble methods after evaluation.

A.V Ramana et. al.[3] used machine learning methods to find the best classification algorithm. They tested the model against UCI and Mendeley phishing datasets. Researchers achieved 97.51% and 98.45% accuracy after evaluation.

Sajjad Jalil et. al.[14] suggested using the full URL, protocol scheme, and hostname to identify phishing sites. TF-IDF classifies suspicious words and brands. The Random Forest model performed best in their experiments. The 30-feature model achieved 92.2% and 94.65% accuracy on Kaggle datasets. On benchmark datasets, the proposed model outperformed existing methods.

3. URL and attacking techniques

This section, discussed about the techniques that attackers use to avoid detection by security mechanisms and system administrators. First, we should understand the basic structure of a web page's URL. In the example shown in Figure 1 the first part of this process is the establishment of the second level domain, which is a type of organization name that is commonly used in the server hosting. The second part of this process is the placement of the top-level domain, which is a representation of the Internet's root zone. The third part of this process is the creation of the web page's host name, which is a representation of the domain name.

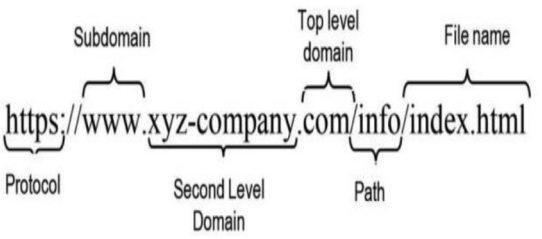


Figure 1

Components of URL

The internal address of a web page is represented by the path that the page takes to get to the server as well as the name of the page written out in HTML format. Even though the name of the second level domain, or SLD, is typically connected to a business or line of work, an adversary can easily find it and use it for phishing if they so choose.

By expanding the path and file names associated with the second level domain (SLD), an attacker can generate an unlimited number of web pages. The actions of the attackers will determine how the inner address design of these web pages is laid out. The domain name of a website is made up of two parts: the top-level domain (also known as the TLD) and the second level domain (also known as the SLD). The URL of a website is one of its most important parts (SLD). If a domain name is found to be being used for phishing, the associated IP address can be blocked so that users are unable to access the websites that are associated with the domain. An attacker can increase the performance of an attack by using various techniques, such as the combination of words and random characters. These methods can also help increase the vulnerability of the victims. Due to these attacks, detection mechanisms should be aware of them

4. Proposed Methodology

Phishers are mainly focused on obtaining sensitive information from an individual by tricking them into entering their cage by replicating a legitimate website. They use various strategies to hide their illegal activities, such as URL obscuring. This paper aims to provide a simple and effective way to identify phishing URLs. Use of the URL string, rather than going to the website itself, is all that is required to carry out this method. This section provides an overview of the various features and performance measurements that the proposed model possesses, as shown in Figure 2.

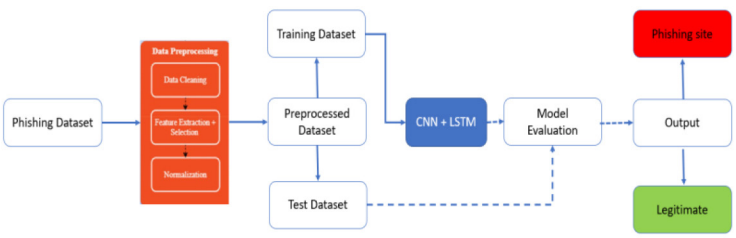


Figure 2

proposed architecture

4.1 *Dataset*

We used standard dataset from Kaggle. Correlation heatmap, Graph of how data is distributed and the how features are related to each other and confusion matrix are depicted by Figure 3 and Figure 4 respectively.

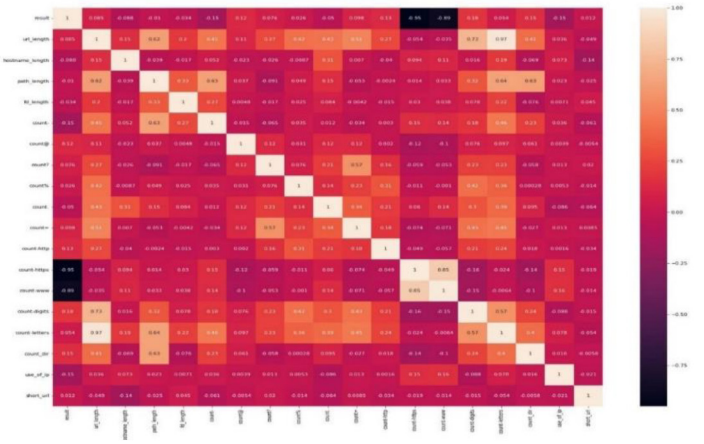


Figure 3

Correlation heatmap

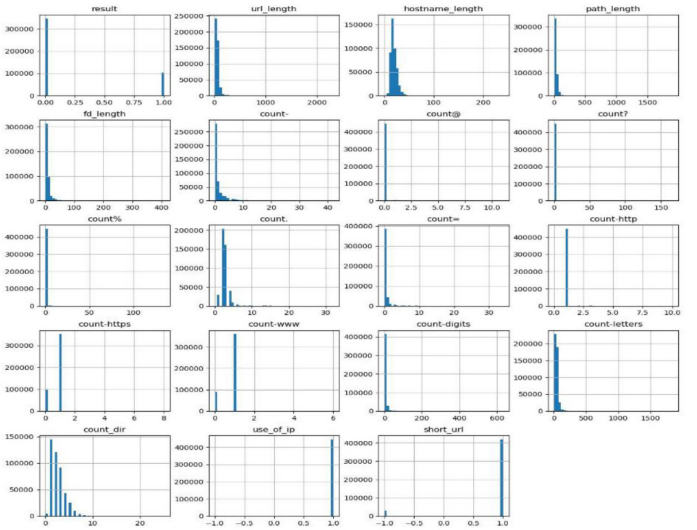


Figure 4

Graph of how data is distributed and the how features are related to each other

4.2 *Pre-Processing*

The initial step in any machine learning process is data preprocessing. This process involves converting the datasets into feature vectors, which can be used by models to interpret them easily. After the datasets have been extracted, the preprocessing steps are then applied to them. Some of the redundant URLs and noise in the data are removed from the datasets.

Redundant and noisy URLs have been removed from the Kaggle datasets. The redundant ones were found in D1 and D2 datasets, which are being further processed. There are also some instances of noise in the data, such as some URLs that are in quotes or separated by commas. Before extracting features from the data, these abnormalities should be taken out.

The process of feature engineering is very important in order to improve the performance of a phishing detection method. Although the previous generation of features was extensive, it took a lot of time to extract them. To solve this issue, we have taken advantage of the URL string itself to extract the features. This method makes the process faster and more efficient.

4.3 *Natural Processing Language*

Data processing can be used to extract various features from a web page. Through the use of natural language processing, these features can be obtained. Usually, these are extracted using the English language. However, they can also be easily adapted to other languages. Most of the time, choosing and designing these features is very trivial. Most of the time, the functions that are selected are mainly focused on identifying phishing threats. The selected features should also be able to parametrize the web page's URL.

The task of finding each word in a web page is trivial. In this process, the URL was analyzed by taking into account special characters such as "?", "/," and "=." Then, a list of words was compiled, which can be combined with two or more distinct words to create a meaningful sentence. The latter is often used by attackers to convince the victims that the page is legitimate.

Different techniques are used by attackers to trick the users into visiting a malicious website. For instance, they can use the public domain names of well-known companies such as Google, Apple, and Gmail. In addition to these, they can also use the keywords secure, login, and account to target their users. This paper presents a variety of additional features that can help prevent people from visiting a phishing website.

Two types of categories for input parameter:

1. Non-natural language.
2. Natural language.

Non-natural language Parameter consist of features that we can extract from the URL are:

1. Length Features
   - Length of URL, Hostname, Path, First Directory, Top-Level Domain
2. Count Features
   - Count of '-', '@', '?', '%', '.', '=', 'http', 'www', 'Digits', 'Letters', 'Number Of Directories'.
3. Binary Features
   - Use of IP or not
   - Use of Shortening URL or not

4.4 *Implemented algorithm*

- **CNN**

  The CNN model contains a number of layers, each of which is dependent on the workings of the convolution process. The first layer, which is the layer that is the lowest in the network, is the one that is in charge of locating the concealed features in the image. The identification of the fundamental characteristics is the responsibility of the subsequent layers, which are the most complicated. Image processing is the application that sees the most frequent use of this model. The convolution layer, the max-pooling layer, the dropout layer, and the multilayer perceptron layer are the four layers that make up the CNN architecture. The topmost layer has a comprehensive connection to the architecture.

- **LSTM**

  The LSTM is an artificial neural network that can be used in deep learning and artificial intelligence. Unlike the feedforward neural networks, which are usually used in these applications, LSTM has feedback connections. This allows it to process large amounts of data. The term LSTM refers to an analogy between a standard RNN and a short-term memory. The connection between the network

and the long-term memory is similar to how synaptic strengths can change in response to training. The activation patterns of the network change once every time-step, similar to how the patterns in the brain's electric firing system change during a short-term memory. The main objective of the LSTM architecture is to provide a short-term memory for RNNs, which can last thousands of timesteps.

An LSTM network is an extension to an RNN model that manages situations in which the model fails to perform as expected. The output of the prior stage is utilised as an input for the stage that is currently being performed. Despite their widespread application in the field of speech recognition, RNNs are unable to retain any information they have previously processed. The RNN model has been further developed into the LSTM network, which is intended to remember long-run dependencies rather than the traditional RNNs.

5. Results

5.1 *Confusion Matrix*

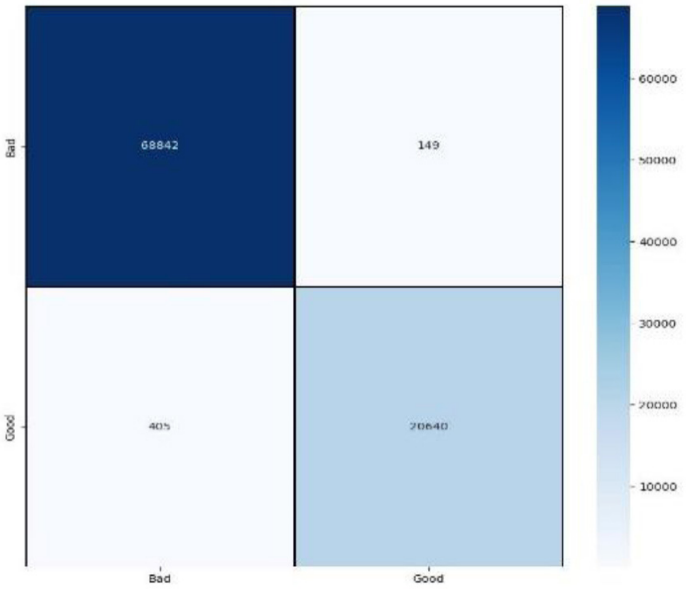


Figure 5
Confusion matrix

5.2 *Model architecture*

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 batch_normalization (BatchN  (None, 174, 1)           4
 ormalization)

 conv1d (Conv1D)             (None, 168, 64)           512

 batch_normalization_1 (Batc  (None, 168, 64)          256
 hNormalization)

 max_pooling1d (MaxPooling1D  (None, 84, 64)           0
 )

 dropout (Dropout)           (None, 84, 64)            0

 lstm (LSTM)                 (None, 100)               66000

 dropout_1 (Dropout)         (None, 100)               0

 flatten (Flatten)           (None, 100)               0

 dense (Dense)               (None, 512)               51712

 batch_normalization_2 (Batc  (None, 512)              2048
 hNormalization)

 dropout_2 (Dropout)         (None, 512)               0

 dense_1 (Dense)             (None, 512)               262656

 batch_normalization_3 (Batc  (None, 512)              2048
 hNormalization)

 dropout_3 (Dropout)         (None, 512)               0

 dense_2 (Dense)             (None, 1)                 513

=================================================================
Total params: 385,749
Trainable params: 383,571
Non-trainable params: 2,178
```

5.3 *Model Accuracy and Model loss*

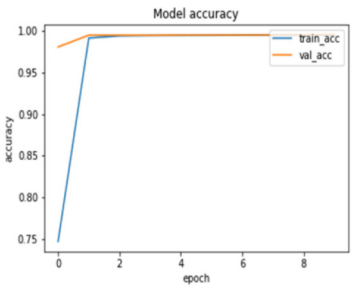


Figure 6
Model accuracy

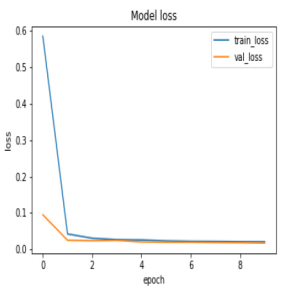


Figure 7
Model loss

5.4 *Precision vs. Recall and FPR (False Positive rate) vs TPR (True Positive Rate)*

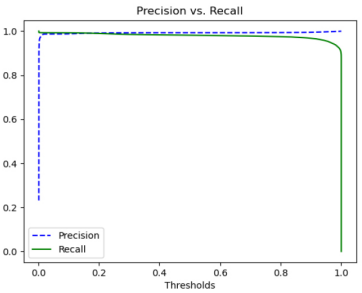


Figure 8

Precision vs Recall

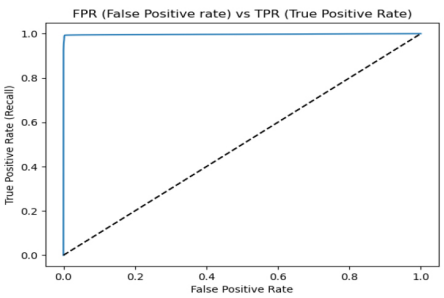


Figure 9

FPR vs TPR

5.5 *Evaluation Matrix*



Figure 10

Evaluation matrix for good vs bad URLs

6. Discussion and Conclusion

Figure 5 represent the confusion matrix between good and bad URLs, Figure 6 represent the model architecture. Figure 7 and Figure 8 represent the model accuracy and loss. Figure 9 and Figure 10 represent precision vs recall and TPR vs FPR respectively. The paper proposes a hybrid model of CNN-LSTM that combines the features of both character embedding and hand-crafted natural language processing (NLP) features for detecting

phishing URLs. The proposed models can find deep connections between the characters and their meanings. The results of the study as shown in Figure 10 revealed that the proposed models provided better performance than the other ones.

References

[1] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 5, pp. 2015-2028 (2019), doi: 10.1007/s12652-018-0798-z.

[2] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, "A hybrid DNN–LSTM model for detecting phishing URLs," *Neural Comput. Appl.*, vol. 0123456789 (2021), doi: 10.1007/s00521-021-06401-z.

[3] A. V. Ramana, K. L. Rao, and R. S. Rao, "Stop-Phish: an intelligent phishing detection method using feature selection ensemble," *Soc. Netw. Anal. Min.,* vol. 11, no. 1, pp. 1-9 (2021), doi: 10.1007/s13278-021-00829-w.

[4] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," *Neural Comput. Appl.,* vol. 28, no. 12, pp. 3629-3654 (2017), doi: 10.1007/s00521-016-2275-y.

[5] https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/" .

[6] J. Anitha and M. Kalaiarasu, "A new hybrid deep learning-based phishing detection system using MCS-DNN classifier," *Neural Comput. Appl.,* vol. 34, no. 8, pp. 5867-5882 (2022), doi: 10.1007/s00521-021-06717-w.

[7] L. Lakshmi, M. P. Reddy, C. Santhaiah, and U. J. Reddy, "Smart Phishing Detection in Web Pages using Supervised Deep Learning Classification and Optimization Technique ADAM," Wirel. Pers. Commun., vol. 118, no. 4, pp. 3549–3564 (2021), doi: 10.1007/s11277-021-08196-7.

[8] M. G. Hr, A. Mv, S. Gunesh Prasad, and S. Vinay, "Development of anti-phishing browser based on random forest and rule of extraction framework," Cybersecurity, vol. 3, no. 1, pp. 1-14 (2020), doi: 10.1186/s42400-020-00059-1.

[9] M. Somesha, A. R. Pais, R. S. Rao, and V. S. Rathour, "Efficient deep learning techniques for the detection of phishing websites," Sadhana - Acad. Proc. Eng. Sci., vol. 45, no. 1, pp. 1-18 (2020), doi: 10.1007/s12046-020-01392-4.

[10] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," Neural Comput. Appl., vol. 31, no. 8, pp. 3851-3873 (2019), doi: 10.1007/s00521-017-3305-0.

[11] R. S. Rao and A. R. Pais, "Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 9, pp. 3853-3872 (2020), doi: 10.1007/s12652-019-01637-z.

[12] S. Anupam and A. K. Kar, "Phishing website detection using support vector machines and nature-inspired optimization algorithms," Telecommun. Syst., vol. 76, no. 1, pp. 17-32 (2021), doi: 10.1007/s11235-020-00739-w.

[13] S. Das Guptta, K. T. Shahriar, H. Alqahtani, D. Alsalman, and I. H. Sarker, "Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques," Ann. Data Sci. (2022), doi: 10.1007/s40745-022-00379-8.

[14] S. Jalil, M. Usman, and A. Fong, "Highly accurate phishing URL detection based on machine learning," *J. Ambient Intell. Humaniz. Comput.*, no. 0123456789 (2022), doi: 10.1007/s12652-022-04426-3.

[15] S. Priya, S. Selvakumar, and R. L. Velusamy, "Evidential theoretic deep radial and probabilistic neural ensemble approach for detecting phishing attacks," *J. Ambient Intell. Humaniz. Comput.*, no. Kaspersky (2020, 2021), doi: 10.1007/s12652-021-03405-4.

[16] U. A. Butt, R. Amin, H. Aldabbas, S. Mohan, B. Alouffi, and A. Ahmadian, "Cloud-based email phishing attack using machine and

deep learning algorithm," Complex Intell. Syst. (2022), doi: 10.1007/s40747-022-00760-3.

[17] V. Sapkal and N. More, "An Improved Classification Model For Identifying The Phishing Attacks," vol. 18, no. 6, pp. 7056-7062 (2021).

[18] Amit Kumar Gupta, Pushpa Gothwal, Dinesh Goyal & Carlos M. Travieso-Gonzalez. IoT-Galvanized pandemic special E-Toilet for generation of sanitized environment, *Journal of Discrete Mathematical Sciences and Cryptography* (2022), DOI: 10.1080/09720529.2022.2068607.

[19] Amit Kumar Gupta, Vijander Singh, Priya Mathur & Carlos M. Travieso-Gonzalez. Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in Indian scenario, *Journal of Interdisciplinary Mathematics*, (2020), DOI: https://doi.org/10.1080/09720502.2020.1833458.

[20] Jitendra Singh Yadav, Amit Kumar Gupta & Arjit Saxena. A review on gender identification using machine learning based on fingerprints, *Journal of Information and Optimization Sciences*, 40:5, 1121-1129 (2019), DOI: https://doi.org/10.1080/02522667.2019.1638002.