A Comprehensive Systematic Review of Neural Networks and Their Impact on the Detection of Malicious Websites in Network Users

https://doi.org/10.3991/ijim.v17i01.36371

Javier Gamboa-Cruzado[1(✉)], Juan Briceño-Ochoa[1], Marco Huaysara-Ancco[1], Alberto Alva-Arévalo[2], Caleb Ríos-Vargas[2], Magaly Arangüena Yllanes[3], Liset S. Rodriguez-Baca[1]
[1] Universidad Autónoma del Perú, Lima, Perú
[2] Universidad Nacional de San Martín, Tarapoto, Perú
[3] Universidad Nacional José María Arguedas, Apurímac, Perú
jgamboa65@hotmail.com

Abstract—The large branches of Machine Learning represent an immense support for the detection of malicious websites, they can predict whether a URL is malicious or benign, leaving aside the cyber attacks that can generate for network users who are unaware of them. The objective of the research was to know the state of the art about Neural Networks and their impact for the Detection of malicious Websites in network users. For this purpose, a systematic literature review (SLR) was conducted from 2017 to 2021. The search identified 561 963 papers from different sources such as Taylor & Francis Online, IEEE Xplore, ARDI, ScienceDirect, Wiley Online Library, ACM Digital Library and Microsoft Academic. Of the papers only 82 were considered based on exclusion criteria formulated by the author. As a result of the SLR, studies focused on machine learning (ML), where it recommends the use of algorithms to have a better and efficient prediction of malicious websites. For the researchers, this review presents a mapping of the findings on the most used machine learning techniques for malicious website detection, which are essential for a study because they increase the accuracy of an algorithm. It also shows the main machine learning methodologies that are used in the research papers.

Keywords—machine learning, neural network, web site detection, malicious web sites, algorithms, systematic literature review

1 Introduction

Nowadays, websites offer services of all kinds to network users, such as e-mail, social networks, online shopping, among others. These websites store users' confidential information. Fraudsters always try to steal users' confidential information by using misleading URL text [41]. Researchers have been applying different types of algorithms for example Sequential Minimum Optimization (SMO), logistic regression and naive

bayes, decision tree, K nearest neighbors, among others. These algorithms achieve reliable and accurate results when detecting malicious websites [10]. In order to evaluate the performance of the models and algorithms, some experiments are carried out using a set of comparative data: accuracy and area under the receiver operating characteristic (ROC), receiver operating characteristic (ROC) and area under the curve (AUC) [41]. In this work, we identified the algorithms most commonly used by researchers from different studies with the aim of detecting malicious websites.

In this SLR we obtained papers related to the topic, but not focused on the prediction of malicious websites with a comparative approach using neural networks. However, the papers reviewed propose algorithms for the prediction of these websites thus giving their effectiveness for this detection process.

In the general study for the prediction process Sahingoz, Buber, Demir and Diri [45] propose to use two lists, Whitelist (whitelist) and Blacklist (blacklist), to classify legitimate and malicious websites. Whitelist-based website detection systems create safe and legitimate websites to provide the necessary information. Every website that is not on the whitelist is considered malicious. In the study of the prediction process there are a number of algorithms that give different results for different varieties of malicious websites.

Authors Gandotra and Gupta [15] use machine learning algorithms SVM, Random Forest (RF), Neural Network, Logistic Regression and Naïve Bayes (NB); to differentiate suspicious websites from benign ones.

Authors P. Yang, Zhao and Z. Yang [47] argue that deep learning (Deep Learning) is a research direction of neural networks that can discover hidden information within complex data through level-by-level learning. CNN is an artificial deep feedback deep neural network. Compared with traditional back propagation neural networks, CNNs adopt a weight sharing network structure similar to that of a biological neural network, and its neurons are sparsely connected, which reduces the complexity of the network model and improves the training performance.

According to Haider and Singh [1], phishing is a deception technique that aims to steal sensitive personal information such as passwords, credit cards, identity theft and other fraudulent activities by an individual or a group. Intruders can take this information using phishing techniques (e.g., when a user enters their data on a phishing website, their data is stolen and they are then redirected to the original site).

Al-Milli and Hammo [43] claim that convolutional neural network (CNN) is one of the most successful methods used recently in classification problems. CNN is used for complex classification problems. More specifically, CNN is used in the image processing domain.

The uniqueness of this research is the use of the Mendeley tool useful for the management of the papers, as well as the use of artificial intelligence for the generation of the bigrams, trigrams and bibliometric networks that show relevant information, as well as a comparison between the keywords in the reviewed papers.

The main objective is to determine the current state of the art of worldwide experimental research on Neural Networks and their influence on the Detection of Malicious Websites in network users. The structure of the paper is organized as follows; section II presents an interpretation of previous research and what the study aims to achieve.

Section III details the methodology to be used for the systematic literature review and this was developed according to Kitchenham and Charters [83]. Section IV shows the results of each question and also a comparison with a review paper. Section V finally gives conclusions and recommendations.

2 Review methodology

2.1 Review protocol

This research followed the model and steps proposed by Kitchenham and Charters [83]; it covered the following: research questions, sources of information, identified studies, exclusion criteria, quality assessment, data extraction, and synthesis of findings (Figure 1).

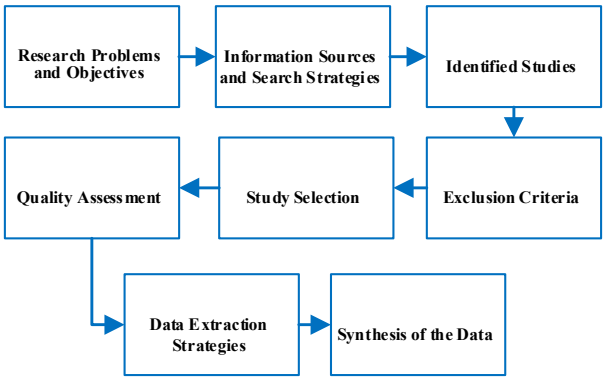

Fig. 1. RSL process

2.2 Research problems and objectives

For the SLR, research questions (RQs) were developed that are necessary for the data search, extraction and analysis strategy. Each research question has an objective, which are shown in Table 1.

Table 1. Research questions and objectives

Research Question	Objetive
RQ1: What types of algorithms are being considered to detect malicious websites in web users?	Identify the most considered types of algorithms for detecting malicious websites in network users
RQ2: What machine learning methodologies are used to detect malicious websites?	Report the machine learning methodologies that are used for malicious website detection
RQ3: What are the criteria for measuring the overall effectiveness of neural networks with machine learning?	Determine what are the criteria for measuring the overall effectiveness of neural networks with machine learning

RQ4: What are the most commonly used and relevant keywords about neural networks and their influence on the process of detecting malicious websites in network users?	Determine which are the most used and relevant keywords, on neural networks and their influence on the process of detecting malicious websites in network users.
RQ5: What are the keywords that show co-occurrence in neural network research and their influence on the process of malicious website detection in network users?	Detect which keywords present co-occurrence in neural network research and their influence on the process of detecting malicious websites in network users

2.3 Information sources and search strategies

Figure 2 shows the sources used in the search for research papers.

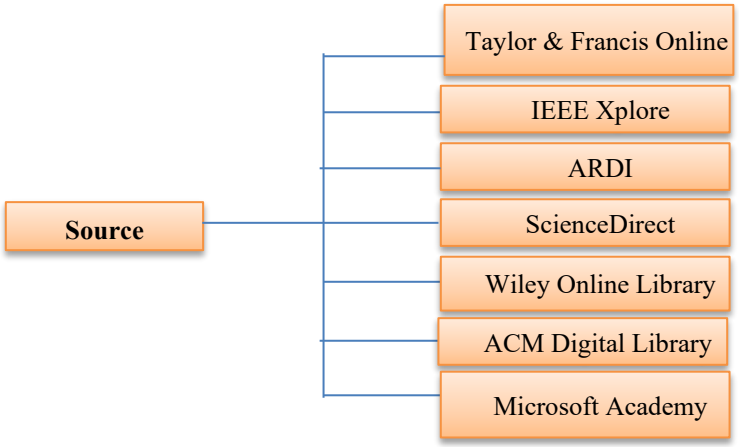

Fig. 2. Sources of information

For the search strategy of the studies, the thematic descriptors and their synonyms were identified (See Table 2).

Table 2. Search descriptors

Descriptor	Variable
neural network / machine learning	Independent
website / URL detection	Dependent

For the search of the papers, equations were used for each source of information, as shown in Table 3.

Table 3. Information sources and search equations

Source	Search equation
Taylor & Francis Online	[[All: "neural network"] OR [All: "machine learning"]] AND [[All: "website detection"] OR [All: URL*]]
IEEE Xplore	("Neural network" OR "Machine Learning") AND ("website detection" OR URL*)
ARDI	("Neural network" OR "Machine Learning") AND (" website detection" OR URL*)
Science Direct	("neural network" OR "machine learning") AND (website detection OR URL OR URLS)
Wiley Online Library	""neural network" OR "machine learning"" anywhere and ""website detection" OR URL*" anywhere
ACM Digital Library	[[All: "neural network"] OR [All: "machine learning"]] AND [[All: website detection] OR [All: URL*]
Microsoft Academic	("Neural network" OR "Machine Learning") AND ("website detection" OR "URL*")

2.4 Identified studies

The papers identified in each source are shown in Figure 3.

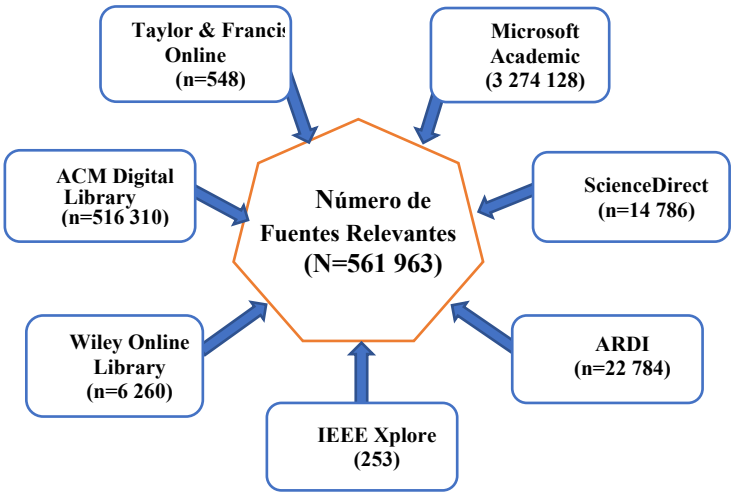

Fig. 3. Number of relevant studies

2.5 Exclusion criteria

Al Some exclusion criteria (EC) were necessary for the filtering and selection of the papers:

CE1.The papers are older than 5 years old
CE2. The papers are not written in English
CE3. The papers are not published in Conferences or Journals
CE4. The papers are repeated
CE5. The titles and keywords of the papers are not very appropriate
CE6. There is not enough information to make the assessment
CE7. The abstracts of the papers are not very relevant

2.6 Studio selection

Initially, 561963 papers were obtained, based on the search performed using the keywords relevant to the study.

Then, a series of selection and filtering steps were applied, as shown in Figure 4.

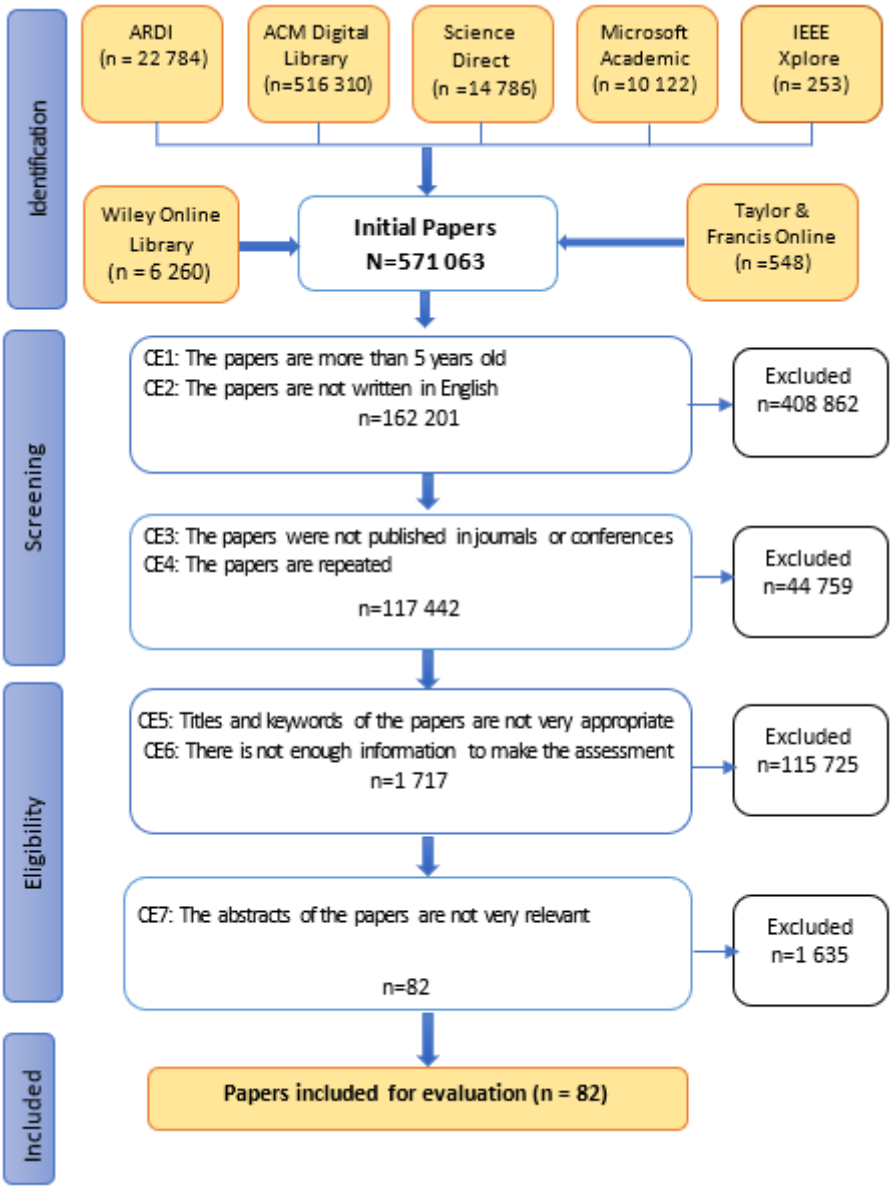

Fig. 4. PRISMA flow chart

2.7 Quality assessment

An important phase in rigorously evaluating the final list of papers was the quality assessment criteria (QAs). These were applied to measure the quality of the research papers. Seven QAs were identified:

QA1. Is the purpose of the research clearly explained?
QA2. Are the research findings clearly explained?
QA3. Is the paper well organized?
QA4. Does the document include practical experiments?
QA5. Are data collection and measurements adequately described?
QA6. Are the results of the experiments performed clearly identified and reported?
QA7. Is the document considered useful?

In general, the use of these 7 QAs ensures that these findings could make a valuable contribution to the review. Each of the 7QAs was rated on a dichotomous scale. Therefore, the research described in the papers is understandable and their results can be relied upon. Of the studies evaluated for quality assurance, the 82 papers have been retained.

2.8 Strategies of data extraction

For the final data extraction, the list of total papers was integrated in order to answer the research questions formulated.

The information extracted from each paper included the following: ID of the paper, title of the paper, URL, source, year, country, number of pages, language, publication type, publication name, authors, affiliation, number of citations, abstract, keywords, sample size. The Mendeley desktop tool was used to perform the data extraction.

2.9 Synthesis of the data

The data collected for research questions RQ1-RQ5 were tabulated and presented as quantitative data that were used to develop a statistical comparison between the various solutions for each question, taking into account research conducted in the last 5 years.

3 Results and discussion

3.1 General description of studies

Eighty-two papers were selected for the review, which were evaluated for their quality and the data extraction and analysis strategy was applied. Figure 5 shows the number of papers published per year. This indicates that the year in which the most papers were published was 2020.

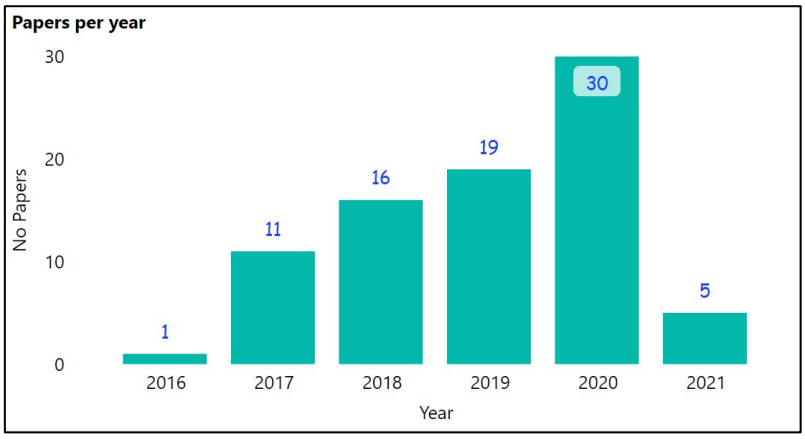

Fig. 5. Papers by year

Figure 6 details the number of papers by country, this shows that the country with the highest number of papers for the research was India as well as China. It can be seen that India and China are the countries that have a greater focus on the detection of malicious websites with machine learning by the number of papers published and the countries that have less participation are Jordan and Slovenia.

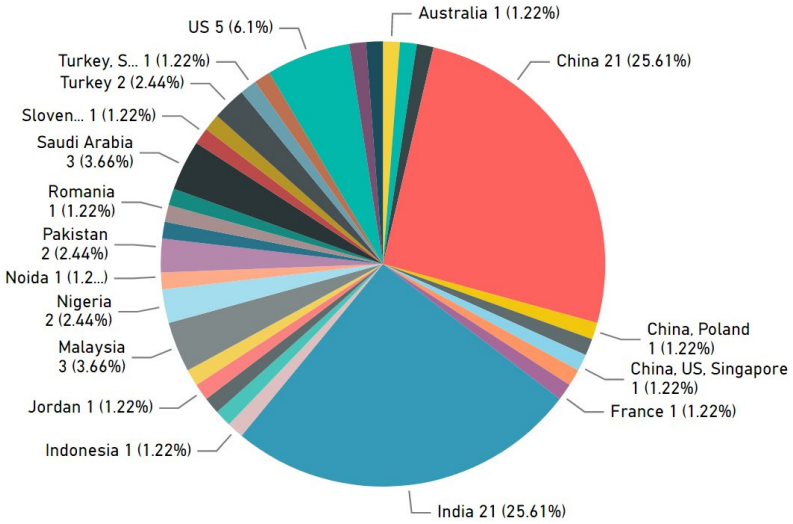

Fig. 6. Number of papers by country

Figure 7 shows the papers by type of publication. As shown in the figure, a total of 82 papers were found, of which 40% were published in Conferences and the remaining 60% were published in Journals.

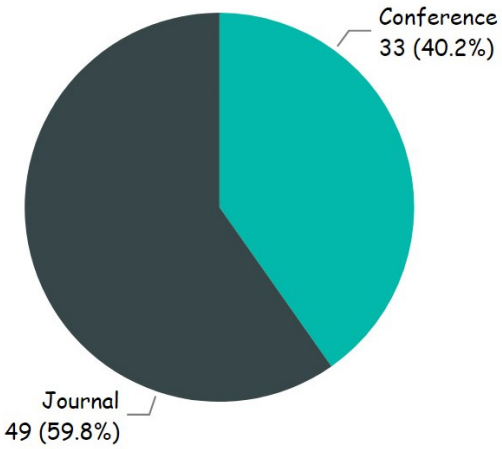

Fig. 7. Papers by publication type

Figure 8 shows the most common words in the abstracts of the papers reviewed. It can be seen that the most used word was phishing, this was even the most common word in the keywords. This is followed by the words learning and detection.

Fig. 8. Most recurrent words in the abstract

3.2 Answers to research questions

The following are the answers (results) for each of the RQs formulated, which were obtained after an exhaustive review of the 82 papers selected.

RQ1: What types of algorithms are being applied to detect malicious websites in network users? The results shown in Table 4 indicate which were the most used machine learning algorithms in the area of malicious website detection in network users. In this research 10 algorithms were considered, it is observed that the one with the highest percentage of references are neural networks with 23%, it can be said that it is the most popular in this area of malicious website detection, the second most applied type of algorithm was Super Vectore Machine with 19%.

Table 4. Machine learning algorithms

ML algorithm	Reference	Qty. (%)
Neural network	[1][3][4][5][7][8][9][16][18][19][20][23][26][21] [22][28][34][44] [54] [57] [64][69][82]	24 (23)
K-nearest neighbors	[4][8][11][12][15][25][26][21][30] [35]	10 (10)
Decision tree	[4][10][11][14][25][30][35] [44] [54] [65]	10 (10)
Logistic regression	[7][10][11][21][30][35][47] [65] [77]	10 (10)
Random forest	[4][8][10][11][15][25][30] 54][65] [69]	10 (10)
Naive Bayes	[7][8][14][15][16][26][30]35][44][46][47][54][57] [62]	14 (14)
Super Vector Machine	[4][7][8][11][12][14][15][16][25][21][28][35][44] [46] [47][54] [57] [69] [77]	19 (19)
Extreme machine learning (ELM)	[21][79]	2 (2)
Ripper algorithm	[60]	1 (1)
Cultural Algorithm	[1]	1 (1)

According to authors Badawi and Jourdan [89], in their research question they refer what mechanisms are available to detect cybercriminal activities; it answers that the models that offer the best results are random forest with an accuracy of 99%, another algorithm model that also stands out in their answer is SVM with an accuracy of 97.9%.

According to the authors Gheewala and Patel [86], neural networks are a widely used data mining algorithm for detecting malicious websites and it in turn depends on the proper selection of its features, such as model performance and the level of prediction accuracy of the algorithm.

RQ2: What machine learning methodologies are used to detect malicious websites? To answer the question the following machine learning methodologies in neural networks were considered: convolutional neural network and artificial neural network as the most considered in the research papers, this is because these methodologies offer better results in the area of malicious website detection, on the other hand convolutional

neural network was mentioned in many papers that have more than 5 citations (See Table 5).

Table 5. Machine learning methodologies

ML Methodologie	Reference	Qty. (%)
Convolutional Neural Network (CNN)	[9][7][23][26][28][35][43][47][5][54][64][69][77][82]	14 (34)
Deep Neural Network	[10][23][41][46][47] [57][64]	7 (17)
Artificial Neural Network	[1][3][4][11][16][18][19][20][21][22][34][44][49][54][57]	15 (36)
Recurrent Neural Network	[5][7][9][28][54][82]	5 (13)

From the point of view of authors Odeh, Keshta and Abdelfattah [85], they argue that convolutional neural networks and long short-term memory (LSTM), are widely used techniques for website phishing detection, but when both are used in a study both CNN and LSTM get better results. A clear example of this is that CNN learns from URL features and sends them to LSTM for final resolution. This approach motivates other researchers to design a model using a combination of deep learning models.

RQ3: What are the criteria for measuring the overall effectiveness of neural networks with machine learning? To answer this question, the criteria for measuring the effectiveness of Machine Learning were taken into account. The most commonly used criteria are the ROC curve and AUC to determine the accuracy in the tests performed statistically based on the machine learning algorithms proposed in the papers studied (See Table 6).

Table 6. Criteria for measuring ML effectiveness

Criteria	Reference	Qty. (%)
ROC CURVE	[5][7][10][19][30][57] [69][82][2][3][9][36] [43]	12 (43)
AUC	[4][7][10][19][28][82] [2][3][36][41][43][51]	12 (43)
PCA	[8][14][16] [9]	4 (14)

According to Dou, Khalil, Khreishah, Al-Fuqaha and Guizani [88], ROC curves can be used as comparisons between two or more models or techniques of algorithms thus giving greater efficiency in terms of the result, it is also possible to interpret that in the combination of these algorithms a result greater than 90% is achieved taking into account the characteristics given in the combination.

RQ4: What are the most frequently used and relevant keywords about neural networks and their influence on the process of detecting malicious websites in network users? Figure 9 shows the keywords with more frequency in the papers, in first place is "machine learning", this word is repeated in 25 papers, in addition, this word was key when searching the papers and it is also the independent variable in Table 3. In second place is the keyword "phishing" with 17 repetitions, a very common word when searching for papers related to the detection of malicious websites in network users.

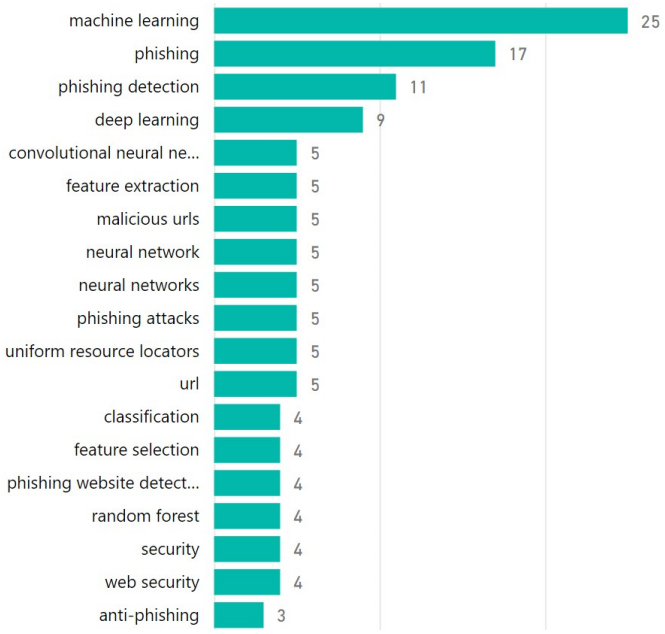

Fig. 9. Keyword repetitions

The authors Mondal, Maheshwari, Pai, and Biwalkar [87], identify "machine Learning" and "pishing" as very common keywords when searching for papers on malicious website detection.

One can also display the most repeated words as a word cloud in Figure 10.

Fig. 10. Cloud of most repeated keywords

RQ5: What are the keywords that show co-occurrence in neural network research and their influence on the process of malicious website detection in network users? In order to answer this RQ, a biometric network of keywords as they appear together in the reviewed research has been elaborated (see Figure 11).

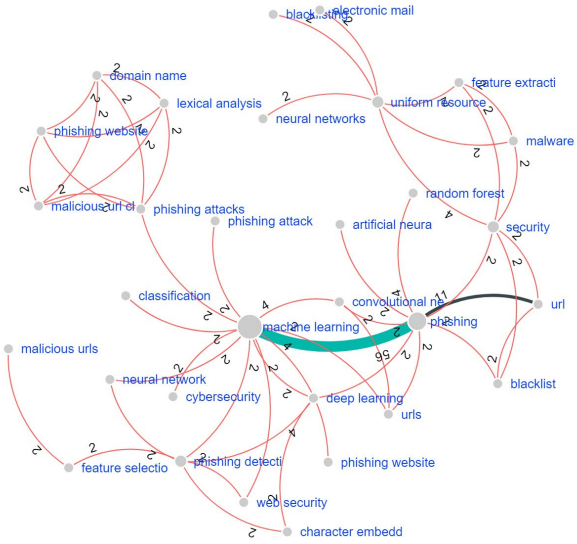

Fig. 11. Keyword bibliometric network

In the bibliometric network it is possible to visualize the keywords most used by the authors together, as well as it can also be visualized that "phishing" and "machine learning" are together in 56 reviewed papers, also "phishing" and "url" are together in 11 papers. Then it is deduced that the words have a relationship, the first one which is phishing and machine learning its relationship is problem-solution because nowadays machine learning method is used for prediction of malicious websites (phishing). The second relationship of phishing and url, its relationship is of similarity because phishing is a technique of deception, in this case it would be through a misleading url impersonating someone close or a particular company in order to steal information.

4 Conclusion

This study has used the SRL methodology proposed by Kitchenham and Charters [83], whose purpose was to respond to the problems posed, so high quality papers that address the RQs formulated have been evaluated. As a result, 82 research papers have been identified, which answer the research questions. These papers were carefully selected by applying exclusion criteria and had a quality assessment. It should be noted that the Mendeley tool was used, which was very helpful for the management of the papers analyzed. In the results section and its discussions, the 5 RQs formulated have been answered, using statistical graphs, tables and the novel bibliometric networks. These were very helpful in answering the research questions and a comparison has been made with many review papers.

Therefore, future research should consider reviewing more recent publications on the process of detecting malicious websites. This will benefit to optimize the inquiry on the topic of malicious website detection process in order to be able to have a wider scope and depth on the topics of cybersecurity.

5 Acknowledgment

We would like to thank the Universidad Autónoma del Perú for its support to our work carried out, specifically the career of de Systems Engineering.

6 References

[1] A. Haider and R. Singh, “Phishing URL Detection using Neural Network Optimized by Cultural Algorithm,” Int. J. Comput. Sci. Eng., vol. 6, no. 7, pp. 860–863, 2018, https://doi.org/10.26438/ijcse/v6i7.860863

[2] A. A. Orunsolu, A. S. Sodiya, and A. T. Akinwale, “A predictive model for phishing detection,” J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2020, https://doi.org/10.26438/ijcse/v6i7.860863

[3] A. Basit, M. Zafar, A. R. Javed, and Z. Jalil, “A Novel Ensemble Machine Learning Method to Detect Phishing Attack,” Proc. - 2020 23rd IEEE Int. Multi-Topic Conf. INMIC 2020, no. November, 2020, https://doi.org/10.1109/INMIC50486.2020.9318210

[4] A. Subasi, E. Molah, F. Almkallawi, and T. J. Chaudhery, "Intelligent phishing website detection using random forest classifier," 2017 Int. Conf. Electr. Comput. Technol. Appl. ICECTA 2017, vol. 2018-Janua, pp. 1–5, 2017, https://doi.org/10.1109/INMIC50486.2020.9318210

[5] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. Gonzalez, "Classifying Phishing URLs Using Recurrent Neural Networks," eCrime Res. Summit, eCrime, pp. 1–8, 2017. https://doi.org/10.1109/ECRIME.2017.7945048

[6] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J. P. Niyigena, "An effective phishing detection model based on character level convolutional neural network from URL," Electron., vol. 9, no. 9, pp. 1–24, 2020, https://doi.org/10.3390/electronics9091514

[7] A. A. Ubing, S. K. B. Jasmi, A. Abdullah, N. Z. Jhanjhi, and M. Supramaniam, "Phishing website detection: An improved accuracy through feature selection and ensemble learning," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 252–257, 2019, https://doi.org/10.14569/IJACSA.2019.0100133

[8] A. Zamir et al., "Phishing web site detection using diverse machine learning algorithms," Electron. Libr., vol. 38, no. 1, pp. 65–80, 2020, https://doi.org/10.1108/EL-05-2019-0118

[9] A. Vazhayil, R. Vinayakumar, and K. Soman, "Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks," 2018 9th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2018, pp. 1–6, 2018, https://doi.org/10.1109/ICCCNT.2018.8494159

[10] S. Anwar et al., "Countering Malicious URLs in Internet of Things Using a Knowledge-Based Approach and a Simulated Expert," IEEE Internet Things J., vol. 7, no. 5, pp. 4497–4504, 2020, https://doi.org/10.1109/JIOT.2019.2954919

[11] A. Ghimire, A. K. Jha, S. Thapa, S. Mishra, and A. M. Jha, "Machine learning approach based on hybrid features for detection of phishing URLs," Proc. Conflu. 2021 11th Int. Conf. Cloud Comput. Data Sci. Eng., no. January, pp. 954–959, 2021, https://doi.org/10.1109/Confluence51648.2021.9377113

[12] Y. C. Chen, Y. W. Ma, and J. L. Chen, "Intelligent Malicious URL Detection with Feature Analysis," Proc. - IEEE Symp. Comput. Commun., vol. 2020-July, 2020, https://doi.org/10.1109/ISCC50000.2020.9219637

[13] C. L. Tan, K. L. Chiew, N. Musa, and D. H. Abang Ibrahim, "Identifying the Most Effective Feature Category in Machine Learning-based Phishing Website Detection," Int. J. Eng. Technol., vol. 7, no. 4.31, pp. 1–6, 2018.

[14] A. Crisan, G. Florea, L. Halasz, C. Lemnaru, and C. Oprisa, "Detecting Malicious URLs Based on Machine Learning Algorithms and Word Embeddings," Proc. - 2020 IEEE 16th Int. Conf. Intell. Comput. Commun. Process. ICCP 2020, pp. 187–193, 2020, https://doi.org/10.1109/ICCP51029.2020.9266139

[15] E. Gandotra and D. Gupta, "Improving Spoofed Website Detection Using Machine Learning," Cybern. Syst., vol. 52, no. 2, pp. 169–190, 2021, https://doi.org/10.1080/01969722.2020.1826659

[16] E. Zhu, D. Liu, C. Ye, F. Liu, X. Li, and H. Sun, "Effective phishing website detection based on improved BP neural network and dual feature evaluation," Proc. - 16th IEEE Int. Symp. Parallel Distrib. Process. with Appl. 17th IEEE Int. Conf. Ubiquitous Comput. Commun. 8th IEEE Int. Conf. Big Data Cloud Comput. 11t, pp. 759–765, 2019. https://doi.org/10.1109/BDCloud.2018.00114

[17] E. Zhu, C. Ye, D. Liu, F. Liu, F. Wang, and X. Li, "An effective neural network phishing detection model based on optimal feature selection," Proc. - 16th IEEE Int. Symp. Parallel Distrib. Process. with Appl. 17th IEEE Int. Conf. Ubiquitous Comput. Commun. 8th IEEE

Int. Conf. Big Data Cloud Comput. 11t, pp. 781–787, 2019. https://doi.org/10.1109/BDCloud.2018.00117
[18] E. Zhu, Y. Ju, Z. Chen, F. Liu, and X. Fang, "DTOF-ANN: An Artificial Neural Network phishing detection model based on Decision Tree and Optimal Features," Appl. Soft Comput. J., vol. 95, p. 106505, 2020, https://doi.org/10.1016/j.asoc.2020.106505
[19] E. Zhu, Y. Chen, C. Ye, X. Li, and F. Liu, "OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network," IEEE Access, vol. 7, pp. 73271–73284, 2019, https://doi.org/10.1109/ACCESS.2019.2920655
[20] F. P. Motlagh and A. K. Bardsiri, "Detecting Fake Websites Using Swarm Intelligence Mechanism in Human Learning," Int. J. Eng., vol. 31, no. 10, pp. 1642–1650, 2018, https://doi.org/10.5829/ije.2018.31.10a.05
[21] F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han, and J. Q. Wang, "The application of a novel neural network in the detection of phishing websites," J. Ambient Intell. Humaniz. Comput., vol. 0, no. 0, pp. 1–15, 2018, https://doi.org/10.1007/s12652-018-0786-3
[22] S. Gayathri, "Phishing websites classifier using polynomial neural networks in genetic algorithm," 2017 4th Int. Conf. Signal Process. Commun. Networking, ICSCN 2017, pp. 16–19, 2017, https://doi.org/10.1109/ICSCN.2017.8085736
[23] G. Vrbančič, I. Fister, and V. Podgorelec, "Swarm Intelligence Approaches for Parameter Setting of Deep Learning Neural Network," pp. 1–8, 2018, https://doi.org/10.1145/3227609.3227655
[24] S. Gupta and S. Sachdeva, "Invitation or Bait? Detecting Malicious URLs in Facebook Events," 2018 11th Int. Conf. Contemp. Comput. IC3 2018, pp. 1–6, 2018, https://doi.org/10.1109/IC3.2018.8530525
[25] G. Harinahalli Lokesh and G. BoreGowda, "Phishing website detection based on effective machine learning approach," J. Cyber Secur. Technol., vol. 5, no. 1, pp. 1–14, 2021, https://doi.org/10.1109/IC3.2018.8530525
[26] H. huan Wang, L. Yu, S. wei Tian, Y. fang Peng, and X. jun Pei, "Bidirectional LSTM Malicious webpages detection algorithm based on convolutional neural network and independent recurrent neural network," Appl. Intell., vol. 49, no. 8, pp. 3016–3026, 2019, https://doi.org/10.1007/s10489-019-01433-4
[27] H. Yuan, Z. Yang, X. Chen, Y. Li, and W. Liu, "URL2Vec: URL modeling with character embeddings for fast and accurate phishing website detection," Proc. - 16th IEEE Int. Symp. Parallel Distrib. Process. with Appl. 17th IEEE Int. Conf. Ubiquitous Comput. Commun. 8th IEEE Int. Conf. Big Data Cloud Comput. 11t, pp. 265–272, 2019. https://doi.org/10.1109/BDCloud.2018.00050
[28] I. Yilmaz, A. Siraj, and D. Ulybyshev, "Improving DGA-Based malicious domain classifiers for malware defense with adversarial machine learning," 4th IEEE Conf. Inf. Commun. Technol. CICT 2020, no. January, 2020, https://doi.org/10.1109/CICT51604.2020.9311925
[29] J. Yuan, G. Chen, S. Tian, and X. Pei, "Malicious URL detection based on a parallel neural joint model," IEEE Access, vol. 9, pp. 9464–9472, 2021, https://doi.org/10.1109/ACCESS.2021.3049625
[30] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, and B. S. Bindhumadhava, "Phishing website classification and detection using machine learning," 2020 Int. Conf. Comput. Commun. Informatics, ICCCI 2020, pp. 20–25, 2020, https://doi.org/10.1109/ACCESS.2021.3049625
[31] N. Sanglerdsinlapachai and A. Rungsawang, "Using domain top-page similarity feature in machine learning-based web phishing detection," 3rd Int. Conf. Knowl. Discov. Data Mining, WKDD 2010, pp. 187–190, 2010, https://doi.org/10.1109/WKDD.2010.108

[32] K. M. Zubair Hasan, M. Z. Hasan, and N. Zahan, "Automated Prediction of Phishing Websites Using Deep Convolutional Neural Network," 5th Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2019, pp. 1–4, 2019, https://doi.org/10.1109/IC4ME247184.2019.9036647

[33] Y. Liang and X. Yan, "Using deep learning to detect malicious URLs," Proc. - IEEE Int. Conf. Energy Internet, ICEI 2019, pp. 487–492, 2019, https://doi.org/10.1109/ICEI.2019.00092

[34] M. E. Pratiwi, T. A. Lorosae, and F. W. Wibowo, "Phishing Site Detection Analysis Using Artificial Neural Network," J. Phys. Conf. Ser., vol. 1140, no. 1, 2018, https://doi.org/10.1088/1742-6596/1140/1/012048

[35] A. R. Abbas, S. Singh, and M. Kau, "Detection of phishing website using machine learning Approach," Lect. Notes Networks Syst., vol. 89, pp. 1307–1314, 2020, doi: 10.1007/978-981-15-0146-3_128. https://doi.org/10.1088/1742-6596/1140/1/012048

[36] M. Maktabar, A. Zainal, M. A. Maarof, and M. N. Kassim, "Content based fraudulent website detection using supervised machine learning techniques," Adv. Intell. Syst. Comput., vol. 734, pp. 294–304, 2018, https://doi.org/10.1007/978-3-319-76351-4_30

[37] A. S. Manjeri, R. Kaushik, A. Mnv, and P. C. Nair, "A Machine Learning Approach for Detecting Malicious Websites using URL Features," Proc. 3rd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2019, pp. 555–561, 2019, https://doi.org/10.1109/ICECA.2019.8821879

[38] M. Sameen, K. Han, and S. O. Hwang, "PhishHaven - An Efficient Real-Time AI Phishing URLs Detection System," IEEE Access, vol. 8, pp. 83425–83443, 2020, https://doi.org/10.1109/ACCESS.2020.2991403

[39] M. Korkmaz, O. K. Sahingoz, and B. DIri, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis," 2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020, 2020, https://doi.org/10.1109/ICCCNT49239.2020.9225561

[40] M. V. Kunju, E. Dainel, H. C. Anthony, and S. Bhelwa, "Evaluation of phishing techniques based on machine learning," 2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019, no. Iciccs, pp. 963–968, 2019, https://doi.org/10.1109/ICCS45141.2019.9065639

[41] M. Al-Janabi, E. De Quincey, and P. Andras, "Using supervised machine learning algorithms to detect suspicious URLs in online social networks," Proc. 2017 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2017, pp. 1104–1111, 2017, https://doi.org/10.1145/3110025.3116201

[42] M. Aydin, I. Butun, K. Bicakci, and N. Baykal, "Using Attribute-based Feature Selection Approaches and Machine Learning Algorithms for Detecting Fraudulent Website URLs," 2020 10th Annu. Comput. Commun. Work. Conf. CCWC 2020, no. Dm, pp. 774–779, 2020, https://doi.org/10.1109/CCWC47524.2020.9031125

[43] N. Al-Milli and B. H. Hammo, "A Convolutional Neural Network Model to Detect Illegitimate URLs," 2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020, pp. 220–225, 2020, https://doi.org/10.1109/ICICS49469.2020.239536

[44] N. Hosseini, F. Fakhar, B. Kiani, and S. Eslami, "Enhancing the security of patients' portals and websites by detecting malicious web crawlers using machine learning techniques," Int. J. Med. Inform., vol. 132, no. March, 2019, https://doi.org/10.1016/j.eswa.2018.09.029

[45] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Expert Syst. Appl., vol. 117, pp. 345–357, 2019, https://doi.org/10.1016/j.eswa.2018.09.029

[46] P. S. Ezekiel, O. E. Taylor, and F. B. D.- Okuchaba, "A Model to Detect Phishing Websites using Support Vector Classifier and a Deep Neural Network Algorithm," Ijarcce, vol. 9, no. 6, pp. 188–194, 2020, https://doi.org/10.17148/IJARCCE.2020.9632

[47] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," IEEE Access, vol. 7, no. c, pp. 15196–15209, 2019, https://doi.org/10.1109/ACCESS.2019.2892066

[48] A. K. Singh and N. Goyal, "A Comparison of Machine Learning Attributes for Detecting Malicious Websites," in 2019 11th International Conference on Communication Systems and Networks, COMSNETS 2019, 2019, vol. 2061, pp. 352–358, https://doi.org/10.1109/COMSNETS.2019.8711133

[49] P. Saravanan and S. Subramanian, "A Framework for Detecting Phishing Websites using GA based Feature Selection and ARTMAP based Website Classification," Procedia Comput. Sci., vol. 171, no. 2019, pp. 1083–1092, 2020, https://doi.org/10.1109/COMSNETS.2019.8711133

[50] R. Rajalakshmi, H. Tiwari, J. Patel, A. Kumar, and R. Karthik, "Design of Kids-specific URL Classifier using Recurrent Convolutional Neural Network," Procedia Comput. Sci., vol. 167, no. 2019, pp. 2124–2131, 2020, https://doi.org/10.1016/j.procs.2020.03.260

[51] R. Haidar and S. Elbassuoni, "Website navigation behavior analysis for bot detection," Proc. - 2017 Int. Conf. Data Sci. Adv. Anal. DSAA 2017, vol. 2018-Janua, pp. 60–68, 2017, https://doi.org/10.1016/j.procs.2020.03.260

[52] R. Kumar, X. Zhang, H. A. Tariq, and R. U. Khan, "Malicious URL detection using multi-layer filtering model," 2016 13th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. ICCWAMTIP 2017, vol. 2018-Febru, pp. 97–100, 2017, https://doi.org/10.1016/j.procs.2020.03.260

[53] R. Verma and A. Das, "What's in a URL: Fast feature extraction and malicious URL detection," IWSPA 2017 - Proc. 3rd ACM Int. Work. Secur. Priv. Anal. co-located with CODASPY 2017, pp. 55–63, 2017, https://doi.org/10.1016/j.procs.2020.03.260

[54] S. Shivangi, P. Debnath, K. Saieevan, and D. Annapurna, "Chrome Extension for Malicious URLs detection in Social Media Applications Using Artificial Neural Networks and Long Short Term Memory Networks," 2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018, pp. 1993–1997, 2018, https://doi.org/10.1016/j.procs.2020.03.260

[55] S. Nayak, D. Nadig, and B. Ramamurthy, "Analyzing Malicious URLs using a Threat Intelligence System," Int. Symp. Adv. Networks Telecommun. Syst. ANTS, vol. 2019-Decem, pp. 6–9, 2019, https://doi.org/10.1109/ANTS47819.2019.9118051

[56] S. Wen, Z. Zhao, and H. Yan, "Detecting malicious websites in depth through analyzing topics and web-pages," ACM Int. Conf. Proceeding Ser., pp. 128–133, 2018, https://doi.org/10.1145/3199478.3199500

[57] S. G. Selvaganapathy, M. Nivaashini, and H. P. Natarajan, "Deep belief network based detection and categorization of malicious URLs," Inf. Secur. J., vol. 27, no. 3, pp. 145–161, 2018, https://doi.org/10.1080/19393555.2018.1456577

[58] S. Singhal, U. Chawla, and R. Shorey, "Machine Learning Concept Drift based Approach for Malicious Website Detection," 2020 Int. Conf. Commun. Syst. NETworkS, COMSNETS 2020, pp. 582–585, 2020, https://doi.org/10.1109/COMSNETS48256.2020.9027485

[59] S. Sindhu, S. P. Patil, A. Sreevalsan, F. Rahman, and A. N. Saritha, "Phishing detection using random forest, SVM and neural network with backpropagation," Proc. Int. Conf. Smart Technol. Comput. Electr. Electron. ICSTCEE 2020, pp. 391–394, 2020, https://doi.org/10.1109/COMSNETS48256.2020.9027485

[60] S. Thakur, E. Meenakshi, and A. Priya, "Detection of malicious URLs in big data using RIPPER algorithm," RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018-Janua, pp. 1296–1301, 2017, https://doi.org/10.1109/COMSNETS48256.2020.9027485

[61] S. Gupta and A. Singhal, "Phishing URL detection by using artificial neural network with PSO," 2nd Int. Conf. Telecommun. Networks, TEL-NET 2017, vol. 2018-Janua, pp. 1–6, 2018, https://doi.org/10.1109/TEL-NET.2017.8343553
[62] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018-Janua, pp. 300–301, 2018, https://doi.org/10.1109/TEL-NET.2017.8343553
[63] T. Li, G. Kou, and Y. Peng, "Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods," Inf. Syst., vol. 91, p. 101494, 2020, https://doi.org/10.1109/TEL-NET.2017.8343553
[64] T. Shibahara et al., "Malicious URL sequence detection using event de-noising convolutional neural network," IEEE Int. Conf. Commun., 2017, https://doi.org/10.1109/ICC.2017.7996831
[65] V. Patil, P. Thakkar, C. Shah, T. Bhat, and S. P. Godse, "Detection and Prevention of Phishing Websites Using Machine Learning Approach," Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–5, 2018, https://doi.org/10.1109/ICCUBEA.2018.8697412
[66] F. Vanhoenshoven, G. Napoles, R. Falcon, K. Vanhoof, and M. Koppen, "Detecting malicious URLs using machine learning techniques," 2016 IEEE Symp. Ser. Comput. Intell. SSCI 2016, 2017, https://doi.org/10.1109/SSCI.2016.7850079
[67] W. Ali, "Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 9, pp. 72–78, 2017, https://doi.org/10.14569/IJACSA.2017.080910
[68] W. Ali and A. A. Ahmed, "Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting," IET Inf. Secur., vol. 13, no. 6, pp. 659–669, 2019, https://doi.org/10.1049/iet-ifs.2019.0006
[69] W. Wei, Q. Ke, J. Nowak, M. Korytkowski, R. Scherer, and M. Woźniak, "Accurate and fast URL phishing detector: A convolutional neural network approach," Comput. Networks, vol. 178, no. April, 2020, https://doi.org/10.1016/j.comnet.2020.107275
[70] S. Mondal, D. Maheshwari, N. Pai, and A. Biwalkar, "A Review on Detecting Phishing URLs using Clustering Algorithms," 2019 6th IEEE Int. Conf. Adv. Comput. Commun. Control. ICAC3 2019, pp. 1–6, 2019, https://doi.org/10.1109/ICAC347590.2019.9036837
[71] J. Mao et al., "Detecting Phishing Websites via Aggregation Analysis of Page Layouts," Procedia Comput. Sci., vol. 129, pp. 224–230, 2018, https://doi.org/10.1016/j.procs.2018.03.053
[72] W. Bai, "Phishing website detection based on machine learning algorithm," Proc. - 2020 Int. Conf. Comput. Data Sci. CDS 2020, pp. 293–298, 2020, https://doi.org/10.1109/CDS49703.2020.00064
[73] W. Yang, W. Zuo, and B. Cui, "Detecting Malicious URLs via a Keyword-Based Convolutional Gated-Recurrent-Unit Neural Network," IEEE Access, vol. 7, no. c, pp. 29891–29900, 2019, https://doi.org/10.1109/CDS49703.2020.00064
[74] X. Xiao, D. Zhang, G. Hu, Y. Jiang, and S. Xia, "CNN–MHSA: A Convolutional Neural Network and multi-head self-attention combined approach for detecting phishing websites," Neural Networks, vol. 125, pp. 303–312, 2020, https://doi.org/10.1016/j.neunet.2020.02.013
[75] X. Yan, Y. Xu, B. Cui, S. Zhang, T. Guo, and C. Li, "Learning URL Embedding for Malicious Website Detection," IEEE Trans. Ind. Informatics, vol. 16, no. 10, pp. 6673–6681, 2020, https://doi.org/10.1109/TII.2020.2977886

[76] X. D. Hoang, "A website defacement detection method based on machine learning," Lect. Notes Networks Syst., vol. 63, pp. 116–124, 2019, https://doi.org/10.1007/978-3-030-04792-4_17

[77] X. Yu, "Phishing Websites Detection Based on Hybrid Model of Deep Belief Network and Support Vector Machine," IOP Conf. Ser. Earth Environ. Sci., vol. 602, no. 1, pp. 0–9, 2020, https://doi.org/10.1088/1755-1315/602/1/012001

[78] Y. Xue, Y. Li, Y. Yao, X. Zhao, J. Liu, and R. Zhang, "Phishing sites detection based on Url Correlation," Proc. 2016 4th IEEE Int. Conf. Cloud Comput. Intell. Syst. CCIS 2016, pp. 244–248, 2016, https://doi.org/10.1109/CCIS.2016.7790262

[79] Y. Shi, G. Chen, and J. Li, "Malicious Domain Name Detection Based on Extreme Machine Learning," Neural Process. Lett., vol. 48, no. 3, pp. 1347–1357, 2018, https://doi.org/10.1007/s11063-017-9666-7

[80] Y. Huang, J. Qin, and W. Wen, "Phishing URL Detection Via Capsule-Based Neural Network," Proc. Int. Conf. Anti-Counterfeiting, Secur. Identification, ASID, vol. 2019-Octob, no. December 2016, pp. 22–26, 2019, https://doi.org/10.1109/ICASID.2019.8925000

[81] Y. Lian et al., "Detecting Malicious Web Requests Using an Enhanced TextCNN," in Proceedings - 2020 IEEE 44th Annual Computers, Software, and Applications Conference, COMPSAC 2020, 2020, no. 61872011, pp. 768–777, https://doi.org/10.1109/COMPSAC48688.2020.0-167

[82] Z. Chen, Y. Liu, C. Chen, M. Lu, X. Zhang, and School, "Malicious URL Detection Based on Improved Multilayer Recurrent Convolutional Neural Network Model", Security and Communication Networks, vol. 2021, Article ID 9994127, 13 pages, 2021. https://doi.org/10.1155/2021/9994127

[83] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in SE," *Guidel. Perform. Syst. Lit. Rev. SE*, pp. 1–44, 2007, [Online]. Available: https://userpages.uni-koblenz.de/~laemmel/esecourse/slides/slr.pdf

[84] A. Arshad, A. U. Rehman, S. Javaid, T. M. Ali, J. A. Sheikh, and M. Azeem, "A Systematic Literature Review on Phishing and Anti-Phishing Techniques," pp. 163–168, 2021, [Online]. Available: http://arxiv.org/abs/2104.01255

[85] A. Odeh, I. Keshta, and E. Abdelfattah, "Machine LearningTechniquesfor Detection of Website Phishing: A Review for Promises and Challenges," 2021 IEEE 11th Annu. Comput. Commun. Work. Conf. CCWC 2021, no. January, pp. 813–818, 2021, https://doi.org/10.1109/CCWC51732.2021.9375997

[86] S. Gheewala and R. Patel, "Machine Learning Based Twitter Spam Account Detection: A Review," Proc. 2nd Int. Conf. Comput. Methodol. Commun. ICCMC 2018, no. Iccmc, pp. 79–84, 2018, https://doi.org/10.1109/ICCMC.2018.8487992

[87] S. Mondal, D. Maheshwari, N. Pai, and A. Biwalkar, "A Review on Detecting Phishing URLs using Clustering Algorithms," 2019 6th IEEE Int. Conf. Adv. Comput. Commun. Control. ICAC3 2019, pp. 1–6, 2019, https://doi.org/10.1109/ICAC347590.2019.9036837

[88] Z. Dou, I. Khalil, A. Khreishah, A. Al-Fuqaha, and M. Guizani, "Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection," IEEE Commun. Surv. Tutorials, vol. 19, no. 4, pp. 2797–2819, 2017, https://doi.org/10.1109/COMST.2017.2752087

[89] E. Badawi and G. V. Jourdan, "Cryptocurrencies emerging threats and defensive mechanisms: A systematic literature review," IEEE Access, vol. 8, pp. 200021–200037, 2020, https://doi.org/10.1109/ACCESS.2020.3034816

7 Authors

Dr. Javier Gamboa-Cruzado works at the Faculty of Systems Engineering of the Universidad Autónoma del Perú, Lima, Peru. He is Doctor in Systems Engineering and Doctor in Administrative Sciences. He has published several articles in international journals and conferences. His research interests are in machine learning, big data, the internet of things, natural language processing, and business intelligence (email: jgamboa65@hotmail.com).

Juan Briceño-Ochoa is graduate of the Faculty of Engineering and Architecture at the Universidad Autónoma del Perú, Peru. His research interests are in web systems, internet of things, and big data (email: jbricenoo@autonoma.edu.pe).

Marco Huaysara-Ancco is graduate at the Faculty of Engineering and Architecture at the Universidad Autónoma del Perú, with extensive experience in database management and development of mobile applications (email: mhuaysaraa@autonoma.edu.pe).

Dr. Alberto Alva Arévalo is a Professor at the Faculty of Systems Engineering and Informatics of the Universidad Nacional de San Martín, Peru. He is a Systems Engineer, and has a Master's Degree in Sciences with a Mention in Information Technology. His research interests are related to Information Technology and Communications (Email: aalva@unsm.edu.pe).

Dr. Caleb Ríos Vargas is a University Professor, Works Consultant, is a Civil Engineer by profession, graduated from the Universidad Nacional de San Martín - Tarapoto and has a Master of Science with a mention in Transportation Engineering from the National University of Engineering of Lima - Peru, holds a Doctorate in Business Management from the National University of San Martín. His interests are related to research and technological innovation (email: crios@unsm.edu.pe).

Mg. Magaly Arangüena Yllanes is working at the Faculty of Engineering of the José María Arguedas National University, Apurímac, Peru. She has a Master's degree in Magister Scientiae in Computer Science, mention in Management of Information and Communications Technologies. She has published articles in national magazines. Her research interests are Data and Information Analytics and Visualization, Governance and Business Intelligence (email: magalyyllanes@gmail.com).

Dra. Liset S. Rodriguez-Baca, Systems Engineer, graduated in Education, Master in Systems Engineering with Mention in Management and Management in Information Technology, Master in Strategic Business Management, Doctor in Education Sciences. Director of the Professional School of Systems Engineering at the Universidad Autónoma del Perú (email: liset.rodriguez@autonoma.pe).

Article submitted 2022-09-25. Resubmitted 2022-11-13. Final acceptance 2022-11-13. Final version published as submitted by the authors.