How to Cite:
Varuna, W. R., Harshini, M., & Baby, K. T. (2022). An intelligent forecasting system for unauthorized URL identification using deep learning. *International Journal of Health Sciences*, *6*(S1), 7832–7842. https://doi.org/10.53730/ijhs.v6nS1.6725

An intelligent forecasting system for unauthorized URL identification using deep learning

W. Rose Varuna
Assistant Professor, Department of Information Technology, Bharathiar University, Coimbatore, Tamilnadu, India
Email: hvaruna@gmail.com

M. Harshini
Student (II M.Sc), Department of Information Technology, Bharathiar University, Coimbatore, Tamilnadu, India
Email: harshininyedha@gmail.com

K.T. Baby
Student (II M.Sc), Department of Information Technology, Bharathiar University, Coimbatore, Tamilnadu, India
Email: babythangavel09@gmail.com

***Abstract*---**A cyber-attack is a destructive action carried out in cyberspace by an intruder, a group, an association, or a country that targets computer information systems, infrastructures, personal computers, and servers in order to carry out harmful operations or impair their functioning. Attackers produce and sell a wide range of tools for launching web application attacks and increasing the quantity of destructive attacks. Dynamic classifiers continuously watch the web page's activities, whereas classifiers download the page, extract the attributes, and do classification analysis. This article covers how to identify a fraudulent person using a URL and other pertinent data. The process of selecting features from a URL is accomplished using PCA, and the collected characteristics are used to classify URL types. The categorization is done with 98.5 percent accuracy using the DBN approach, which distinguishes between normal and benign users. Other existing techniques are outperformed by the feature selection technique PCA with DBN classification.

***Keywords*---**Cyber-attack, URL, classification, neural network, malicious user, PCA, and accuracy.

Introduction

The cyber-attack is a type of disgusting activity in the cyberspace by an individual, group, association and nations that targets the computer information systems, infrastructures, personal computers, servers to perform malicious activities or to degrade their performance. The cyber-attack belongs to various categories like espionage, cyber warfare, cyber terrorism etc. Unfortunately web is used as a platform to carry out criminal activities to steal the valuable information (like online trading / banking credentials) of the user [1]. The attacker exploits the user's machine and uses it for further attacks through remote administration. Web users are suffering due to various web attacks like phishing, Bot attacks, DDoS, etc. The user visiting the websites for various purposes like online banking, downloading video, sending e-mail and reading newspaper becomes the victim of the cyber-attacks [2].

The lack of awareness and an increase in the number of web users are the main factors for the cyber criminals to generate various malicious attacks. They attract the user to visit malicious web sites to carry out attack. When the user visits the malicious websites they become the victim of the phishing attack. The attacker exploits the weakness and browser plug-inns to inject the malware in the victim's computer [3]. To improve the success rate to malicious activities the criminals employ various methods like free software download, pay-per install service, malware and spam mails. These attacks target information, infrastructure and financial gain. By using these attacks the cyber criminals steal the crucial information from an organization like username, login credentials, significant information of the target organization and military details. The attack against infrastructure happens in various levels. It spoils from the user's machine to the entire network [4, 5].

Currently the attackers also target the industrial control systems for automated industrial machines. The attacks against infrastructure are increasing in the present scenario. Another type of attacks is carried for financial gain targeting an individual or an organisation.Various types of cyber-attacks are growing rapidly. The cyber terrorism is carrying out violence in cyber space. It leads to loss of human life. It is also considered as an act of internet usage for terrorist activities [6]. It includes damaging the large infrastructure of computer networks using computer viruses [7, 8].

The cyber warfare involves the internet as a battle space. The attackers hack the government's network to steal data or destroy the entire network. The cyber bulling involves spreading gossips using e-mail or societal networking web sites, posting embracing pictures, videos and generating bogus identities [9]. The spying of others on the internet is called espionage. It is a process of acquiring the private information without the knowledge of the proprietor. It is known as intelligent information gathering. It is also considered as an effort associated with nations to spy their enemies for military purpose. But this approach is vastly used by the cyber criminals against the internet users [10].

The cyber sabotage is a critical risk in the cyber world. It damages the information or equipment. It is also known as disruptive hacking. These attacks

are of serious concern to the user. In the past, to carry out the malicious activity the attacker required unique skills. They should have knowledge in networking, linux, virtualization, scripting, database application, and security concepts and technologies [11]. The intention behind the attack is to prove the skills and become fame. But in the current scenario most of the attacks are carried out to attain the financial gain or to take revenge [12].

Cyber-attacks are carried out between rivalry nations. The attackers develop and sell various tools to launch attacks in web application and increase the malicious attacks. Nowadays the criminals also use the cloud as platform to carry out cyber-attacks. Hence the cyber-attacks are disruptive and causes harm to users. The society depends on cyber space to carry out various activities and therefore it is essential to protect the cyber space from various types of attacks [13].Most of the classification techniques consume more time to detect malicious web pages and also affect the performance of the client's system.

The classifiers download the page, extract the features, and perform an analysis for classification and dynamic classifiers continuously monitor the behaviour of the web page. Since the classification techniques are complex and to overcome the above issues, the method to detect the malicious URLs is essential. It is a light weight approach. The static analysis is performed with significant features efficiently. The dynamic behaviour based detection for malicious URLs consumes less time when compared with dynamically detecting malicious web pages [14]. Apart from that, malicious URL detection prevents the attack before it causes any damage to the user. So Malicious URL detection is an apt approach to prevent criminal activities in cyber space [16].

Contribution

The main contribution of the research work is

- To develop an optimization approach for selecting optimal features from the dataset.
- To develop neural network for minimizing the error during the process of training.

The remainder of the article is organized as follows: the related works are discussed in Section 2, the proposed methodology is detailed in Section 3, the results are illustrated in Section 4, and the article is concluded in Section 5.

Related Works

Detecting malicious URL is an essential task. The detection method analyses various features, data collection methods and classification techniques. Bhagyashree E. Sananse and Tanuja K [16] proposed feature based approach to classify the malicious URLs. Lexical features, WHOIS features, PageRank and Alexa rank and PhishTankbased features are used for classification. Web mining heuristics on Random Forest algorithm is used to classify phishing URLs. S. CarolinJeeva and Elijah Blessing Rajsingh [17] analyse the features of the URL using associative rule mining algorithm.

The features are transport layer security, unavailable top level domain in the URL and keyword in the path token of the URL which are found to be sensible indicators for phishing URL. A. Le et al[18] obtains new lexical feature and heuristics algorithm to protect against the attack. Fuqiang Yu [19] proposed a malicious URL recognition using BM pattern matching method. This method compares URL source code with the virus features in the database to classify the URL is genuine or malicious.

A malicious URL detection using instant messaging is proposed by D.J.Guan et al [20]. This method analyses the anomalies of URL messages and sender's behaviour. Malicious behaviours are clustered in several behavioural patterns. It identifies the malicious features of the URL. This method is not accurate in detecting malicious URLs.Occurrence of redundant and replicated feature can influence the classification.Incidence of error or complex training process can influence the performance of fake URL classification accuracy. By considering the drawbacks, an efficient framework is developed and discussed in subsequent section.

Proposed Methodology

The process of URL acquisition and the classification of acquired URL attained for fake URL detection that is discussed in this section.

Pre-processing

The internet creates a lot of text data, some of which may contain HTML elements. These HTML elements add no value to text data and are solely used to ensure that the browser renders it correctly. The research work utilise the raw text underneath to encompass all of the utilisation of cases and it can be discussing in this research work (the text is based on one of the Amazon food review or dataset's comments). This research has changed the wordings to account for all possible scenarios). Expanded, contracted, and stop words are removed whereby the lemmatization is attained to normalize the URL. The normalization of the data has eliminated all of the "noise" from the textIn a text, a word may appear in many forms, such as walk and walked (past participle) or price and prices (plural). Text normalisation transforms word variants into the root version from same term.

Feature Selection-Principal Component Analysis

The pre-processed URL is considered for feature selection that is done using PCA. The process of feature selection is accomplished with the PCA and the main intent is to minimize the dimensionality of URL. For the eigenvector T of a covariance matrix in PCA and the feature extraction outcome with relevant to T of vector of arbitrary sample AS is given as,

$$z = AS^{v}T = \sum_{i=1}^{DS} AS_i T_i$$

where $T=[T_1,T_2,....T_{DS}]^v$, $AS=[AS_1,AS_2,....AS_{DS}]^v$, and DS denote the dimensionality of sample vectors.

It is obvious that the absolute value of T_i=(i=1,2....N) may be used to statistically analyse the relevance of the i^{th} feature element of samples to the feature extraction outcome. The smaller the exact value of T_i, the less the significance of the i^{th} feature aspect of samples is easier to understand. Certainly, if the complete value of T_k is small enough, eliminating AS_kT_k from $\sum_{i=1}^{DS} AS_iT_i$ will almost not take impact on the feature extraction outcome. When a feature aspect isn't necessary for feature extraction, the procedure might assume it's not relevant in the original set.

As a consequence, if T_k has a small adequate absolute value, it may assume that the k^{th} feature part of samples is unimportant and can be removed. It is also recognised that there are always many eigenvectors, thus recommends considering several eigenvectors when assessing the relevance of a single feature component.

Classification

The building block of Deep Belief Networks (DBNs) is Restricted Boltzmann Machines (RBMs). RBMs is considered as an energy based models, with the utilization of energy function Eg(vi,hl) is determined as

$$E(vi, hl) = -b'_{vi}vi - b'_{hl}hl - hl'Wvi$$

where hidden and visibly connected units are depicted by weights (W) that is denoted as hl and vi respectively. The occurrence of bias in the hidden and visible layers are denoted as bhi and bvi.

Block Gibbs sampling, in which visible units are sampled consecutively with specified hidden unit values are utilised to get samples from an RBM. Furthermore, given the apparent unit values, hidden units are sampled concurrently. As a result, every single point in the Markov chain is given as:

$$hl^{(n+1)} = \sigma(W'vi^{(n)} + b_{hl})$$
$$vi^{(n+1)} = \sigma(Whl^{(n+1)} + b_{vi})$$

where the sigmoid function (σ) is applied to the activations of the (n + 1)th visible and hidden unit. To effectively sample the tweets from p(vi, hl) during the process of learning, many methods have been created for RBMs, the most successful of which is the contrastive divergence (CD-k) technique.

Deep Belief Networks (DBNs) may be formed by stacking and training RBMs greedily. DBNs are graphical scheme that learn to derive a deep hierarchical description for the training data using machine learning. They are utilized in the following model to represent the joint distribution of the observable vector ov and the l number of hidden layers hl^z:

$$p(vi, hl^1, \dots\dots\dots, hl^l) = (\prod_{z=0}^{l-2} P(hl^z | hl^{z+1}) P(hl^{l+1}, hl^l)$$

where conditional distribution for the visible units in the classification on the hidden unit of RBM at the level z is denoted as vi=hl0 and $P(h^{z-1} | h^z)$ and visible-hidden joints in the RBM top-level is denoted as $P(h^{l-1} | h^l)$. The block diagram of proposed methodology is given in Figure 1.

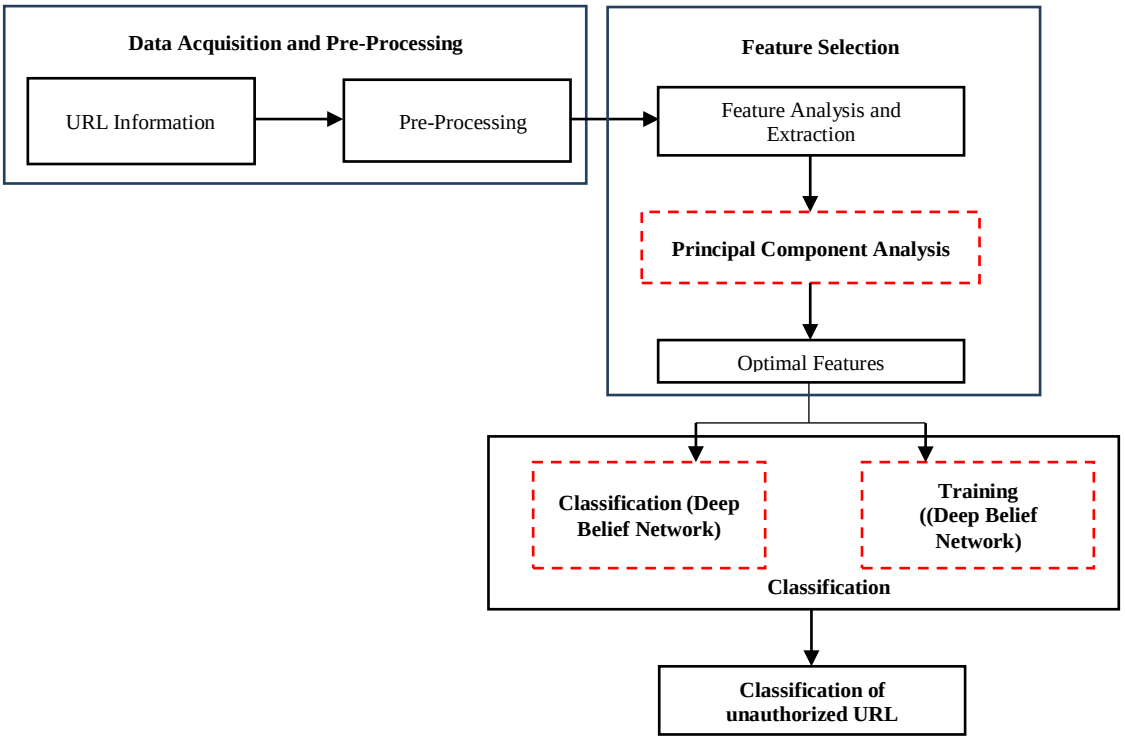

Figure 1. Block Diagram of URL Classification

Result and Discussion

The URL given as input is collected using the browser. The URL is first checked against the system's restricted profile. The browser prohibits the URL from being processed further and alerts the user if the URL fits the profile. However, the URL is sent to the anomaly detector for investigation. This method keeps the system from repeating the process of assessing the same URL and improves the efficiency of the client computer on which it is installed. The proposed approach is compared with Random Forest (RF) and Content based Method (CBM) [16] whereby the evaluation is attained using performance metrics namely Accuracy, Precision, and Recall.

Accuracy

The classification accuracy of the URL is calculated by dividing the number of appropriate fake URL identifications by the total number of URL. Comparison of accuracy is given in Table 1 and Figure 2. The estimation of the accuracy is given as,

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative(TN)}{True\ Positive(TP) + True\ Negative(TN) + False\ Positive(FP) + False\ Negative(FN)}$$

Table 1
Comparison of URL Classification Accuracy

S.No	Iteration	RF	CBM	DBN
1	50	84	77	94
2	100	84.5	79	96
3	150	86	81	97
4	200	87	83	98
5	250	88	85	98.5

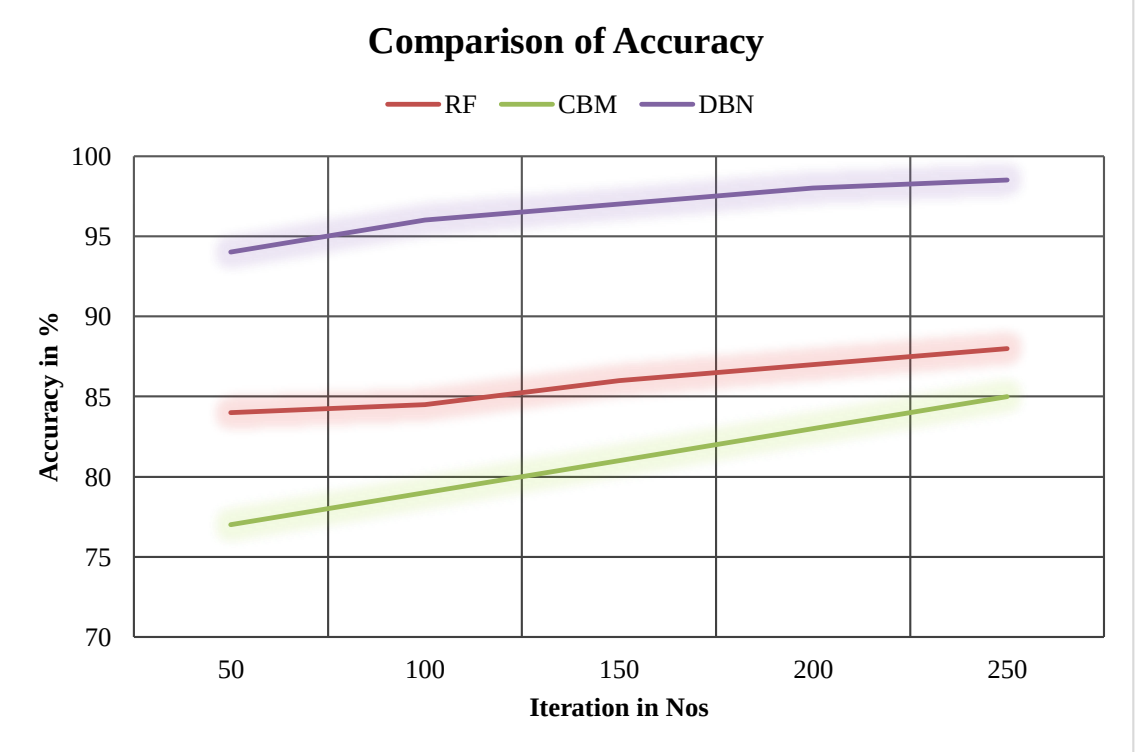

Figure 2. Comparison of URL Classification Accuracy

Precision

The quantitative rate with positive results, also known as precision, reflects the reliability of the prediction and the relevance of the feature found. Comparison of precision is given in Table 2 and Figure 3. It is equated as

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Table 2
Comparison of URL Classification Precision

S.No	Iteration	RF	CBM	DBN
1	50	81	74	94.5
2	100	82	74	95
3	150	84	76	95.5
4	200	84.5	77	96
5	250	85	77.5	97

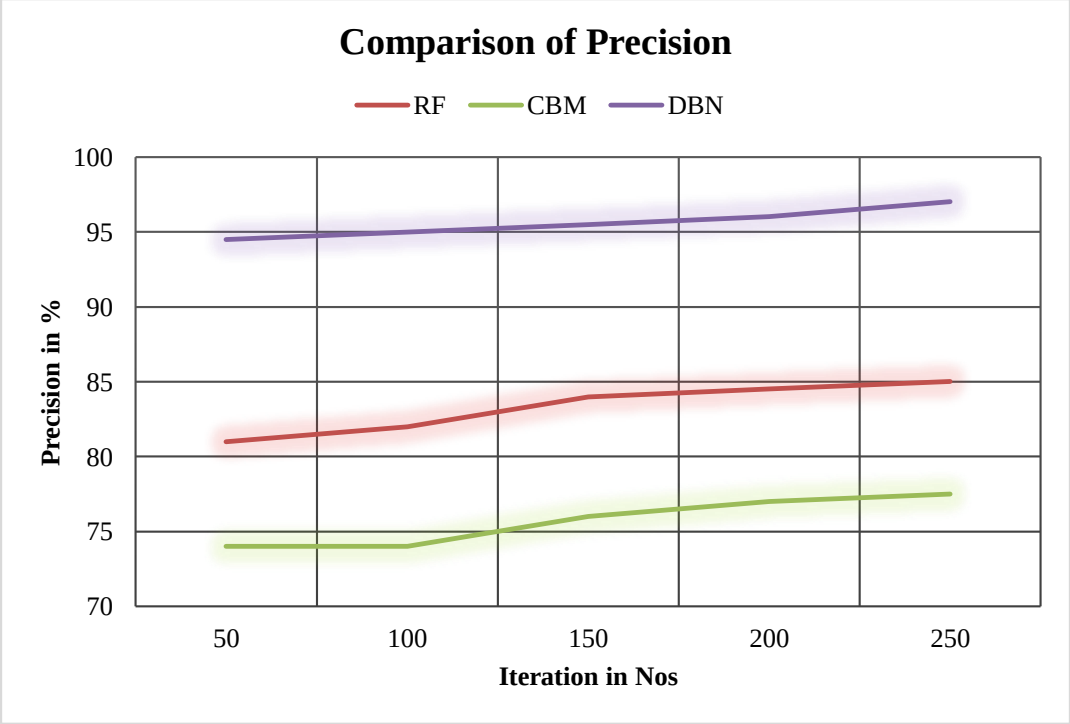

Figure 3. Comparison of URL Classification Precision

Recall

The associated Fake URL among the substantially retrieved occurrences make up the rate of recall. Comparison of recall is given in Table 3 and Figure 4. It is calculated as

$$Recall = \frac{TP}{TP + FN}$$

Table 3
Comparison of URL Classification Recall

S.No	Iteration	RF	CBM	DBN
1	50	83	76	89
2	100	83.5	76.5	90
3	150	84	77	92
4	200	85	78	93
5	250	86	80	96

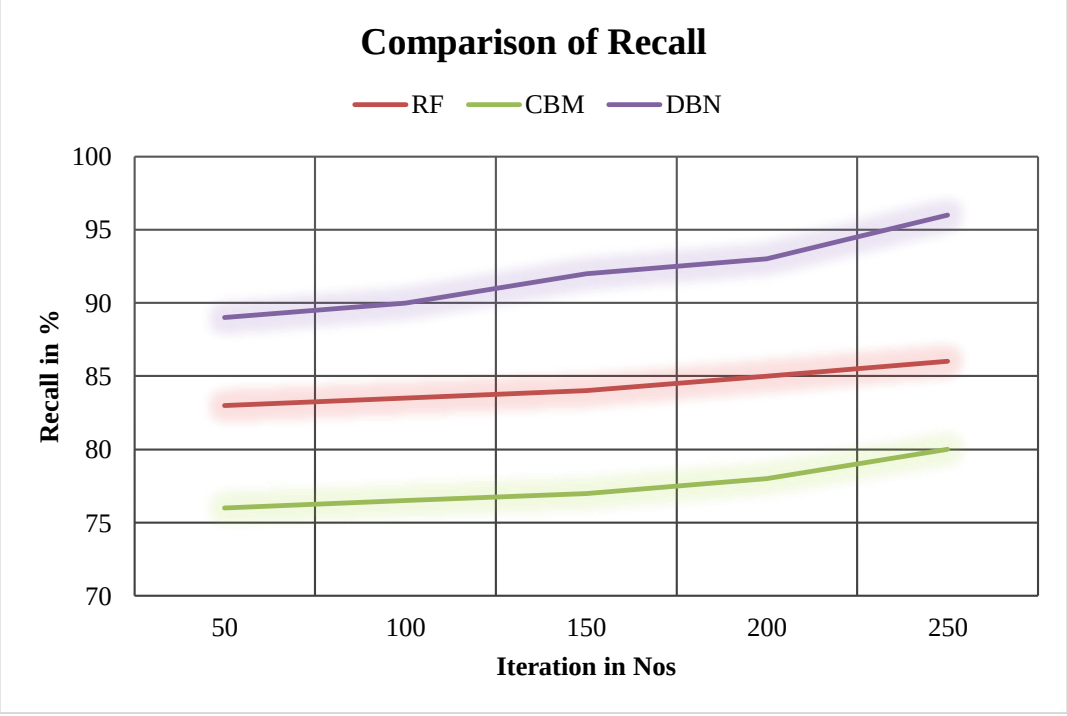

Figure 4. Comparison of URL Classification Recall

From the observation of the above tables and figures, it is identified that the proposed approach is highly effective where the performance of DBN outperforms the existing approaches namely RF as well as CBM.

Conclusion

Attackers create and offer a variety of tools to launch web application assaults and boost the number of harmful attacks. Classifiers download the page, extract the characteristics, and execute classification analysis, whereas dynamic classifiers continually monitor the web page's behavior. This article discusses about the identification of malicious user with the URL and its relevant

information. The process of feature selection from the URL is attained using PCA and the acquired features are utilized for classification of type of user. The classification is attained by the technique DBN and the classification of normal as well as benign user is accomplished with 98.5% accuracy. The feature selection technique PCAwith DBN classification outperforms other existing approaches.

Reference

1. Dixit, P., &Silakari, S. (2021). Deep learning algorithms for cybersecurity applications: A technological and status review. *Computer Science Review*, *39*, 100317.
2. Ulven, J. B., &Wangen, G. (2021). A systematic review of cybersecurity risks in higher education. *Future Internet*, *13*(2), 39.
3. Kumar, E. Boopathi, and V. Thiagarasu. "Segmentation using Fuzzy Membership Functions: An Approach." *IJCSE, ISSN* (2017): 2347-2693.
4. Kumar, E. Boopathi, and V. Thiagarasu. "Comparison and Evaluation of Edge Detection using Fuzzy Membership Functions." *International Journal on Future Revolution in Computer Science & Communication Engineering (IJFRCSCE), ISSN*: 2454-4248.
5. Fosch-Villaronga, E., & Mahler, T. (2021). Cybersecurity, safety and robots: Strengthening the link between cybersecurity and safety in the context of care robots. *Computer Law & Security Review*, *41*, 105528.
6. Lee, I. (2021). Cybersecurity: Risk management framework and investment cost analysis. *Business Horizons*, *64*(5), 659-671.
7. Marinos, N. (2021). *Cybersecurity: Federal Actions Urgently Needed to Better Protect the Nation's Critical Infrastructure* (No. GAO-22-105530).
8. Sadiq, A., Anwar, M., Butt, R. A., Masud, F., Shahzad, M. K., Naseem, S., &Younas, M. (2021). A review of phishing attacks and countermeasures for internet of things-based smart business applications in industry 4.0. *Human Behavior and Emerging Technologies*, *3*(5), 854-864.
9. Butnaru, A., Mylonas, A., &Pitropakis, N. (2021). Towards lightweight url-based phishing detection. *Future internet*, *13*(6), 154.
10. Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. *Computer Communications*, *175*, 47-57.
11. Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. *Computer Communications*, *175*, 47-57.
12. Chen, D., Wawrzynski, P., &Lv, Z. (2021). Cyber security in smart cities: a review of deep learning-based applications and case studies. *Sustainable Cities and Society*, *66*, 102655.
13. Rosenberg, I., Shabtai, A., Elovici, Y., &Rokach, L. (2021). Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, *54*(5), 1-36.
14. Raja, A. S., Vinodini, R., &Kavitha, A. (2021). Lexical features based malicious URL detection using machine learning techniques. *Materials Today: Proceedings*, *47*, 163-166.

15. Biswas, B., Mukhopadhyay, A., Bhattacharjee, S., Kumar, A., &Delen, D. (2022). A text-mining based cyber-risk assessment and mitigation framework for critical analysis of online hacker forums. *Decision Support Systems*, *152*, 113651.
16. Barraclough, P. A., Fehringer, G., & Woodward, J. (2021). Intelligent cyber-phishing detection for online. *Computers & Security*, *104*, 102123.
17. RAMESH, K., BENNET, M. A., VEERAPPAN, J., & RENJITH, P. (2021, April). Performance Metric System for Malicious URL Data using Revised Random Forest Algorithm. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1188-1191). IEEE.
18. Bhagyashree, E., &Tanuja, K. (2015). Phishing URL detection: a machine learning and web mining-based approach. *International Journal of Computer Applications*, *123*(13), 46-50.
19. Jeeva, S. C., &Rajsingh, E. B. (2016). Intelligent phishing url detection using association rule mining. *Human-centric Computing and Information Sciences*, *6*(1), 1-19.
20. Le, A., Markopoulou, A., &Faloutsos, M. (2011, April). Phishdef: Url names say it all. In *2011 Proceedings IEEE INFOCOM* (pp. 191-195). IEEE.
21. Yu, F. (2015). Malicious url detection algorithm based on bm pattern matching. *International Journal of Security and Its Applications*, *9*(9), 33-44.
22. Shrivastava, V., Damodaran, S. S., &Kamble, M. (2020). Adalward: a deep-learning framework for multi-class malicious webpage detection. *Journal of Cyber Security Technology*, *4*(3), 153-195.
23. Yookesh, T. L., et al. "Efficiency of iterative filtering method for solving Volterra fuzzy integral equations with a delay and material investigation." *Materials today: Proceedings* 47 (2021): 6101-6104.
24. E. B. Kumar and V. Thiagarasu, "Color channel extraction in RGB images for segmentation," *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, 2017, pp. 234-239, doi: 10.1109/CESYS.2017.8321272.