A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment

Brij B. Gupta [a,b,*], Krishna Yadav [a], Imran Razzak [c], Konstantinos Psannis [d], Arcangelo Castiglione [e], Xiaojun Chang [f]

[a] National Institute of Technology Kurukshetra, Kurukshetra, 136119, Haryana, India
[b] Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan
[c] Deakin University, Australia
[d] University of Macedonia, Greece
[e] University of Salerno, Fisciano, Salerno, Italy
[f] Monash University Clayton Campus, Australia

ARTICLE INFO

Keywords:
Accuracy
Blacklist
Features
Machine learning
Phishing
Real-time
Random forest
URLs

ABSTRACT

In recent times, we can see a massive increase in the number of devices that are being connected to the internet. These devices include but are not limited to smartphones, IoT, and cloud networks. In comparison to other possible cyber-attacks, these days, hackers are targeting these devices with phishing attacks since it exploits human vulnerabilities rather than system vulnerabilities. In a phishing attack, an online user is deceived by a seemingly trusted entity to give their personal data, i.e., login credentials or credit card details. When this private information is leaked to the hackers, this information becomes the source of other sophisticated attacks. In recent times many researchers have proposed the machine learning-based approach to solve phishing attacks; however, they have used a large number of features to develop reliable phishing detection techniques. A large number of features requires large processing powers to detect phishing, which makes it very much unsuitable for resource constrained devices. To address this issue, we have developed a phishing detection approach that only needs nine lexical features for effectively detecting phishing attacks. We used ISCXURL-2016 dataset for our experimental purpose, where 11964 instances of legitimate and phishing URLs are used. We have tested our approach against different machine learning classifiers and have obtained the highest accuracy of 99.57% with the Random forest algorithm.

1. Introduction

There has been an exponential growth in the number of organizations, and new technologies are continually being explored for broader applicability. Every day millions of new websites are being developed that have login portals to extract credentials from users. As these websites are in large volume, it is becoming challenging to verify their credibility. According to a report published by Dofo [1], more than 5.16 million domain names were registered in the month of April 2020. As the number of users is increasing, it becomes easy for attackers to lure more users. In most cases, phishing attacks start with fraudulent emails that appear to come from a legitimate source. This email consists of malicious links that are redirected to a fake website when clicked by a user. Afterward, the user ends up giving their confidential information like login id, passwords, and credit card details. Sometimes, attackers perform phishing attacks to disseminate malware in the network.

The word phishing was first introduced in the 1990s via America Online. The hacker back then constructed an algorithm that was used to generate the credit card numbers. They use these generated credit card numbers to register an AOL account. When there was a match between the generated and real credit card numbers, they used to create an account whose motive was to ultimately spam other people in AOL's community. Later on, when online users were informed about this scam, hackers switched their phishing platform from messenger to email as it was easy to create an exploitable email, and it was very difficult to catch hackers through emails. Modern days phishing is not only limited to email and websites, but also with the links that appear in online ads, status updates, tweets, and Facebook posts. Some of these links are fraudulent where credentials are being massively stolen.

Fig. 1 represents a traditional/generic phishing scenario (i.e., mass-email phishing campaigns). Fig. 1 shows that an attacker hosts a

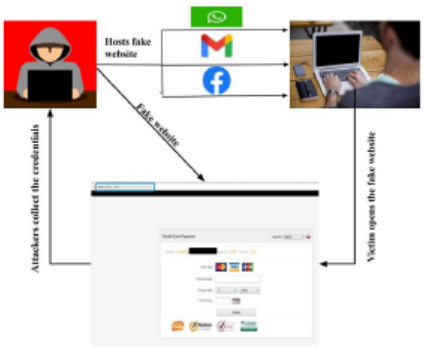

Fig. 1. Traditional Phishing attack.

fake website and sends it to the victim via emails or any other messaging platform. The fake website hosted seems very appealing and authentic to the users. Sometimes hackers also display some lucrative messages inside the hosted fake website, which makes users enter their credentials. The hosted website is controlled by the attackers, and all the entered credentials are redirected to the attackers. Sometimes phishing attacks are used by users to install the malware in the victim's machine. This malware will change the victims machine to the botnet [2,3]. These botnets are used by the attackers to launch DDoS and several other attacks (i.e. Phishing, SQL injection, XSS attacks, authentication/authorization issues, etc.) [4–8].

In 2019, a security threat report of Symantec [9] suggested that one out of 170 URLs in 2018 is malicious and used for performing phishing attacks. As per the report [9], phishing attacks have caused a loss of $1.4 billion in 2018 and $26 billion from June 2016 to July 2019. Check Point security report [10] suggested that 64% of organizations have experienced a phishing attack in 2018. Moreover, IBM also reported [11] that the loss from phishing attacks is not only limited to the revenue generated by the organization but also leads to losing customers, brand, and reputation loss. The report also shows that organizations have lost about 1% of their customers due to a data breach. Phishing attacks were the root cause of these data breaches. Authors at [12] have identified that most of the users are unable to identify the correct websites and select the fake website based on their content and seemingly professional look. In [13], authors believed that people were propelled in sharing their credentials, assuming that email senders already had enough information about them.

Several software have been developed using blacklisting and heuristic approaches to prevent phishing attacks, such as Google Safe Browsing, McAfee SiteAdvisor, Netcraft Anti-phishing Toolbar, Spoof Guard [14]; still, victims are being phished, and credentials are being stolen. With the increase in the novelty of phishing attacks, the blacklisting approach is no more appropriate for phishing detection. Minor changes or mismatch in URLs from the URLs in the blacklist database, i.e., replacing Top-Level Domain, directory path, brand name, Query String substitution, etc., in URLs, can make malicious links undetectable. In [15], authors believed that earlier proposed phishing detection approaches such as heuristic and visual similarity-based techniques time, and intervention of third parties information. In this paper, we have proposed an anti-phishing solution that can detect phishing URLs in a real-time environment without requiring any third party information and also with a very low response time. In our approach, we focused on achieving high accuracy with a limited number of features to detect phishing attacks. Thus, we studied the most important features in the literature and came up with nine lexical-based features to develop a highly accurate phishing detection approach. We evaluated our approach with several machine learning classifiers and obtained the highest accuracy of 99.57%. To provide readers a good insight about our features, we applied different algorithms, such as Spearman correlation, K best, and Random forest, and calculated feature importance scores.

The rest of the paper is organized as follows: The related work is discussed in Section 2, and the proposed approach is discussed in Section 3. Section 4 provides details about the implementation and analysis of results. Finally, Section 5 concludes the paper and discusses some future work.

2. Related work

In this section, we discuss various significant works and existing approaches for phishing attack detection that have been proposed in the literature. Phishing attack detection approaches can be classified into two categories; user education-based and software-based. Software-based approaches can be further categorized into four categories; the blacklisting method, visual similarity method, machine learning method, and hybrid method. Before getting a detailed insight into the existing works proposed in the literature that detect phishing, it is very important to understand some terminology of an URL. Section 2.1 clearly discusses the URL terms, and the rest of the subsection discusses the existing phishing detection approaches.

2.1. Anatomy of an URL

The Uniform Resource Locators are used to identify a webpage that is represented in Fig. 2. It has seven different parts, i.e., Protocol, Domain name, Path, Parameter, Subdomain, Top-level domain, and Query. A protocol defines how our web browser should communicate with a web server. Some of the very common protocols are HTTPS (Hypertext Transfer Protocol Secure), FTP (File Transfer Protocol), POP (Post Office Protocol), SMTP (Simple Mail Transfer Protocol), and IMAP (Internet Message Access Protocol). Further, a domain name is a unique reference that identifies a web site on the internet. The path refers to a unique location where a file or directory exists in a web server, e.g., /home/address/image.jpeg. A subdomain is a subdivision of the main domain name. For example, mail.stanford.edu and cs.iitb.ac.in, are subdomains of stanford.edu and iitb.ac.in. A domain name always includes the TLD(top-level domain); in the case of stanford.edu, edu is a top-level domain. A query is generally found in dynamic web pages. A query is always followed by a question mark. When a client makes a request for a page on a server, it takes a query string and runs the program. For example, https://example.com/over/path/there?name=jason. In this URL, name=jason is a query.

2.2. User education

At the present time, many of the users that are connected to the internet are unaware of Internet security and cyber-threats. These days hackers take advantage of users' limited knowledge to launch different

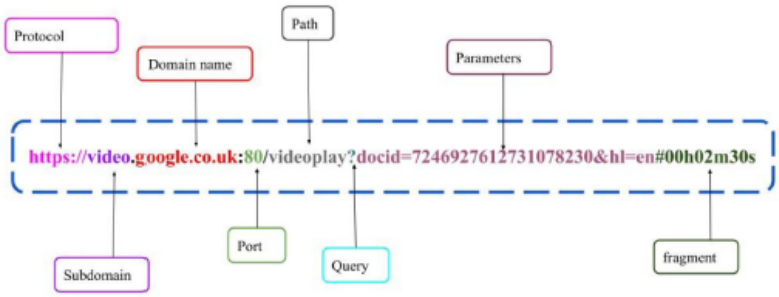

Fig. 2. Anatomy of an URL.

2.3. *Blacklisting phishing URLs*

A classical technique for detecting phishing URLs is to blacklist phishing or malicious URIs. In this approach, a database is maintained containing a large number of phishing URLs, domains, and IP addresses. When a user visits a new URL, the visited URL is searched in the database. If the visited URL is found, then the URL is classified as malicious, otherwise not. For the proper detection of phishing URLs, the blacklist database needs to be updated frequently. Authors at [20] have suggested that the blacklist database needs to be updated every 12 h of the initial phishing test, and in [21], authors have observed that many attackers make minor modifications to URL such as domain name, file path, Query String which lets phishing URL go undetected by the employed blacklist-based system. Authors at [22] suggested that blacklist features alone do not perform well; however, when it is conjugated with other features such as lexical and host-based, it may produce a great result. Due to the simplicity of the blacklist method, it is still widely used in many systems to detect phishing attacks.

2.4. *Visual similarity-based techniques*

Most of the users clicked on fake websites just by looking at the appearance of the URL since it looks like an authentic one. People do not pay much attention to the URL and SSL (Secure Socket Layer) certificates of websites. Visual similarity-based phishing technique utilizes features like HTML tags, CSS, image logo to find the similarity. If the similarity between suspicious websites and legitimate websites exceeds a certain threshold, then the suspicious website is categorized as phishing. The information extraction process from images has been clearly described in [23]. On the basis of extracted information, similarity can be calculated. Authors at [24] have proposed a visual similarity-based phishing detection scheme using images and CSS with a target website finder and have obtained an accuracy of 80%. Authors at [25] proposed a WhiteNet approach where a database of whitelisted pages is maintained based on their URLs. The embeddings of the phishing web pages are compared to the whitelist pages and the decision is made based on visual similarity. Phishing webpages were found to be visually similar to the whitelist. The disadvantage of a visual similarity-based scheme is that it cannot detect newly launched phishing websites.

Fig. 3(a). Legitimate Facebook webpage.

Fig. 3(b). Phishing webpage of Facebook.

2.5. *Machine learning based techniques*

In recent times, many authors have proposed machine learning approaches for phishing detection, which includes a vast database of phishing and legitimate websites. The features related to the URL, page content, DNS, etc., are extracted, and the new dataset is made

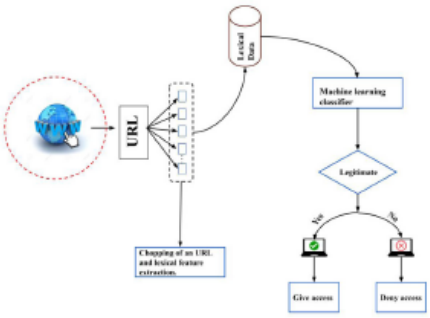

Fig. 4. Proposed Phishing detection architecture.

different machine learning algorithms where accuracy of 97.98% has been obtained with random forest. Authors at [26] have used URL and content-based features and have achieved an accuracy of 96.58%. However, their proposed approach has one disadvantage of high false-positive rates. Authors at [27] have presented a phishing detection approach where they extracted 19 features at the client-side to distinguish malicious links from legitimate ones. They have used 2141 phishing pages and 1918 legitimate web pages to prepare a dataset and have obtained an accuracy of 99.39%. Their proposed approach is available for client-side desktop applications and can be used for real-time phishing detection. In literature, many approaches have also used the natural language processing method for phishing detection. In [28], authors have utilized the NLP method to detect phishing emails and obtained an accuracy of 95%. They have used a blacklisted bag of words with a dataset that contains 5009 and 5000 instances of phishing and legitimate emails.

2.6. Hybrid features based technique

The hybrid approach uses the combination of several features such as URL-based, Content-based, and Domain-based features for phishing detection. In [29], authors have suggested that the combination of several features may boost up the accuracy rate of phishing URLs detection. They have discussed a machine learning-based hybrid approach and image checking approaches for the detection of phishing websites [29]. They have used hyperlink based features, third-party based features, and URL obfuscation features and have obtained an accuracy of 99.55%.

3. Proposed approach

3.1. Design objectives

The main objective of this research is to develop a machine learning-based phishing detection system that can help users to check the legitimacy and maliciousness of an URL within a minimum amount of time. We aim to develop a mechanism that can extract the feature vectors from URL whenever a user visits that URL. Afterward, the feature vectors are pre-processed and fed into several machine learning algorithms to validate the legitimacy of an URL. The other objectives

- *Low response time*: We aim at building an anti-phishing approach that can detect phishing URLs as early as possible. The low response time will give hackers comparatively less time to steal the credentials.
- *Detection of new phishing websites*: We aim at building an anti-phishing system that outperforms the current blacklisting techniques for phishing detection and can detect new phishing websites in the coming future.
- *Scalability: We aim at developing a phishing detection approach that can be embedded in a device with constrained resources, i.e., IoT, to the devices powered up by multiple CPUs and GPUs.*

3.2. Architecture of proposed approach

The architecture of our proposed system is shown in Fig. 4. In Fig. 4, we can see that whenever a user visits a new URL, the URL is chopped into different segments. We developed a feature extraction mechanism that obtains different lexical based information from URL. Our developed feature extraction mechanism is briefly discussed in Section 3.3. Our feature extraction mechanism consists of multiple scripts written in python. The lexical information obtained is then mapped into respective features, and the instance of an URL is prepared that contains all the information necessary to classify the URL as legitimate or phishing. The data from an instance of an URL is then preprocessed, and the preprocessed data is fetched into machine learning algorithms to check the legitimacy of an URL. If the website is found legitimate, then it is given access to the clients to use, otherwise, the website is blocked.

3.3. Features extraction algorithm

In this section, we clearly discuss our lexical features and the algorithms we have used to extract lexical data from the URLs. The most contributing lexical features were developed by analyzing several available lexical features proposed by different researchers [30,31]. We applied several algorithms to calculate feature importance and came up with the optimal number of features. In our features, we have expanded the number of top-level domains and have included new domains that could present in phishing URLs. Moreover, we have also expanded the number of delimiters in a delimiters list. Expanding the list has increased the feature importance that can be seen in Fig. 8 and ultimately accuracy. The lexical data extracted were then used to make a dataset to train machine learning classifiers.

1. ***No.of token in a domain***: Tokenizing is the process of chopping the URL into several pieces. A token refers to a segment that has been broken down into sequence. In this feature, we break down URLs into the tokens and take their count. Generally, it is seen that phishing URLs are longer and contain a greater number of a token count. For e.g., URL: http://clubeamigosdopedrosegundo.com.br/last/
   Tokenized URL=['http', 'clubeamigosdopedrosegundo', 'com', 'br', 'last']. Algorithm 1 gives an idea about the way of tokenizing URLs and taking their count. In algorithm 1, tokenize function helps in creating a token which is stored in a list of

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research work is being supported by sponsored project grant YFRF, under the project Visvesvaraya Ph.D. Scheme of Ministry of Electronics & Information Technology, Government of India and being implemented by Digital India Corporation.

References

[1] Domain registered report available at: https://dofo.com/blog/domain-industry-report-april-2020/ Last accessed on May 11, 2020.
[2] Amrita Dahiya, Brij B. Gupta, A reputation score policy and Bayesian game theory based incentivized mechanism for DDoS attacks mitigation and cyber defense, Future Gener. Comput. Syst. 117 (2021) 193–204.
[3] A. Al-Nawasrah, A.A. Almomani, S. Atawneh, M. Alauthman, A survey of fast flux botnet detection with fast flux cloud computing, Int. J. Cloud Appl. Comput. (IJCAC) 10 (3) (2020) 17–53.
[4] C. Esposito, M. Ficco, et al., Blockchain-based authentication and authorization for smart city applications, Inf. Process. Manage. 58 (2) (2021) 102468.
[5] S. Kaushik, C. Gandhi, Ensure hierarchal identity based data security in cloud environment, Int. J. Cloud Appl. Comput. (IJCAC) 9 (4) (2019) 21–36.
[6] Q. Zheng, X. Wang, M.K. Khan, W. Zhang, et al., A lightweight authenticated encryption scheme based on chaotic scml for railway cloud service, IEEE Access 6 (2017) 711–722.
[7] O.O. Olakanmi, A. Dada, An efficient privacy-preserving approach for secure verifiable outsourced computing on untrusted platforms, Int. J. Cloud Appl. Comput. (IJCAC) 9 (2) (2019) 79–98.
[8] C.L. Stergiou, K.E. Psannis, et al., IoT-based big data secure management in the fog over a 6G wireless network, IEEE Internet Things J. (2020).
[9] The Security threat report of Symantec is available at https://docs.broadcom.com/doc/istr-24-2019-en Last accessed on May 11, 2020.
[10] Check Point security report available at: https://www.phishingbox.com/assets/files/images/Check-Point-Research-Information-Security-Report-2018.pdf. Last accessed on May 11, 2020.
[11] The phishing loss report produced by IBM is available at: https://www.ibm.com/security/data-breach. Last accessed on May 11, 2020.
[12] R. Dhamija, J.D. Tygar, M. Hearst, Why phishing works, in: Proceedings of ACM Conference on Human Factors in Computing Systems (CHI2006), 2006, 581-59.
[13] J.S. Downs, M.B. Holbrook, L. Cranor, Decision strategies and susceptibility to phishing, in: Proceedings of the Second Symposium on Usable Privacy and Security (SOUPS 2006), 2006, pp. 79–90.
[14] Information about existing anti-phishing software is available at https://en.wikipedia.org/wiki/Anti-phishing_software Last accessed on May 15, 2020.
[15] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, An empirical analysis of phishing blacklists, 2009, https://kilthub.cmu.edu/articles/An_Empirical_Analysis_of_Phishing_Blacklists/6469805.
[16] O.K. Sahingoz, E. Buber, O. Demir, B. Diri, Machine learning base phishing detection from URLs, Expert Syst. Appl. 117 (2019) 345-357.
[17] A.K. Jain, B.B. Gupta, Towards detection of phishing websites on client-side using machine learning based approach, in: Telecommunication Systems, Springer, 2018.
[18] M. Moghimi, A.Y. Varjani, New rule-based phishing detection method, Expert Syst. Appl. 53 (2016) 231–242.
[19] S. Afroz, R. Greenstadt, Phishzoo: Detecting phishing websites by looking at them, in: Fifth IEEE International Conference on Semantic Computing, 2011.
[20] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, An empirical analysis of phishing blacklists, 2009, https://kilthub.cmu.edu/articles/An_Empirical_Analysis_of_Phishing_Blacklists/6469805.
[21] P. Prakash, M. Kumar, R.R. Kompella, Phishnet: predictive blacklisting to detect phishing attacks, in: Mini-Conference at IEEE INFOCOM, 2010.
[22] J. Ma, L.K. Saul, S. Savage, G.M. Voelker, Beyond blacklists: learning to detect malicious web sites from suspicious URLs, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.
[25] S. Abdelnabi, K. Krombholz, M. Fritz, WhiteNet Phishing website detection by visual whitelists, in: Cryptography and Security, 2019, (arXiv).
[26] V. Patil, P. Thakkar, C. Shah, T. Bhat, Detection and prevention of phishing websites using machine learning approach, in: Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.
[27] A.K. Jain, B.B. Gupta, Towards detection of phishing websites on client-side using machine learning based approach, in: Telecommunication Systems, Springer, 2018.
[28] T. Peng, I. Harris, Y. Sawa, Detecting phishing attacks using natural language processing and machine learning, in: IEEE 12th International Conference on Semantic Computing(ICSC), 2018.
[29] R.S. Rao, A.R. Pais, Detection of phishing websites using an efficient feature-based machine learning framework, in: Neural Computing and Applications, Vol. 31, Springer, 2019.
[30] Kholoud Althobaiti, Ghaidaa Rummani, Kami Vaniea, A review of human-and computer-facing url phishing features, in: 2019 IEEE European Symposium on Security and Privacy Workshops (EuroS & PW), IEEE, 2019.
[31] Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi, Malicious URL detection using machine learning: A survey, 2017, arXiv preprint arXiv:1701.07179.
[32] M. Maalouf, Logistic regression in data analysis: an overview, Int. J. Data Anal. Tech. Strateg. (2011).
[33] C. Crisci, B. Ghattas, G. Perera, A review of supervised machine learning algorithms and their applications to ecological data, Ecol. Model. 240 (2012) 113-122.
[34] C. Crisci, B. Ghattas, G. Perera, A review of supervised machine learning algorithms and their applications to ecological data, Ecol. Model. 240 (2012) 113-122.
[35] G. Biau, E. Scornet, A random forest guided tour, 2016, pp. 197–227.
[36] C. Croux, C. Dehon, Influence functions of the Spearman and Kendall correlation measures, in: Statical Method and Applications, 2010, pp. 497–515.
[37] D.S. Modha, W.S. Spangler, Feature weighting in k-means clustering, in: Machine Learning, Springer, 2003, pp. 217–237.
[38] B.H. Menze, B.M. Kelm, R. masuch, U. Himmerreich, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, in: BMC Bioinformatics, Springer, 2009, 213.
[39] Legitimate and phishing URLs we have used is available here: https://www.unb.ca/cic/datasets/url-2016.html. Last accessed on May 10, 2020.
[40] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowl. Inf. Syst. (2012) 1–33.
[41] Standard scaling theory available at : https://en.wikipedia.org/wiki/Feature_scaling. Last accessed on July 20, 2020.
[42] I.U. Haq, I. Gondal, P. Vamplew, S. Brown, Categorical features transformation with compact one-hot encoder for fraud detection in distributed environment, in: Australasian Conference on Data Mining, 2018, pp. 69-80.
[43] N. Abdelhamid, A. Ayesh, F. Thabtah, Phishing detection based associative classification data mining, Expert Syst. Appl. 41 (13) (2014) 5948-5959.
[44] K.L. Chiew, E.H. Chang, W.K. Tiong, Utilisation of website logo for phishing detection, Comput. Secur. 54 (2015) 16–26.
[45] G. Xiang, J. Hong, C.P. Rose, L. Cranor, Cantina+ a feature-rich machine learning framework for detection of phishing web sites, ACM Trans. Inf. Syst. Secur. (2011) 21.
[46] A.K. Jain, B.B. Gupta, Towards detection of phishing websites on client-side using machine learning based approach, in: Telecommunication Systems, Springer, 2018.
[47] S. Gupta, B.B. Gupta, PHP-sensor: a prototype method to discover workflow violation and XSS vulnerabilities in PHP web applications, in: Proceedings of the 12th ACM International Conference on Computing Frontiers, 2015, pp. 1–8.
[48] Yazan Ahmad Alsariera, et al., Ai meta-learners and extra-trees algorithm for the detection of phishing websites, IEEE Access 8 (2020) 142532-142542.
[49] Ammara Zamir, et al., Phishing Web Site Detection using Diverse Machine Learning Algorithms, The Electronic Library, 2020.
[50] Nureni Ayofe Azeez, et al., Identifying phishing attacks in communication networks using URL consistency features, Int. J. Electron. Secur. Digit. Forensics 12 (2) (2020) 200-213.
[51] Ankit Kumar Jain, Brij B. Gupta, A machine learning based approach for phishing detection using hyperlinks information, J. Ambient Intell. Humaniz. Comput. 10 (5) (2019) 2015–2028.
[52] A. Tewari, B.B. Gupta, Security, privacy and trust of different layers in Internet-of-Things (IoTs) framework, Future Gener. Comput. (2020).