Journal of King Saud University – Computer and Information Sciences 35 (2023) 590–611

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

A systematic literature review on phishing website detection techniques

Asadullah Safi [a], Satwinder Singh [b,*]

[a] Nangarhar University, Ministry of Higher Education, Afghanistan
[b] Dept. of Computer Science & Technology, Central University of Punjab, Bathinda, Punjab, India

ARTICLE INFO

Article history:
Received 23 June 2022
Revised 24 November 2022
Accepted 3 January 2023
Available online 11 January 2023

Keywords:
Phishing
Phishing Detection
Deep Learning
Cyber Security
Machine Learning

ABSTRACT

Phishing is a fraud attempt in which an attacker acts as a trusted person or entity to obtain sensitive information from an internet user. In this Systematic Literature Survey (SLR), different phishing detection approaches, namely Lists Based, Visual Similarity, Heuristic, Machine Learning, and Deep Learning based techniques, are studied and compared. For this purpose, several algorithms, data sets, and techniques for phishing website detection are revealed with the proposed research questions. A systematic Literature survey was conducted on 80 scientific papers published in the last five years in research journals, conferences, leading workshops, the thesis of researchers, book chapters, and from high-rank websites. The work carried out in this study is an update in the previous systematic literature surveys with more focus on the latest trends in phishing detection techniques. This study enhances readers' understanding of different types of phishing website detection techniques, the data sets used, and the comparative performance of algorithms used. Machine Learning techniques have been applied the most, i.e., 57 as per studies, according to the SLR. In addition, the survey revealed that while gathering the data sets, researchers primarily accessed two sources: 53 studies accessed the PhishTank website (53 for the phishing data set) and 29 studies used Alexa's website for downloading legitimate data sets. Also, as per the literature survey, most studies used Machine Learning techniques; 31 used Random Forest Classifier. Finally, as per different studies, Convolution Neural Network (CNN) achieved the highest Accuracy, 99.98%, for detecting phishing websites.

Contents

* Corresponding author.
E-mail addresses: asad.nu.it@gmail.com (A. Safi), satwinder.singh@cup.edu.in (S. Singh).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2023.01.004
1319-1578/

1. Introduction

Phishing is a social engineering attack (Paliath et al., 2020; Nakamura and Dobashi, 2019; Zabihimayvan and Doran, 2019) identified as the most common method used by cybercriminals to get access to an internet user's personal information such as credit card information, usernames, and passwords (Ramana et al., 2021; Faris and Yazid, 2021). Sometimes, attackers perform phishing attacks to disseminate malware in the network (Gupta et al., 2021). There are various types of phishing attacks, well-known among them and not limited to spoofing, malware-based phishing, DNS-based phishing, data theft, email/spam, web-based delivery, and phone phishing, as shown in Fig. 1 (Kathrine et al., 2019). Phishing attacks have many forms and usually involve a variety of communication channels, such as email, instant messages, quick response (QR) codes (Geng et al., 2018), and social media. Attackers commonly mimic well-known banks, credit card agencies, or well-known e-commerce websites to intimidate or convince users to log in to the phishing website and provide credentials they may regret later. A user may, for example, receive an instant message indicating an issue with their bank account and be directed to a website that appears identical to the bank's website. The client inserts their credentials without hesitation into the relevant fields, falling into the trap of the attackers. Criminals keep track of this data and exploit it to access the user's legitimate accounts (Liu et al., 2021). According to the Internet Crime Complaint Center (IC3) report in 2020, the FBI received 791,790 complaints of suspected internet crime, an increase of more than 300,000 complaints compared to 2019 data (FBI, "FBI Releases the Internet Crime Complaint Center, 2020). In the literature, various techniques were proposed to identify the phishing website, Lists-Based, Visual Similarity, Heuristic, Machine Learning (Somesha et al., 2020; Nakamura and Dobashi, 2019), and Deep Learning techniques (Basit et al., 2020).

Fig. 1. Types of phishing attacks.

List Based: Browsers like Microsoft Edge, Firefox, and Google Chrome utilize List Based methods to detect phishing websites. Whitelisting and blacklisting are two types of List Based approaches. The whitelist contains a list of valid URLs that browsers can access, which means if the URL is in the whitelist, the browser can download the web page. At the same time, the blacklisting database includes phishing or fraudulent URLs that stop browsers from downloading the web pages. The major disadvantage is that a minor modification in the URL is enough to bypass the List Based techniques and to prevent new phishing URLs, these lists must frequently be updated (Yang et al., 2021).

***Visual Similarity*:** This approach assesses the suspect and authentic websites based on various visual characteristics. Because the phishing web page appears to be highly similar to its legitimate page, these tools compare similarities: This approach uses CSS, text layout, source code, the website logo, screenshots of the web page, and other visual elements. Because these techniques compare the suspect web page to previously visited or saved web pages, they cannot detect zero-hour phishing attacks (Jain and Gupta, 2018).

Heuristic: The heuristic approach uses features derived from the phishing website. This strategy is based on several attributes that can differentiate a phishing website from a genuine one. These methods gather data from various sources, such as URLs, text content, DNS, digital certificates, and website traffic. The feature set, training samples, and classification algorithms all influence the success of this method. One of the advantages of this technique is that it can detect Zero-hour phishing attacks (Jain and Gupta, 2018).

Machine Learning: Nowadays, Machine Learning is a prevalent approach for detecting phishing websites (Sindhu et al., 2020). Common attributes such as URL information, website structure, and JavaScript features are collected to represent phishing URLs

and related websites. Then, based on those features, phishing data sets are obtained. After that, Machine Learning classifiers are trained to detect the phishing website based on those features (Zhu et al., 2020). This technique works very well with Big Data sets (having high Velocity, Variety, Volume, Value, and Veracity). Machine Learning-based classifiers achieved more than 99% accuracy, which proved to be the most effective method (Alkawaz et al., 2021).

Deep Learning: According to recent developments in Deep Learning methods, Deep Neural Network should perform better than conventional Machine Learning techniques in detecting phishing websites. Some well-known Deep Learning algorithms used for phishing detection are the Deep Neural Network, Recurrent Neural Network, Feed-Forward Deep Neural Network, limited Boltzmann machine, Convolutional Neural Network, deep belief network, and deep auto-encoder (Basit et al., 2020).

a. Motivation

The motivation for this review is derived from the fact that a detailed insight study is required for studying phishing website detection techniques. A search of the relevant literature did not yield a clear overview of all the major approaches in this area. A collaborated work that includes the techniques, data sets, and algorithms used in phishing website detection was not available in a systematic format. There was a need to study this area and give an accredited overview.

The purpose of this research is to evaluate and contribute to the following:

1. Discover the best phishing website detection techniques to help the security manager easily select the top technique among the anti-phishing approaches for their security systems.
2. A good review paper is needed that focuses on the techniques, data sets, and algorithms used by the scholars in the regarded area.

This remaining paper is structured as follows: There are six sections. Section 2 presents some background information on related works; Section 3 describes the research methodology; Section 4 reviews the selected 80 research papers on phishing website detecting techniques; Section 5 presents the results and discussion of 80 research articles; Section 6 presents the conclusion.

2. Background of related work

Many authors have explored the detection of phishing websites. However, only a few have conducted a systematic literature review on the topic, as described below.

Qabajeh et al. (Qabajeh et al., 2018) recently worked on conventional vs automated phishing detection techniques. The conventional anti-phishing methods include raising awareness, educating users, conducting periodic training or workshop, and using a legal perspective. The Computerized or automated anti-phishing approaches talks about list-based and Machine Learning Based techniques. More importantly, the paper compares these approaches' similarities, positive and negative elements from the user and performance perspectives. According to this study, Machine Learning and rule induction are suitable for combating phishing attacks. The limitations of this work are: the review is based on 67 research items, and the study does not include Deep Learning techniques for phishing website detection.

Zuraiq & Alkasassbeh (Zuraiq and Alkasassbeh, 2019) carried out a comprehensive review of current phishing detection methods. The study discusses anti-phishing techniques such as Heuristic, Content Based, and Fuzzy rule-based approaches. The study indicated that there are better methods for identifying phishing websites. The background of the work is based on research conducted between 2013 and 2018. The drawbacks of this work are that it analyzed only 18 studies and did not include Machine Learning, List Based and Deep Learning approaches for phishing website detection.

Kunju et al. (Kunju et al., 2019) used a survey method to detect phishing attacks. The research provides several phishing attack detection solutions and methodologies. According to the research, many of the proposed solutions were found to be insufficient in providing solutions to phishing attacks. The literature in this work includes only 14 studies which are in the period between 2007 and 2019. The study discusses only Machine Learning techniques for phishing website detection.

Benavides et al. (Benavides et al., 2020) conducted a systematic review to analyze different approaches of other researchers for detecting phishing attacks by applying Deep Learning algorithms. In conclusion, there is still a significant gap in the area of Deep Learning algorithms for phishing attack detection. The literature in this work includes only 19 studies published between 2014 and 2019. Only research articles with the essential topics of phishing and Deep Learning are considered in this paper.

Athulya & Praveen (Athulya and Praveen, 2020) addressed different phishing attacks, phishers' most recent phishing tactics, and anti-phishing strategies. In addition, the article aims to raise awareness regarding phishing attacks and strategies used for phishing detection. According to this study, the best way to prevent phishing attacks is to educate users about the different types of phishing attacks. Users can choose the best security software tools or applications to detect phishing attacks, such as anti-phishing browser extensions. The literature in this work is based on nine research items. The study does not include Deep Learning techniques for phishing website detection.

Basit et al. (Basit et al., 2020) reported a survey on artificial intelligence-based phishing detection techniques. The authors used statistical phishing reports to examine the harm and trends of phishing attempts. In the paper, Antiphishing evaluations are classified into four categories: Machine Learning, Hybrid Learning, Scenario-based and Deep Learning. The research shows that Machine Learning procedures produce the best results compared to other approaches. The work is based on literature published in the last ten years and analyzed only 21 research items.

Kathrine et al. (Kathrine et al., 2019) presented a framework to detect and prevent different types of phishing attacks. According to this study, Machine Learning based algorithms effectively detect true positive results. The limitations of this study are: the literature in this work discussed only 11 studies, and the research does not include Deep Learning techniques for mitigating phishing websites.

Korkmaz et al. (Korkmaz, 2020) proposed a review work for selecting features that can be used in URL-based phishing detection systems. This research aims to create a general survey resource for scientists who work on web page classification or network security. This study's limitation is that the work discussed only five studies in the literature.

Arshad et al. (Arshad et al., 2021) presented different types of phishing and anti-phishing techniques in their study. The SLR evaluated that phone phishing, Email Spoofing, spear phishing, and Email Manipulation are the frequently used phishing techniques. According to this study, the highest Accuracy was achieved through Machine Learning approaches. The research is limited by the fact that it is based on only 20 studies.

Catal et al. (Catal et al., 2022) worked on a systematic literature review, which answered nine research questions. The study's main aim is to identify, assess, and synthesize the results of Deep Learning approaches for phishing detection. According to this study, Supervised ML algorithms were applied in 42 studies out of 43. The most used algorithm was DNN, and the best performance was given by DNN and Hybrid DL algorithms. The work only discusses Deep Learning related studies for phishing detection.

Table 1 shows only three SLRs published in the last five years about phishing website detection techniques in the five influential journals selected in the current study.

Table 1
Summary of review papers on phishing detection systems.

Author & Year	SLR or not	Analyzed articles	Aim	Main findings	Limitations
Qabajeh et al., 2018	No	67	This review paper compares traditional anti-phishing methods, which includes raising awareness, educating users, conducting periodic training or workshop, and using a legal perspective. The Computerized anti-phishing techniques talk about list-based and machine-learning techniques.	Machine Learning and rule induction are suitable to combat phishing due to their high detection rate and, more importantly, the easy-to-understand outcomes.	Sixty-seven studies were analyzed in work, and the research did not discuss Deep Learning techniques.
Zuraiq and Alkasassbeh, 2019	No	18	This study examines several phishing detection methods, including heuristic, content-based, and fuzzy rule-based methods.	The study indicated no perfect method for identifying phishing websites.	The work analyzed only 18 studies and did not include Machine Learning, List-based, and Deep Learning approaches.
Kunju et al., 2019	No	14	This paper provides an overview of many Machine Learning algorithms for identifying phishing websites, including kNN, Naive Bayes, Decision tree, SVM, Neural Network, and Random Forest.	According to this study, phishing website detection using a single approach is insufficient.	The literature in this work included only 14 studies discussing Machine Learning techniques.
Benavides et al., 2020	Yes	19	This systematic literature review aimed to evaluate various other scholars' proposals for identifying phishing attacks using Deep Learning algorithms.	In conclusion, there is still a significant gap in the area of Deep Learning algorithms for phishing attack detection.	This work includes 19 studies, and only research articles on phishing and Deep Learning are considered in this study.
Athulya and Praveen, 2020	No	9	The research addressed phishing attacks, phishers' most recent phishing tactics, and anti-phishing techniques. In addition, the article aims to raise awareness regarding phishing attacks and strategies used for phishing detection.	According to this research, the best method to mitigate phishing attacks is to raise awareness for the users and select the best anti-phishing security software tool.	The literature in this work is based on nine research items, and the study does not include Deep Learning techniques for phishing website detection.
Basit et al., 2020	No	21	For phishing detection, the study examines Artificial intelligence approaches such as Machine Learning, Hybrid Learning, Scenario-based, and Deep Learning.	The study proved that Machine Learning procedures give the best results.	The work analyzed only 21 research items.
Kathrine et al., 2019	No	11	This work presents different phishing attacks with the latest prevention approaches. This paper proposed a framework to detect and prevent phishing attacks.	According to this study, Machine Learning-based algorithms effectively detect true positive results.	The work discussed only 11 studies, and the research does not include Deep Learning techniques for mitigating phishing websites.
Korkmaz, 2020	No	5	Proposed a review work on selecting features that can be used in URL-based phishing detection systems.	According to research, URL-based detection strategies are preferred to increase detection speed.	This study's limitation is that the work discussed only five studies in the literature.
*Arshad et al., 2021	Yes	20	In this study, different types of phishing and anti-phishing techniques are presented.	They evaluated that phone phishing, Email Spoofing, spear phishing, and email manipulation are the frequently used phishing techniques. The study analyzed that Machine Learning approaches have the highest Accuracy.	The work is based on only 20 studies.
*Catal et al., 2022	Yes	43	The work answers nine research questions. The main aim is to synthesize, assess, and analyses Deep Learning techniques for phishing detection.	According to this study, 42 studies applied Supervised ML algorithms out of 43 studies. The most used algorithm was DNN, and the best performance was given by DNN and Hybrid DL algorithms.	The work only discusses Deep Learning related studies for phishing detection.

* Papers were excluded from the inclusion–exclusion criteria; these were very much related to the topic included in the study.

3. Methodology

The systematic literature review is a research process that follows a set of rules. The paper follows the methodology introduced by Singh & Kaur (Singh and Kaur, 2018), Singh et al. (Singh and Beniwal, 2021), Kitchenham et al. (Kitchenham et al., 2010), and Brereton et al. (Brereton et al., 2007). The review methodology includes constructing research questions, identifying the list of electronic databases to be explored, data collection, data analysis, discussion on findings, and a comparison study of final selected research articles once all exclusion criteria have been applied. This systematic literature review aims to find the best approach, data set, and algorithm researchers employ for phishing website detection.

3.1. Methodology of the review

As discussed in the above para study will start by designing research questions and then explore the databases used for detection and analysis by comparing the findings of other literature as a part of the review methodology. The procedure includes searching primary and secondary databases, implementing inclusion–exclusion criteria, analyzing results, and discussions are all part of the

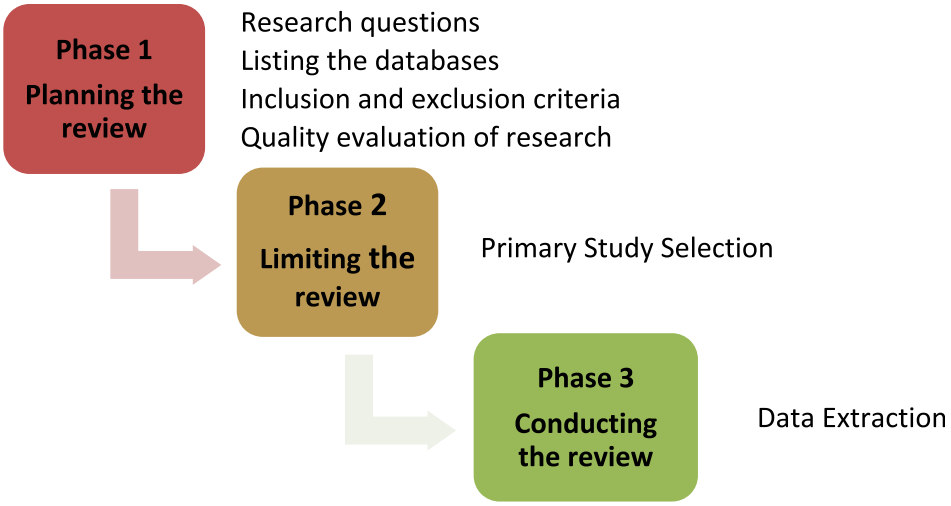

Fig. 2. Phases of Systematic Literature Review.

Fig. 3. Search exclusion criteria.

Table 2
Research questions.

RQ 1.	What are the phishing website detecting techniques, and which technique has been used in most studies?
RQ 2.	What are the different data sets used by the researchers to detect phishing websites, and which dataset has been used so far in most studies?
RQ 3.	Which algorithms have been used by authors, and which algorithm has been used mainly by the researcher?
RQ 4.	Which algorithm has the best Accuracy when it comes to detecting phishing attacks?

process, as shown in Fig. 2. Only electronic databases are explored for the literature survey, which includes the most reputable journals, conference proceedings, and research thesis. During the initial search, 537 papers were found, and only 80 research items were selected after applying the inclusion–exclusion criteria (see Fig. 3).

3.2. Research questions

Table 2 lists the research questions designed after a discussion among a team of four people that includes experts from related fields. The prime objective of the discussion in a team is to reveal the various phishing approaches, data sets utilized in the relevant studies, the algorithms used in the area, and the highest Accuracy achieved by the implemented algorithms.

3.3. Search the relevant documents

A comprehensive view is required to perform a systematic review. As a result, before beginning the review, an appropriate selection of databases should be identified to deliver relevant results based on the keywords quickly. We selected the following five databases for systematic review.

(a) ACM Digital Library (https://dl.acm.org).
(b) IEEE Explore (https://ieeexplore.ieee.org).
(c) Elsevier (https://www.elsevier.com).
(d) Springer (https://link.springer.com).
(e) Other Articles (Indexed in Scopus Journals (https://scopus.com)).

3.3.1. Source of review

(a) Review Articles.
(b) Conference Proceedings.

(c) Technical Reports that have been Published.
(d) Books Chapters.
(e) Researchers' theses.

3.4. Important keywords for research

The study explored all primary and additional sources of information for the given list of keywords. The search was conducted for papers published from January 2017 to February 2022. The following keywords were used to search the research items in each source:

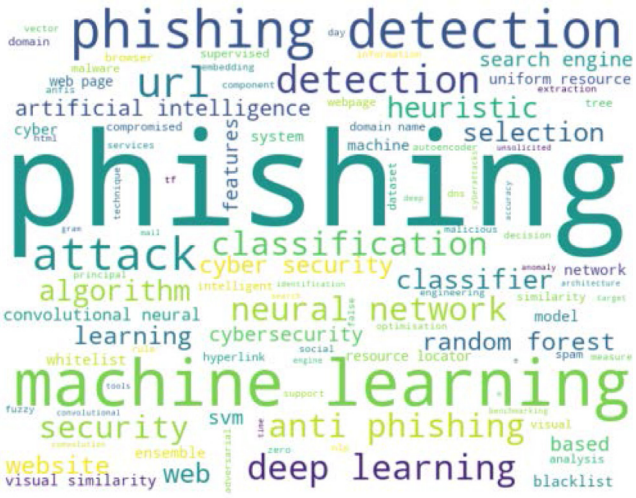

Fig. 4. Word cloud for the keywords of the selected research items.

Fig. 5. Word cloud for the titles of the research items.

(a) Phishing
(b) Phishing Detection
(c) Deep Learning
(d) Cyber Security
(e) Machine Learning

3.5. Inclusion and exclusion criteria

The inclusion–exclusion criteria were used at three levels. Unrelated papers are eliminated after each stage or level. The primary search included papers from the computer science and engineering disciplines. However, because the term "Machine Learning" is interdisciplinary, articles from other fields (such as medical science, food processing, material sciences, biomechanics, nanotechnology, and so on) were included in the list, and such types of papers were excluded from the study. Only papers written in English were considered for inclusion. The systematic review included research publications published between January 2017 and February 2022. The same research papers from multiple libraries are discarded. Serial research articles with minor changes published by the same authors are considered. If research is initially discussed at a conference and then issued in a journal, both sources are considered, with the latest version included in the study. The systematic review passes through three levels to reach the final set of research papers. As given in Fig. 3, a total of 537 publications were collected. After performing exclusion rules, 120 articles were included in the literature. Later on, from the set of these articles, 100 articles were selected based on their keywords and abstract reading. Finally, based on reading the complete text of the publications, 80 research papers were selected at the third level.

3.6. Quality evaluation of research

After inclusion–exclusion criteria had been finalized for the selection of articles and to meet the search's quality criteria, it was determined that the review would only be conducted on papers that had been approved at the scientific level and were in the field of Computer Science. The following indexed databases and repositories are selected: Elsevier, Scopus indexing journal, ACM, IEEE Explore, and Springer portal. Further, three documents are designed Appendices A–C to assure the quality parameter, which is set through inclusion–exclusion criteria. The purpose of these documents is to focus on the criteria set for Literature Survey. Quality assurance was done on the basis of these three Appendices by the Professor having expertise in the area of Cyber Security. The review process starts from Appendix A, and the reviewer moves to Appendix B after the satisfactory evaluation of Appendix A, then similarly to Appendix C.

3.7. Topical association

The Word Cloud technique shows how the articles are closely related according to the topical association theme. In most cases, word clouds are used to summarize text documents. The bigger and bolder word represents its frequency and importance in a

Fig. 6. Phishing website detection techniques.

Table 3
Primary studies on Heuristic techniques.

Ref	Applied Approach	Used Algorithm	Used Data set	Main Findings	Limitations/Challenges	Citation
Kumar et al., 2018	Heuristic & Machine Learning	Random Forest Multilayer Perceptron	UCI ML Repository 2949 legitimate emails 1378 spam emails 11,000 URL instances 30 features	Using the Random Forest classifier, the system can detect phishing and spam emails with an accuracy of 97.7% and 89.2%, respectively.	A study has used only two classifiers. The same dataset is used for training and testing. Challenges: for phishing detector, it has attained high Accuracy over the limited 88,73 instances.	9
Rao and Pais, 2019	Heuristic & Machine Learning	Support Vector Machine Random Forest Logistic Regression AdaBoost J48 Tree Multilayer Perceptron Sequential Minimal Optimization	PhishTank (2119 phishing websites), Alexa (1407 legitimate sites)	With an accuracy of 99.55%, the Random Forest algorithm outperformed the other algorithms.	Experiments were conducted with a limited dataset of 35,26 instances.	161
Babagoli et al., 2019	Heuristic & Machine Learning	Support Vector Machine Harmony search	UCI Machine Learning Repository 11,055 web pages 30 features	The study claims that Harmony search has a greater accuracy rate of 94.13% for training and 92.80% for testing operations.	The approach was tested on a limited number of data sets, and the data set used in this work contained only 11,055 instances. The work experimented with only two algorithms.	76
Gupta et al., 2021	Heuristic & Machine Learning	Support Vector Machine Random Forest K-Nearest Neighbor Logistic Regression	ISCXURL-2016 11,964 instances	With the Random Forest algorithm, the study obtained the highest Accuracy of 99.57%.	A study has yet to use the different datasets for training and testing to check the robustness of the proposed approach. Challenges: It has attained high Accuracy over the limited nine URL features.	24
Hr et al., 2020	Heuristic & Machine Learning	Random Forest	PhishTank 11,055 instances 30 features	This study achieved 99.36% accuracy in real-time phishing website detection.	The data set used in this work contains only 11,055 instances. The work experimented with only one algorithm.	13
Rao et al., 2019	Heuristic & Machine Learning	XGBoost Random Forest Logistic Regression K-Nearest Neighbor Support Vector Machine Decision Tree	Common Crawl Alexa PhishTank	The model achieved an accuracy of 94.26% using Random Forest.	The proposed model has achieved a slightly lower accuracy score than the other literature models. The model may fail to detect the phishing URLs which use shortening URLs and data URL services.	45
Ding et al., 2019	Visual Similarity, Heuristic & Machine Learning	Logistic Regression	PhishTank Yahoo URLBlacklist DMOZ	According to the experiments' results, this system's Accuracy is 98.90%.	Four data sets used in this work contain several domain name instances. The work experimented with only one algorithm.	40
Rao et al., 2020	Visual Similarity, Heuristic & Machine Learning	Support Vector Machine	PhishTank Alexa 11,000 instances.	The Twin Support Vector Machine classifier (TWSVM) beat the other versions in the experiments with an Accuracy of 98.05%.	The data set used in this work contains only 11,000 instances.	8
Almeida and Westphall, 2020	Heuristic & Machine Learning	–	PhishTank	The recommended approach detected phishing URLs at an average time of 30 s, with Accuracy rates ranging from 73.3% to 97.66%.	It has attained high Accuracy of 97.66% and the lowest of 73.3%, which is low compared to the other works in the literature with the Heuristic Machine Learning approach.	4

word cloud. Here, the word clouds visualize the word frequencies in the keywords shown in Fig. 4 and the word cloud for titles shown in Fig. 5 of the selected research items.

Phishing, Machine Learning, Phishing Detection, Cyber Security, and Deep Learning are among the most often used words in the keywords of the selected studies. In the word cloud, these are the most significant and boldest words shown in Fig. 4. Other words in small font appeared less frequently in the articles about phishing website detection techniques and may be considered less important keywords. Phishing detection, machine learning, phishing website, phishing, phishing attack, and Deep Learning are the most commonly used words in the titles shown in Fig. 5. Words found smaller in size in the cloud have been used in some specific articles as the title. However, they are not the commonly used ones. For both keywords and titles, highly focused words are Phishing, Machine Learning, Phishing Detection, Deep Learning, and Cyber Security.

3.8. Data extraction

During the initial search, 537 research papers were found. After applying the inclusion–exclusion criteria, only 80 works relevant to phishing website detection were selected (see Fig. 3). For this

Table 4
Primary studies on Visual Similarity techniques.

Ref	Applied Approach	Used Algorithm	Used Data set	Main Findings	Limitations/Challenges	Citation
Wang et al., 2020	Visual Similarity	–	PhishTank 60 phishing URLs	The author stated that the proposed OCR approaches could detect phishing websites and overcome the current approaches' drawbacks.	The approach was tested on very few numbers of phishing URLs. The application is device and language-dependent.	5
Ramana et al., 2021	Visual Similarity, Heuristic & Machine Learning	Decision Tree K-Nearest Neighbor AdaBoost Gradient Boost Logistic Regression Random Forest XGBoost	UCI Machine Learning Repository Mendeley. 21,055 instances	The experimental study attained a 97.51% Accuracy with a data set from UCI and using Mendeley's phishing data set for Machine Learning, and they obtained an Accuracy of 98.45%.	Data set 1 (D1) has a total number of 11,055 instances, and data set (D2) has 10,000 instances. The approach uses third-party features, which slows down the process.	2
van Dooremaal et al., 2021	Visual Similarity & Machine Learning	Logistic Regression	OpenPhish PhishTank PhishStats	The proposed system gives an accuracy of 99.20% for target identification and 99.66% for phishing classification on a data set.	The approach depends on search engine-based filtering, a third-party component; it may return different results for the same query over time.	3
Hidayat et al., 2021	Visual Similarity	Fuzzy Soft Set (FSS)	–	The experiment revealed similarity measures that surpassed the others in Accuracy and Recall, with 95.45% and 99.77%, respectively.	FSS has a promising approach to phishing detection in this study, although its response time was similar to the Matching Function.	0
Li et al., 2019	Visual Similarity, Heuristic & Machine Learning	Support Vector Machine Random Forest Decision Tree K-Nearest Neighbor XGBoost Gradient Boosting LightGBM	PhishTank Alexa 2000 web pages	The suggested approach achieves 98.60% accuracy.	The first data set used in the approach contains only 2000 web pages (1000 phishing and 1000 legitimate).	119
Rao and Pais, 2019	Visual Similarity	–	PhishTank Alexa	The system had an accuracy of 98.61%.	The performance of the proposed method gets affected by the services provided by the search engine. Due to the use of zero as the similarity threshold, Jail-Phish may lose some phishing sites. The legitimate sites hosted on free domain hosting providers are classified as phishing.	57
Liu and Fu, 2020	Visual Similarity & Heuristic	The unsupervised feature learning algorithm	PhishTank and OpenPhish (0.5 million malicious URLs),Alexa and DMOZ (1 million legitimate URLs)	The work achieved precision (over 95%).	Approach achieved precision (over 95%), recall (around 84%) As compared to other literary works, the performance achieved is less.	5
Barlow et al., 2020	Visual Similarity & Machine Learning	Neural Network	this data set has a mixture of 25 samples from the Bank of America Phish, PayPal Phish, ABSA Phish, DHL TRACKING Phish, and Microsoft Login Phish.	The system used a combination of neural networks paired with a binary visualization and achieved an overall detection accuracy of 94.16%.	The phishing websites data set contained a mixture of 25 samples which is very limited. The approach has used only one algorithm.	6
Bozkir and Aydos, 2020	Visual Similarity & Machine Learning	Support Vector MachineHistogram of Oriented Gradients (HOG)	PhishTank PhishTank OpenPhish	Experiments showed that the system could achieve a precision of 93.50%, a recall of 77.94%, and an F1 score of 85.02%.	The proposed scheme has some limitations due to the semi-rigid representation of HOG features. The approach used limited data of 204 snapshots for creating a detector for each brand.	28
Jain and Gupta, 2018	Visual Similarity	–	PhishTank OpenPhish Alexa 2000 URLs	The proposed method's performance was evaluated and found to have a 98.05% accuracy rate.	The proposed approach cannot detect the attached malware with web pages. The performance of the proposed approach depends on the search results and the extracted hyperlinks. If an attacker redesigns, the web page containing the approach may incorrectly classify the phishing web pages as legitimate.	45

(*continued on next page*)

Table 4 (*continued*)

Ref	Applied Approach	Used Algorithm	Used Data set	Main Findings	Limitations/Challenges	Citation
Li et al., 2020	Visual Similarity	Vision-based page segmentation (VIPS) algorithm	PhishTank Alexa 20 web pages	The experiments showed that the system has better robustness and Accuracy.	The authors selected only eight legitimate web pages and selected 12 phishing web pages as a dataset. A minimal dataset was used.	0
Jain et al., 2020	Visual Similarity	Term Frequency-inverse Document Frequency (TF-IDF)	PhishTank OpenPhish Alexa 200 instances	The Accuracy value for this approach is 89.0%.	The study has used a minimal number of features (tags), i.e., only five within the body tag of a web page. The approach was tested on a limited dataset, i.e., 100 legitimate sites and 100 phishing sites. The study itself claims a small size of the corpus.	8
Alsariera et al., 2020	Visual Similarity	LogitBoost-Extra Tree (LBET)Rotation Forest-Extra Tree (RoFBET)AdaBoost-Extra Tree (ABET) Bagging-Extra tree (BET)	UCI Machine Learning Repository Kaggle 11,055 instances 30 independent attributes	The LBET model achieved a detection accuracy greater than 97.5%.	The phishing website data set used in this work contains only 11,055 instances.	50

Table 5
Primary studies on List-Based techniques.

Ref	Applied Approach	Used Algorithm	Used Data set	Main Findings	Limitations	Citation
Barraclough et al., 2021	Blacklist-Based, Visual Similarity, Heuristic & Machine Learning	Adaptive neuro-fuzzy inference system (ANFIS) Nave Bayes PART J48 Tree JRip	PhishTank MillerSmiles Relbanks	The best performance was achieved by PART, which reached 99.33% Accuracy in 0.006 s.	The approach has a 0.6 percent error rate, which is relatively high.	14
Maroofi et al., 2020	List Based, Visual Similarity, Heuristic & Machine Learning	Random Forest Logistic Regression	PhishTank OpenPhish APWG URLhaus 38 features	The system produced an accuracy of 97.00% using the Random Forest classifier.	The study has used only two Machine Learning algorithms. It used third - party based features that slow down the process.	11
Azeez et al., 2021	White-list-based & Visual Similarity	–	PhishTank (140 phishing),Alexa (60 legitimate)	After six experiments, the average Accuracy was 96.17% achieved by the system.	The study has used a dataset of 200 sites, 140 phishing, and 60 legitimate web- sites.	12
Nathezhtha et al., 2019	DNS blacklist, Heuristic & Visual Similarity	–	PhishTank Alexa Google	The suggested system detects phishing and zero-day phishing attacks with a 98.90% accuracy.	The approach has used search engine-based features, which take time and slows down the process.	20
Rao and Pais, 2020	List Based, Visual Similarity, Heuristic & Machine Learning	Support Vector Machine Random Forest Decision Tree AdaBoost XGBoost	PhishTank (4097 instances),Google (5438 instances)	The authors suggested an ensemble model that combined Extra-Tree, Random Forest, and XGBoost to examine the significance of both heuristic-based and blacklist-based filters as a single entity, with an accuracy of 98.72%.	The data set used in the work has a small number of instances. Overall, the response time of the system is high.	22

systematic review, research works published from January 2017 to February 2022 were considered. The following is a description of the entire data extraction procedure:

b. One of the researchers reviewed all publications and collected data from all 80 primary papers.
c. The collection of work is validated by one of the most well-known professors (as an independent) in the cyber security domain, who evaluated the articles based on various criteria outlined in the data extraction and quality forms (Appendices A–D).
d. If there is a disagreement between the results, a meeting is scheduled to resolve the problem for selecting an article.

4. Phishing website detection approaches

To identify and prevent phishing attacks, various anti-phishing methods are available. As illustrated in Fig. 6, it is classified into five groups in this work.

The following section will discuss the literature based on phishing website detection techniques.

4.1. Heuristic technique

The Heuristic based approach uses features derived from phishing websites. This approach will help to distinguish between phishing and genuine website. Heuristics features are like right

Table 6
Primary studies on Machine Learning techniques.

Ref	Applied Approach	Used Algorithm	Used Data set	Main Findings	Limitations/Challenges	Citation
Shirazi et al., 2018	Machine Learning & Visual Similarity	Support Vector Machine Naive Bayes K-Nearest Neighbor Gradient boosting Decision tree	PhishTank (1000 URL) Alexa (1000 URL) OpenPhish (2013 URL)	According to this study, the Gradient boosting classifier with 97.00% accuracy produced the best results.	The study has used limited features based on domain name only, i.e., four binaries based and four non-binary based features. The training and testing data set for model evaluation is small and biased.	67
Hannousse and Yahiouche, 2021	Machine Learning, Visual Similarity & Heuristic	Support Vector Machine Decision tree Logistic Regression Random Forest Naive Bayes	PhishTank Alexa OpenPhish Yandex Search API 87 features	The best Accuracy score of 96.61% was achieved using hybrid features and the Random Forest classifier.	In this study, some content-based features for runtime analysis are unsuitable. No Feature Selection technique was used other than the manual selection of 87 features, which may create a bias in feature selection. No percentage of the Train-Test dataset split ratio was given.	7
Rashid et al., 2020	Machine Learning	Support Vector Machine	Alexa, Common Crawl archive (5000 URL)	The suggested method categorizes phishing and legal websites with 95.66% of Accuracy.	The study is very shallow and has used only one classifier, i.e., SVM, and five features for detecting phishing websites. A small data set was collected using GNU and Python scripts. Moreover, only one performance metric, i.e., Accuracy, was used for model evaluation.	11
Basit et al., 2020	Machine Learning	Random Forest K-Nearest Neighbor Decision tree Artificial Neural Network	UCI machine learning repository 11,055 instances 30 features	The combination of K-Nearest Neighbors and Random Forest classifier detects phishing attacks with 97.33% accuracy.	The study has not used multiple data sets to evaluate their ensemble model. Further, the UCI dataset is open source and has normalized features. It does not include the Original URL. The study has also not included any feature selection procedure. The study has picked the open-source data set and existing ML algorithms for their study. It still needs to include the calibration values of each selected ML approach.	25
Stobbs et al., 2020	Machine Learning, Heuristic & List Based	Random Forest Linear regression Neural Network Support Vector Machine	PhishTank Alexa	With a 99.33% Accuracy, Random Forest with PSO for feature selection and TPE for hyperparameter optimization was shown to be the optimum combination.	The study used different ML algorithms but did not give the split ratio of training and testing. All performance parameters could be better with the proposed approaches with two different data sets in comparison to other related work. In the case of the proposed approach, the Precision value could be better. Only Recall and Accuracy are best in comparison to other existing approaches.	3
Sahingoz et al., 2019	Machine Learning & Heuristic	Naive Bayes Random Forest K-Nearest Neighbor AdaBoost K-star SMO Decision tree	Ebbu 2017: Created own data set (73575 URL)	The Random Forest algorithm employs only NLP-based features detecting phishing URLs with a 97.98% accuracy rate.	Testing on multiple data sets still needs to be performed. The study has collected its dataset with its script. Further, in the case of short domain NLP based feature extraction techniques will not be able to detect these short domains.	347
Wu et al., 2019	Machine Learning & Heuristic	Support Vector Machine Decision tree Logistic Regression	PhishTank (5000 URL),DMOZ directory (10,000URL)	The Support Vector Machine produced the highest Accuracy.	The study has used a limited number of ML techniques. Also, no information about the configuration of Hyperparameters of ML algos was given. Only 16 features of URL were used for analysis. The Accuracy achieved, i.e.,89.3%, is much less than the comparative literature. As reported, the study cannot detect the sites imitated with images of legal websites.	18

(*continued on next page*)

Table 6 (*continued*)

Ref	Applied Approach	Used Algorithm	Used Data set	Main Findings	Limitations/Challenges	Citation
Abedin et al., 2020	Machine Learning & Heuristic	K-Nearest Neighbor Logistic Regression Random Forest	Kaggle 11,504 URL 32 attributes	The Random Forest classifier has an accuracy of 97.0%, a recall of 99.0%, and an F1 Score of 97.0% based on observations.	A limited number of ML algos were used in the study, and no information about the configuration of hyperparameters of ML algos was given. All the features listed in the data set were used, and the study intended to use a feature selection technique to reduce the selected features. The study has also not compared the results with any existing study.	4
Saha et al., 2020	Machine Learning	Random Forest Decision tree	Kaggle 11,504 URL 32 attributes	The highest Accuracy of 97.00% was achieved through the Random Forest classifier.	The study has used only two Machine Learning approaches and only a single dataset. They used the PCA feature selection technique for analyzing data set characteristics. The study intended to use CNN for anticipating phishing attacks. They did not compare the results with existing equivalent techniques. It is a very shallow study.	22
Mao, 2018	Machine Learning & Visual Similarity	Support Vector Machine Decision tree	PhishTank 2923 instance	According to the experiment results, both classifiers have more than 93% accuracy	The study has achieved minimal performance compared to other similar literature.It has used a limited no of instances, i.e., 2923, for evaluation with only two classifiers.	36
Sindhu et al., 2020	Machine Learning & Heuristic	Random Forest Support Vector Machine Neural Network	UCI Machine Learning repository11,055 URLs (6157 phishing, 4898 legitimate instances)	The work achieved accuracies of 97.36%, 97.45%, and 97.25%, respectively.	The study has not used multiple data sets to evaluate their model. Further, the UCI dataset is open source and has normalized features. It does not include the Original URL. The study has also not included any feature selection procedure.	9
Kasim, 2021	Machine Learning & Heuristic	Support Vector Machine LightGBM Multilayer Perceptron Convolution Neural Network	ISCXURL-2016 2978 instances and 77 different features	The current technique uses the Light Gradient Boosted Machine model to classify the features encoded with SAE-PCA at a rate of 99.60% accuracy.	The study has done the experiment on a limited dataset of 2978 instances, and PCA has reduced the feature selection from 77 to 20; these are also very limited.	2
Sánchez-Paniagua et al., 1267	Machine Learning & Heuristic	Random Forest K-Nearest Neighbor Support Vector Machine Naive Bayes Logistic Regression	Develop own dataset of 60,000 URLs of Legitimate and phishing sites. The study also used data sets of PWD2016 and Ebbu2017.	The results showed that Random Forest obtained the best result in classifying the index URLs with a 94.59% accuracy.	The results obtained were lower than the other similar literature. The collected dataset has index and login pages that reduce the performance parameters.	8
Butnaru et al., 2021	Machine Learning & Heuristic	Naive Bayes Decision tree Random Forest Support Vector Machine Multi-Layer perceptron	PhishTank 1,00,315 instances and 12 features, including two new proposed features	The highest Accuracy delivered by Optimized Random Forest is 99.29%.	Three out of five classifiers outperformed, and they compared the results with Google Safe Browsing (GSB), a popular protection available through many web browsers.	10
Munir Prince et al., 2021	Machine Learning &	Naive Bayes C4.5 JRip PART K-Nearest Neighbor Random Forest Support Vector Machine	Mendeley 10,000 websites instances, 48 attributes/ features	The study found that Random Forest had the highest Accuracy of 98.36%.	A very limited dataset was used for the evaluation. The study has not used the approach for feature reduction to remove overlapping features.	1
Geyik et al., 2021	Machine Learning	Decision tree Logistic Regression Naive Bayes Random Forest	PhishTank Alexa Common-crawl	The highest Accuracy produced by the Random Forest classifier is 83.0%.	Performance achieved with the dataset is very low compared to other studies with similar classifiers and datasets.	1

Table 6 (*continued*)

Ref	Applied Approach	Used Algorithm	Used Data set	Main Findings	Limitations/Challenges	Citation
Korkmaz et al., 2020	Machine Learning & Heuristic	Logistic Regression K-Nearest Neighbor Decision tree Support Vector Machine Naive Bayes XGBoost Random Forest Artificial Neural Network	PhishTank Alexa Common-crawl	In this work, the highest Accuracy was obtained by Random Forest applied on data set 1, which was 94.59%	Performance achieved with the dataset is very low compared to other studies with similar classifiers and datasets.	22
Patil et al., 2018	Machine Learning, List Based, Visual Similarity & Heuristic	Decision tree Logistic Regression Random Forest	Alexa 9076 websites	In this study, the highest Accuracy produced by the Random Forest classifier is 96.58%.	The study has used a limited number of existing approaches for model testing. Further, it was limited to a single data set which did not prove the robustness of the study. The data set used for evaluation is small compared to the other studies. In the evaluation of the result, a study has minimal false-positive and false-negative results	35
Yadollahi et al., 2019	Machine Learning & Heuristic	Decision tree AdaBoost Kstar Random Forest SMO Naive Bayes XCS	3983 phishing websites and 4021 legitimate websites	The results indicate that XCS achieved an accuracy of 98.39%.	The study has used a minimal data set for evaluation, i.e., 8004 URLs.	19
Palaniappan et al., 2020	Machine Learning, List Based & Heuristic	Logistic Regression	PhishTank Alexa ICANN DNS-BH (20,000 domain names)	Their work achieved an Accuracy of about 60.00% using a Logistic Regression classification algorithm.	Study results are significantly less as compared to many other studies. The study has applied only a single algorithm for evaluation.	20
Ozker and Sahingoz, 2020	Machine Learning & Visual Similarity	Naive Bayes Random Forest Support Vector Machine Logistic Regression K-Nearest Neighbor Decision tree Multilayer perceptron XGBoost	PhishTank 13,791 samples 58 features	The highest Accuracy obtained by Random Forest was 97.91%.	The study has used multiple ML techniques for the identification of Phishing sites but did not mention the reason for using several approaches. Dataset is not sufficient as compared to other studies. Their generated data set is not available for public usage. Feature selection can be applied to consider the essential features.	6
Shirazi et al., 2020	Machine Learning, Visual Similarity & Heuristic	Support Vector Machine Decision tree Gradient boosting K-Nearest Neighbor Random Forest	UCI machine learning repository Mendeley	The highest Accuracy was obtained by Gradient Boosting, which is 95.47%.	UCI dataset is open source and has normalized features. It does not include the Original URL.	12
Chiew et al., 2019	Machine Learning, Visual Similarity & Heuristic	Random Forest Support Vector Machine Naive Bayes C4.5 JRip PART	PhishTank OpenPhish Alexa Common Crawl archive 50,000 phishing and 50,00 valid URLs	The Random Forest classifier achieved the highest Accuracy, which is 96.17%.	The study has achieved comparatively low results for Accuracy as compared to other studies with the same dataset and classifier.	168
Parekh et al., 2018	Machine Learning & Heuristic	Random Forest	PhishTank 31 different URLs features	The accuracy level of the Random Forest was around 95%0.00%.	The study has achieved comparatively low results for Accuracy as compared to other studies with the same dataset and classifier. Also, only 31 different URLs were used for evaluation.	50
Bai, 2020	Machine Learning, Visual Similarity & Heuristic	Logistic Regression Support Vector Machine Naive Bayes Decision tree	PhishTank (3547 phishing web pages)DMOZ (3511 legitimate pages)	The optimal solution obtained by the Logistic Regression was 95.12%.	The study has achieved comparatively low results for Accuracy as compared to other studies with the same dataset and classifier.	8

(*continued on next page*)

Table 6 (*continued*)

Ref	Applied Approach	Used Algorithm	Used Data set	Main Findings	Limitations/Challenges	Citation
Anupam and Kar, 2021	Machine Learning & Heuristic	Support Vector Machine, Grey Wolf Optimizer algorithm, Bat Algorithm, Whale Optimization Algorithm, Firefly Algorithm	PhishTank Yahoo UCI machine learning repository	The results show that the Grey Wolf Optimizer (GWO) Algorithm outperformed all other algorithms in the system, with an accuracy of 90.38%.	UCI dataset is open source and has normalized features. It does not include the Original URL. The study has achieved comparatively low results for Accuracy as compared to other studies with the same dataset and classifier.	13
Suleman and Awan, 2019	Machine Learning & Heuristic	Naive Bayes Iterative Dichotomiser-3 K-Nearest Neighbor Decision tree Random Forest Genetic Algorithms	UCI machine learning repository	The research found that using ID3 along with Yet Another Generating Genetic Algorithm (YAGGA) gives the best Accuracy, 94.99%.	UCI dataset is open source and has normalized features. It does not include the Original URL.	19
Jain et al., 2018	Machine Learning & Heuristic	Support Vector Machine Naive Bayes	PhishTank 33,000 instances 14 features	Experimental results showed 91.28% accuracy in detecting phishing websites using the Support Vector Machine classifier.	The study has achieved comparatively low results for Accuracy as compared to other studies with the same dataset and classifier.	62
Zuhair and Selamat, 2019	Machine Learning, Visual Similarity & Heuristic	Phishing Hybrid Feature-Based Classifier (Naive Bayes Decision tree)	PhishTank Chinese e-Business DMOZ	The underperformance parameter study has achieved TPR equal to 0.984.	The study has not calculated the Accuracy; two existing algorithms were hybridized to obtain the results. The reason for choosing the two was also not given.	4
Ortiz Garces et al., 2019	Machine Learning & List Based Heuristic	Logistic Regression Neural Network	Kaggle 420,464 instances	The work investigated the analysis of anomalous behavior related to phishing web attacks and how machine learning techniques can solve the problem.	Only two algorithms were tested on 420,464 rows of data with two columns. It does not discuss the procedure of feature selection. The study randomly picked the Length, Subdomain, and Google index to identify the suspicious URL.	14

click disabled, '@' symbol in the URL, pop-up windows for passwords, and IP address in the domain part.

In Table 3, studies based on heuristic techniques are analyzed; among them, the study by (Gupta et al., 2021) by applying the Random Forest classifier achieved 99.57% accuracy, which is the highest Accuracy among the other article applying the heuristic approaches.

4.2. Visual Similarity based technique

In this technique, comparing two web pages is done according to the similar contents on the web page. These approaches use text content, formatting, CSS, source code, a screenshot of the web page, the website logo, images, and other visual elements.

In Table 4. Primary studies on Visual Similarity techniques are analyzed. Among them, the study by (Hidayat et al., 2021) achieved 99.77% accuracy with the Fuzzy set Technique, which is the highest Accuracy among the other articles applying the Visual-similarity approaches. Ramana et al. (Ramana et al., 2021) has also achieved comparative equivalent results of 98.45% of Accuracy with 21,055 instances. These are more promising regarding the large dataset and the maximum number of algorithms tested. While considering the reusability in the form of citation of scientific work, the community gives full endorsement to the work of Li. et al. (Li et al., 2019) with a citation of 119.

4.3. List based technique

Browsers like Microsoft Edge, Firefox, and Google Chrome utilize list-based methods to detect phishing websites. A blacklist includes a list of websites declared as spam, and a whitelist is web pages that browsers can access.

In Table 5, primary studies on the List Based technique are analyzed. Among them, the study by (Barraclough et al., 2021) achieved 99.33% accuracy by applying the PART algorithm, which is the highest Accuracy among the other article applying the List Based approaches. In the case of the List Based approach research community mostly cited the work of Rao & Pais (Rao and Pais, 2020) with a citation count of 22. They also achieved comparatively good results, with an Accuracy of 98.72%.

4.4. Machine Learning techniques

This method extracts features and uses machine learning algorithms to classify them. Common attributes such as URL information, website structure, and JavaScript features are collected to represent phishing URLs and related websites. Then, based on those features, phishing data sets are obtained. After that, machine learning classifiers are trained to detect the phishing website based on those features.

In Table 6. Primary studies on Machine Learning techniques are analyzed. Among them, the study by Stobbs et al. in 2020 (Stobbs et al., 2020) achieved 99.33% accuracy by applying the Random Forest algorithm, which is the highest Accuracy among the other article applying the Machine Learning approaches. As per the maximum reusability of work, the study of Sahingoz et al. (Sahingoz et al., 2019) was reused with a citation count of 347. However, the study created its dataset and analyzed it with seven algorithms. The common thing among the study is that they stamp Random Forest as the best algorithm with maximum Accuracy.

4.5. Deep Learning technique

According to recent developments in Deep Learning methods, Deep Neural Network should perform better than conventional

Table 7
Primary studies on Deep Learning techniques.

Ref	Applied Approach	Used Algorithm	Used Data set	Main Findings	Limitations/Challenges	Citation
Bu and Cho, 2021	Deep Learning & Heuristic	Recurrent Neural Network	PhishTank PhishStorm ISCX-URL-2016 222,541 URLs	The work claimed that the sensitivity improved by 3.98% compared to the previous work.	The limitation of the proposed methodology is that it was optimized for character-level features among the various features constituting URLs. Considering the structure of the web address consisting of domains and subdomains, additional performance improvements can be expected.	5
Korkmaz et al., 2021	Deep Learning & Heuristic	Convolutional Neural Network	PhishTank	The work produced 88.90% of Accuracy.	The study has achieved comparatively low results for Accuracy (88.90%).	1
Feng and Yue, 2020	Deep Learning & Heuristic	Recurrent Neural Network	PhishTank Alexa Common Crawl	The research proved that RNN models could achieve a detection accuracy of 99.50%.	Only one algorithm was tested in the study. From a data set of 1.5 million URLs, only 17 features were extracted.	9
Sirigineedi et al., 2020	Deep Learning, Heuristic & Machine Learning	Neural Network K-Nearest Neighbor Logistic Regression Support Vector Machine Gradient boosting Ada-boost Random Forest	GitHub	Using a Neural Network, the model can detect phishing URLs with an accuracy of 96.60%.	The approach has used external URL features based on the Whois command, which slows down the process.	4
Singh et al., 2020	Deep Learning	Convolutional Neural Network	Ebbu2017 GitHub 73,575 URLs	The proposed system achieved an accuracy of 98.0%.	The proposed system supported only 36,02 instances out of 36,400 legitimate URLs and supported only 37,55 phishing URL instances out of 37,175.	10
Feng, 2020	Deep Learning, Heuristic & Visual Similarity	Convolution Neural Network Recurrent Neural Network	PhishTank (21,303 phishing web pages),Alexa (24,800 regular web pages)	This method has a high accuracy of 99.05%.	The data set used in this work contains 24,800 regular web pages and 21,303 phishing web pages.	16
Opara et al., 2020	Deep Learning & Visual Similarity	Convolution Neural Network	PhishTank Alexa 500,000 domains	The work achieved over 93.0% Accuracy.	The proposed approach has used only one classifier and achieved comparatively low accuracy scores.	28
Wei et al., 2020	Deep Learning	Convolution Neural Network	PhishTank (10,604 phishing sites), Common Crawl (10,604 legitimate instances)	This work provided 99.98% accuracy.	The data set used in this work contains 10,604 regular URLs and 10,604 phishing URLs. The approach has used only one algorithm.	100
Yang et al., 2018	Deep Learning, Heuristic & Machine Learning	Convolutional Neural Network Recurrent Neural Network XGBoost	PhishTank dmoztools.net Total 201,0779 URLs	The highest success rate of 98.61 percent was obtained when CNN and LSTM were used together.	The WHOIS information in the URL features may slow down the process.	132
Al-Ahmadi and Alharbi, 2020	Deep Learning, Heuristic & Visual Similarity	Convolutional Neural Network	–	The experiment's results had a 99.67% accuracy rate.	The proposed approach has used only one classifier.	6
Abdelnabi et al., 2020	Deep Learning & Visual Similarity	Convolutional Neural Network	PhishTank Alexa SimilarWeb	The study shows that their technique outperforms previous Visual Similarity phishing detection systems while also resistant to several phishing attacks.	The proposed system focuses on Desktop browsers only. The domain names of the trusted list should update frequently.	32
AlEroud and Karabatis, 2020	Deep Learning & Heuristic	Generative Adversarial Networks	PhishTank DMOZ MillerSmiles	The experiments showed that GAN is highly successful even when used to deceive classifiers designed to defeat complex attacks.	The first data set contains 4898 phishing websites and 6157 regular sites. The second data set consists of only 12 features.	32
Saha et al., 2020	Deep Learning, Heuristic & Machine Learning	Multilayer Perceptron Neural Network	Kaggle (10,000 web pages), 10 features	The model's Accuracy in the training phase was 95.0%, while it was 93.00% in the testing phase.	The data set used in this work has only 10,000 URL samples. It has attained high Accuracy over the limited 10 features of URL.	18

Machine Learning techniques in detecting phishing websites. Some of the well-known Deep Learning algorithms used for phishing detection are the Deep Neural Network (DNN), Feed-Forward Deep Neural Network (FNN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN).

Table 7 shows the primary studies on Deep Learning techniques. Among them, a study done by (Wei et al., 2020 reported 99.98% accuracy with Convolution Neural Network, which is the highest Accuracy among all listed in the table. Regarding citation count for each article, the most referred article is by Yang et al.

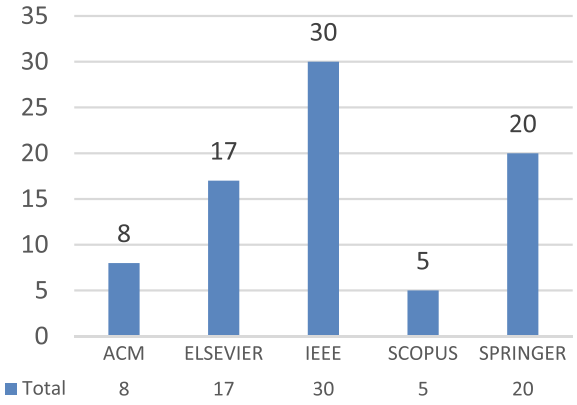

Fig. 7. Number of articles found by publisher.

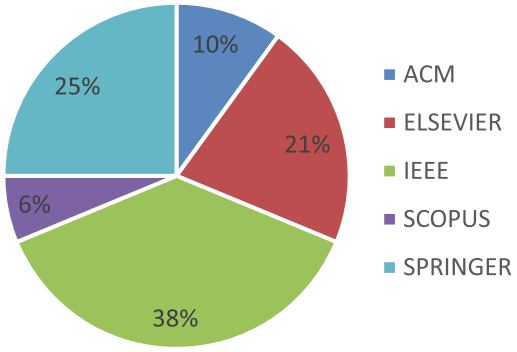

Fig. 8. Percentage of articles found by publisher.

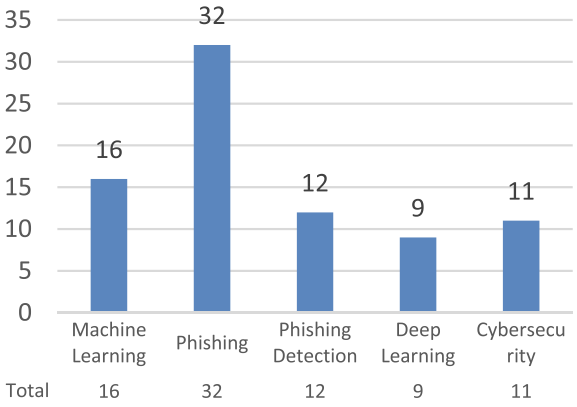

Fig. 9. Number of articles found by keywords.

(Yang et al., 2018) with a citation count 0f 132, this study has considered the 20,10,779 URL highest among all studies and has achieved the Accuracy of 98.61%, which is comparatively good. Both study vote for CNN for best accuracy results.

5. Results and discussion

The findings of the systematic literature review are organized based on the research questions in Table 2. In the current literature survey, after collecting 537 studies, we found only 80 relevant works linked to phishing attacks. These were selected for further critical studies (Fig. 3). From these 80 publications, 30 articles which are 38% of the current literature, are found in IEEE journal, 20 (or 25%) are in Springer, 17 (or 21%) in Elsevier, 8 (or 10%) are

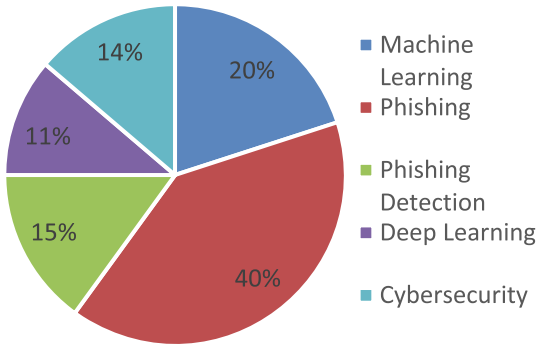

Fig. 10. Percentage of articles found by keywords.

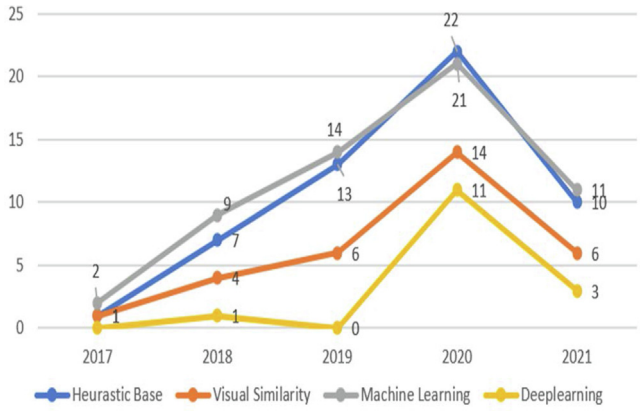

Fig. 11. The division of published articles with year-wise.

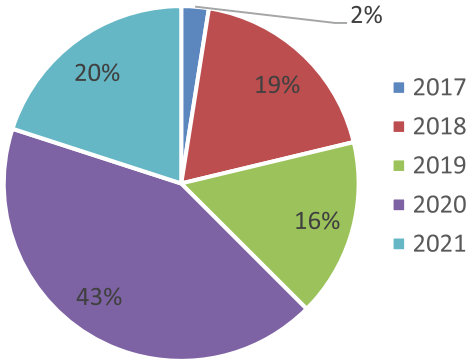

Fig. 12. Percentage of articles published by year.

with ACM and 5 (or 6%) are found with Scopus indexed journals as shown in Fig. 7 and Fig. 8. This shows that IEEE is leading in publication on this topic.

The above study also revealed that most of the publications searched on the mentioned databases are found for the keyword "Phishing" which is 32, and is 40% of the current research work. In contrast, the second highest keyword that retrieved most articles is Machine Learning, with 16 and 20% of the total articles reviewed. The next keyword is Phishing Detection, with 12 articles (or 15%). For the keywords Cyber Security, 11 articles (or 14%), and for the keyword Deep Learning, 9 articles (or 11%) were found, as shown in Fig. 9 and Fig. 10. As per the data, if we ignore the phishing keyword the next keyword which is primarily being used is Machine Learning. This inference machine learning technique is mainly used for detecting phishing sites.

Table 8
Articles versus Approaches.

Approaches	Articles Found	Total Articles Found
Visual-similarity	Ding et al., 2019; Rao et al., 2020; Wang et al., 2020; Ramana et al., 2021; van Dooremaal et al., 2021; Hidayat et al., 2021; Li et al., 2019; Rao and Pais, 2019; Liu and Fu, 2020; Barlow et al., 2020; Bozkir and Aydos, 2020; Jain and Gupta, 2018; Li et al., 2020; Jain et al., 2020; Alsariera et al., 2020; Barraclough et al., 2021; Maroofi et al., 2020; Azeez et al., 2021; Nathezhtha et al., 2019; Rao and Pais, 2020; Shirazi et al., 2018; Hannousse and Yahiouche, 2021; Rashid et al., 2020; Mao, 2018; Patil et al., 2018; Ozker and Sahingoz, 2020; Shirazi et al., 2020; Chiew et al., 2019; Bai, 2020; Zuhair and Selamat, 2019; Feng, 2020; Opara et al., 2020; Al-Ahmadi and Alharbi, 2020; Abdelnabi et al., 2020; Sonowal and Kuppusamy, 2020	**35**
Heuristic Based	Kumar et al., 2018; Rao and Pais, 2019; Babagoli et al., 2019; Gupta et al., 2021; Hr et al., 2020; Rao et al., 2019; Ding et al., 2019; Rao et al., 2020; Almeida and Westphall, 2020; Ramana et al., 2021; Li et al., 2019; Liu and Fu, 2020; Barraclough et al., 2021; Maroofi et al., 2020; Nathezhtha et al., 2019; Rao and Pais, 2020; Hannousse and Yahiouche, 2021; Rashid et al., 2020; Stobbs et al., 2020; Sahingoz et al., 2019; Wu et al., 2019; Abedin et al., 2020; Sindhu et al., 2020; Kasim, 2021; Sánchez-Paniagua et al., 1267; Butnaru et al., 2021; Korkmaz et al., 2020; Patil et al., 2018; Yadollahi et al., 2019; Palaniappan et al., 2020; Shirazi et al., 2020; Chiew et al., 2019; Parekh et al., 2018; Bai, 2020; Anupam and Kar, 2021; Suleman and Awan, 2019; Jain et al., 2018; Zuhair and Selamat, 2019; Ortiz Garces et al., 2019; Bu and Cho, 2021; Korkmaz et al., 2021; Feng and Yue, 2020; Sirigineedi et al., 2020; Feng, 2020; Yang et al., 2018; Al-Ahmadi and Alharbi, 2020; AlEroud and Karabatis, 2020; Saha et al., 2020; Sonowal and Kuppusamy, 2020; Adebowale et al., 2019; Feng et al., 2018; Lakshmi et al., 2021; Zouina and Outtaj, 2017	**53**
List Based	Barraclough et al., 2021; Maroofi et al., 2020; Azeez et al., 2021; Nathezhtha et al., 2019; Rao and Pais, 2020; Stobbs et al., 2020; Patil et al., 2018; Palaniappan et al., 2020; Ortiz Garces et al., 2019; Sonowal and Kuppusamy, 2020	**10**
Machine Learning	Kumar et al., 2018; Rao and Pais, 2019; Babagoli et al., 2019; Gupta et al., 2021; Hr et al., 2020; Rao et al., 2019; Ding et al., 2019; Rao et al., 2020; Ramana et al., 2021; van Dooremaal et al., 2021; Li et al., 2019; Barlow et al., 2020; Bozkir and Aydos, 2020; Barraclough et al., 2021; Maroofi et al., 2020; Rao and Pais, 2020; Shirazi et al., 2018; Hannousse and Yahiouche, 2021; Rashid et al., 2020; Basit et al., 2020; Stobbs et al., 2020; Sahingoz et al., 2019; Wu et al., 2019; Abedin et al., 2020; Saha et al., 2020; Mao, 2018; Sindhu et al., 2020; Kasim, 2021; Sánchez-Paniagua et al., 1267; Butnaru et al., 2021; Munir Prince et al., 2021; Geyik et al., 2021; Korkmaz et al., 2020; Patil et al., 2018; Yadollahi et al., 2019; Palaniappan et al., 2020; Ozker and Sahingoz, 2020; Shirazi et al., 2020; Chiew et al., 2019; Parekh et al., 2018; Bai, 2020; Anupam and Kar, 2021; Suleman and Awan, 2019; Jain et al., 2018; Zuhair and Selamat, 2019; Ortiz Garces et al., 2019; Sirigineedi et al., 2020; Yang et al., 2018; Saha et al., 2020; Sonowal and Kuppusamy, 2020; Feng et al., 2018; Zouina and Outtaj, 2017; Abutair and Belghith, 2017; Tupsamudre et al., 2019; Shirazi et al., 2019; Jain and Gupta, 2019; Kitchenham et al., 2010	**57**
Deep Learning	Bu and Cho, 2021; Korkmaz et al., 2021; Feng and Yue, 2020; Sirigineedi et al., 2020; Singh et al., 2020; Feng, 2020; Opara et al., 2020; Wei et al., 2020; Yang et al., 2018; Al-Ahmadi and Alharbi, 2020; Abdelnabi et al., 2020; AlEroud and Karabatis, 2020; Saha et al., 2020; Lakshmi et al., 2021	**14**

5.1. Trends of the research as per the approaches used

The research explored that out of 80 selected articles, 34 were published in 2020, which equates to 43% of the literature used in this work, and 16 articles (or 20%) were published in 2021. In 2018, 15 articles (or 19%) were published, 13 research articles (or 16%) were published in 2019, and in 2017, 2 (or 2%) articles were published. The division of published articles with year-wise detail is shown in Fig. 11, and the percentage of articles published by year is shown in Fig. 12. As per the information available and collected paper, the year 2020 has the maximum number of papers published with the second standing of the year 2021. Upon a deeper look into the publication of 2020, mainly the studies were done in Deep Learning i.e., 11, as compared to Machine Learning for the same year, i.e., 9. Further, in the next year of 2021, Machine Learning studies were counted to 6 compared to 3 for Deep Learning. Deep Learning is not considered by most of the studies for analysis. However, the results of Deep Learning are more promising than Machine Learning, as per Table 10, which is discussed in Research Question 3(*RQ3*).

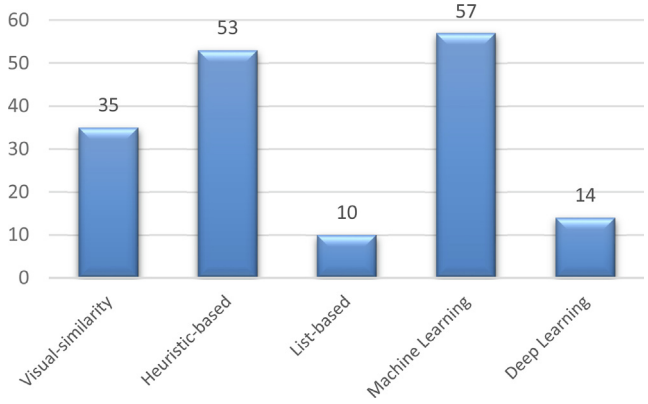

Fig. 13. Number of articles found by each technique.

5.2. Research question discussion

The present systematic literature review aims to address some of the important questions listed in Table 2. This study will try to answer all the research questions raised in Table 2 one by one as follow:

a. ***RQ1.*** *What are the phishing website detecting techniques, and which technique has been used in most studies?*

It was identified as part of a systematic literature review that the main anti-phishing approaches had been classified into five major areas: List Based, Visual-Similarity Based, Heuristic Based, Machine Learning, and Deep Learning based, as shown in Fig. 6 (see Table 8). Different work done by the scientific community under each technique is elaborated in section 4.

To answer this research question, the authors read 80 different publications thoroughly. Concerning the count of studies under each category, few studies used more than one technique for phishing detection. In this case, the study is listed under both techniques. Due to this count, the total number of articles found in column 3 of Table 8 is greater than 80. From those 80 studies, 57 papers applied machine learning approaches for phishing attack detection. As a result, 71.25% of the work has been done with machine learning algorithms among the five mentioned techniques, which is the highest compared to the other phishing detection techniques. Further, 53 (or 66.25%) articles used the Heuristic approach, followed by the Visual Similarity (35 or 43.75%articles) approach, then Deep Learning based (14 or 17.5%) and List based (10 and 12.5% articles) approach. The number of articles found by each technique is shown in Fig. 13.

Table 9
Articles versus Data sets.

Id	Data sets sources	Articles found	Total Articles Found
1	PhishTank	Rao and Pais, 2019; Hr et al., 2020; Rao et al., 2019; Ding et al., 2019; Rao et al., 2020; Almeida and Westphall, 2020; van Dooremaal et al., 2021; Li et al., 2019; Rao and Pais, 2019; Liu and Fu, 2020; Bozkir and Aydos, 2020; Jain and Gupta, 2018; Li et al., 2020; Jain et al., 2020; Barraclough et al., 2021; Maroofi et al., 2020; Azeez et al., 2021; Nathezhtha et al., 2019; Rao and Pais, 2020; Shirazi et al., 2018; Hannousse and Yahiouche, 2021; Stobbs et al., 2020; Wu et al., 2019; Mao, 2018; Sánchez-Paniagua et al., 1267; Butnaru et al., 2021; Geyik et al., 2021; Korkmaz et al., 2020; Palaniappan et al., 2020; Ozker and Sahingoz, 2020; Chiew et al., 2019; Parekh et al., 2018; Bai, 2020; Anupam and Kar, 2021; Jain et al., 2018; Zuhair and Selamat, 2019; Bu and Cho, 2021; Korkmaz et al., 2021; Feng and Yue, 2020; Feng, 2020; Opara et al., 2020; Wei et al., 2020; Yang et al., 2018; Abdelnabi et al., 2020; AlEroud and Karabatis, 2020; Sonowal and Kuppusamy, 2020; Adebowale et al., 2019; Feng et al., 2018; Zouina and Outtaj, 2017; Abutair and Belghith, 2017; Tupsamudre et al., 2019; Shirazi et al., 2019; Kitchenham et al., 2010	**53**
2	Alexa	Rao and Pais, 2019; Rao et al., 2019; Rao et al., 2020; Li et al., 2019; Rao and Pais, 2019; Liu and Fu, 2020; Jain and Gupta, 2018; Li et al., 2020; Jain et al., 2020; Azeez et al., 2021; Nathezhtha et al., 2019; Shirazi et al., 2018; Hannousse and Yahiouche, 2021; Rashid et al., 2020; Stobbs et al., 2020; Geyik et al., 2021; Korkmaz et al., 2020; Patil et al., 2018; Palaniappan et al., 2020; Chiew et al., 2019; Feng and Yue, 2020; Feng, 2020; Opara et al., 2020; Abdelnabi et al., 2020; Zouina and Outtaj, 2017; Abutair and Belghith, 2017; Shirazi et al., 2019; Jain and Gupta, 2019; Kitchenham et al., 2010	**29**
3	UCI Machine Learning Repository	Kumar et al., 2018; Babagoli et al., 2019; Ramana et al., 2021; Alsariera et al., 2020; Basit et al., 2020; Sindhu et al., 2020; Shirazi et al., 2020; Anupam and Kar, 2021; Suleman and Awan, 2019; Feng et al., 2018; Lakshmi et al., 2021; Shirazi et al., 2019; Jain and Gupta, 2019;	**13**
4	Common Crawl	Rao et al., 2019; Rashid et al., 2020; Geyik et al., 2021; Korkmaz et al., 2020; Chiew et al., 2019; Feng and Yue, 2020; Wei et al., 2020	**7**
5	Kaggle	Alsariera et al., 2020; Abedin et al., 2020; Saha et al., 2020; Ortiz Garces et al., 2019; Saha et al., 2020	**5**
6	GitHub	Sirigineedi et al., 2020; Singh et al., 2020	**2**
7	Mendeley	Ramana et al., 2021; Munir Prince et al., 2021; Shirazi et al., 2020; Shirazi et al., 2019	**4**
8	OpenPhish	van Dooremaal et al., 2021; Liu and Fu, 2020; Bozkir and Aydos, 2020; Jain and Gupta, 2018; Jain et al., 2020; Maroofi et al., 2020; Shirazi et al., 2018; Hannousse and Yahiouche, 2021; Chiew et al., 2019; Shirazi et al., 2019	**10**
9	PhishTank	Bozkir and Aydos, 2020	**1**
10	Yandex	Hannousse and Yahiouche, 2021; Sahingoz et al., 2019	**2**
11	ISCXURL-2016	Gupta et al., 2021; Kasim, 2021; Bu and Cho, 2021	**3**
12	Google	Nathezhtha et al., 2019; Rao and Pais, 2020; Feng et al., 2018; Jain and Gupta, 2019	**4**
13	PhishStorm	Bu and Cho, 2021	**1**
14	PhishStats	van Dooremaal et al., 2021	**1**
15	MillerSmiles	Barraclough et al., 2021; AlEroud and Karabatis, 2020; Feng et al., 2018; Jain and Gupta, 2019	**4**
16	Relbanks	Barraclough et al., 2021	**1**
17	Yahoo	Ding et al., 2019; Anupam and Kar, 2021; Jain and Gupta, 2019	**3**
18	URLBlacklist	Ding et al., 2019	**1**
19	DMOZ	Ding et al., 2019; Liu and Fu, 2020; Wu et al., 2019; Bai, 2020; Zuhair and Selamat, 2019; Yang et al., 2018; AlEroud and Karabatis, 2020; Abutair and Belghith, 2017; Tupsamudre et al., 2019	**9**
20	APWG	Maroofi et al., 2020	**1**
21	DNS-BH	Palaniappan et al., 2020	**1**
22	ICANN	Palaniappan et al., 2020	**1**
23	Million Quantcast	Sánchez-Paniagua et al., 1267	**1**
24	Chinese e-Business	Zuhair and Selamat, 2019	**1**
25	SimilarWeb	Abdelnabi et al., 2020	**1**

[1]http://https://www.phishtank.com, [2]http://https://www.alexa.com/topsites, [3]https://archive.ics.uci.edu, [4]http://index.commoncrawl.org, [5]https://https://www.kaggle.com, [6]https://github.com, [7]https://data.mendeley.com/, [8]https://openphish.com, [11] https://www.unb.ca/cic/data sets/url-2016.html,[19]https://https://www.dmoz.org/.

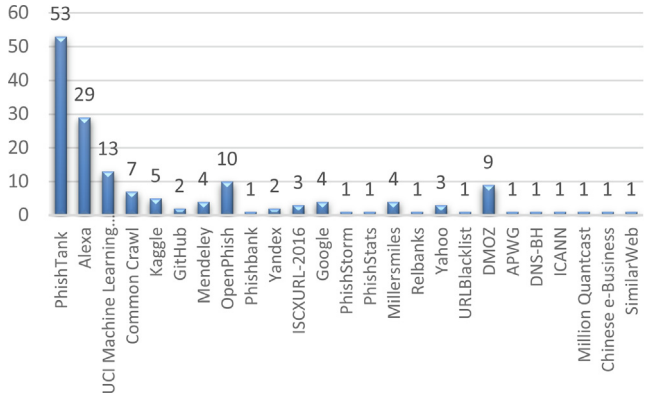

Fig. 14. Articles found by data set.

b. ***RQ2.*** *What are the different data sets used by the researchers to detect phishing websites, and which data set has been used so far in most studies?*

The current work has analyzed 25 data sets that are discussed and used by the researchers to detect phishing websites. For collecting phishing datasets, the PhishTank website was accessed by 53 studies which equal 66.25%, the highest among all the other datasets sources used in the literature. For downloading legitimate data sets, the Alexa website was accessed by 29 research items, 36.25% of all datasets, followed by 13 or (16.25%) UCI Machine Learning repository, 10 or (12.5%) OpenPhish dataset, and so on. The detailed count of data sets used is shown in Table 9 And Fig. 14.

c. ***RQ3.*** *Which algorithms are used by authors for phishing detection, and which algorithm is used mostly by the researcher?*

The study revealed that the researchers mostly use 24 different algorithms to detect phishing websites in the literature. The Random Forest Classifier was used in 31 of 80 studies, which is 38.75% of the work included in this study. Whereas SVM stands second and Decision Tree stands third in usability with 36.25% and 27.5%, respectively, as mentioned in Table 10 (see Fig. 15). CNN reported the highest Accuracy of 99.98%. In contrast, Random

Table 10
Articles versus Algorithm.

Id	Algorithms	Articles found	Total Articles Found	Highest Accuracy
1	Support Vector Machine	Rao and Pais, 2019; Babagoli et al., 2019; Gupta et al., 2021*, Rao et al., 2020; Li et al., 2019; Bozkir and Aydos, 2020; Rao and Pais, 2020; Shirazi et al., 2018; Hannousse and Yahiouche, 2021; Rashid et al., 2020; Stobbs et al., 2020; Wu et al., 2019; Mao, 2018; Sindhu et al., 2020; Kasim, 2021; Sánchez-Paniagua et al., 1267; Butnaru et al., 2021; Munir Prince et al., 2021; Korkmaz et al., 2020; Ozker and Sahingoz, 2020; Shirazi et al., 2020; Chiew et al., 2019; Bai, 2020; Anupam and Kar, 2021; Jain et al., 2018; Sirigineedi et al., 2020; Zouina and Outtaj, 2017; Shirazi et al., 2019; Kitchenham et al., 2010	**29**	97.64%
2	Random Forest	Kumar et al., 2018; Rao and Pais, 2019; Gupta et al., 2021*, Hr et al., 2020; Rao et al., 2019; Ramana et al., 2021; Li et al., 2019; Maroofi et al., 2020; Rao and Pais, 2020; Hannousse and Yahiouche, 2021; Basit et al., 2020; Stobbs et al., 2020; Sahingoz et al., 2019; Abedin et al., 2020; Saha et al., 2020; Sindhu et al., 2020; Sánchez-Paniagua et al., 1267; Butnaru et al., 2021; Munir Prince et al., 2021; Geyik et al., 2021; Korkmaz et al., 2020; Patil et al., 2018; Yadollahi et al., 2019; Ozker and Sahingoz, 2020; Shirazi et al., 2020; Chiew et al., 2019; Parekh et al., 2018; Suleman and Awan, 2019; Sirigineedi et al., 2020; Shirazi et al., 2019; Kitchenham et al., 2010	**31**	99.57%
3	Decision Tree	Ramana et al., 2021; Li et al., 2019; Rao and Pais, 2020; Shirazi et al., 2018; Hannousse and Yahiouche, 2021; Basit et al., 2020; Sahingoz et al., 2019*, Wu et al., 2019; Saha et al., 2020; Mao, 2018; Butnaru et al., 2021; Geyik et al., 2021; Korkmaz et al., 2020; Patil et al., 2018; Yadollahi et al., 2019; Ozker and Sahingoz, 2020; Shirazi et al., 2020; Bai, 2020; Suleman and Awan, 2019; Zuhair and Selamat, 2019; Shirazi et al., 2019; Jain and Gupta, 2019	**22**	97.02%
4	Naive Base	Barraclough et al., 2021*, Shirazi et al., 2018; Hannousse and Yahiouche, 2021; Sahingoz et al., 2019; Sánchez-Paniagua et al., 1267; Butnaru et al., 2021; Munir Prince et al., 2021; Geyik et al., 2021; Korkmaz et al., 2020; Yadollahi et al., 2019; Ozker and Sahingoz, 2020; Chiew et al., 2019; Bai, 2020; Suleman and Awan, 2019; Jain et al., 2018; Zuhair and Selamat, 2019; Kitchenham et al., 2010	**17**	99.33%
5	Logistic Regression	Ramana et al., 2021; Rao and Pais, 2019; Gupta et al., 2021; Ding et al., 2019*, Babagoli et al., 2019; van Dooremaal et al., 2021; Alsariera et al., 2020; Hannousse and Yahiouche, 2021; Wu et al., 2019; Abedin et al., 2020; Sánchez-Paniagua et al., 1267; Geyik et al., 2021; Korkmaz et al., 2020; Patil et al., 2018; Palaniappan et al., 2020; Ozker and Sahingoz, 2020; Bai, 2020; Ortiz Garces et al., 2019; Sirigineedi et al., 2020; Tupsamudre et al., 2019; Kitchenham et al., 2010	**21**	98.90%
6	K-Nearest Neighbor	Gupta et al., 2021*, Ramana et al., 2021; Li et al., 2019; Shirazi et al., 2018; Basit et al., 2020; Sahingoz et al., 2019; Abedin et al., 2020; Sánchez-Paniagua et al., 1267; Munir Prince et al., 2021; Korkmaz et al., 2020; Ozker and Sahingoz, 2020; Shirazi et al., 2020; Suleman and Awan, 2019; Sirigineedi et al., 2020; Abutair and Belghith, 2017; Shirazi et al., 2019	**16**	99.04%
7	AdaBoost	Rao and Pais, 2019*, Ramana et al., , 2021; Rao and Pais, 2020; Sahingoz et al., 2019; Yadollahi et al., 2019; Sirigineedi et al., 2020; Kitchenham et al., 2010	**7**	97.18%
8	XGBoost	Ramana et al., , 2021; Li et al., 2019; Rao and Pais, 2020; Korkmaz et al., 2020; Ozker and Sahingoz, 2020*, Yang et al., 2018	**6**	97.88%
9	Gradient Boosting	Ramana et al., 2021; Li et al., 2019; Shirazi et al., 2018*, Shirazi et al., 2020; Sirigineedi et al., 2020; Shirazi et al., 2019	**6**	97.00%
10	J48 Tree	Rao and Pais, 2019; Barraclough et al., 2021*	**2**	99.33%
11	Linear Regression	Stobbs et al., 2020*	**1**	91.29%
12	LightGBM	Li et al., 2019; Kasim, 2021*	**2**	99.60%
13	PART	Barraclough et al., 2021*, Munir Prince et al., 2021; Chiew et al., 2019	**3**	99.33%
14	JRip	Barraclough et al., 2021*, Munir Prince et al., 2021; Chiew et al., 2019	**3**	99.33%
15	C4.5	Munir Prince et al., 2021*, Chiew et al., 2019; Kitchenham et al., 2010	**3**	97.53%
16	Kstar	Sahingoz et al., 2019*, Yadollahi et al., 2019	**2**	95.27%
17	Multilayer Perceptron	Kumar et al., 2018; Rao and Pais, 2019; Kasim, 2021; Butnaru et al., 2021; Ozker and Sahingoz, 2020; Saha et al., 2020*	**6**	98.40%
18	Artificial Neural Network	Basit et al., 2020*, Korkmaz et al., 2020; Jain and Gupta, 2019	**3**	97.16%
19	Convolution Neural Network	Kasim, 2021; Korkmaz et al., 2021; Singh et al., 2020; Feng, 2020; Opara et al., 2020; Wei et al., 2020*, Yang et al., 2018; Al-Ahmadi and Alharbi, 2020; Abdelnabi et al., 2020	**9**	99.98%
20	Recurrent Neural Network (LSTM + BiLSTM)	Bu and Cho, 2021; Feng and Yue, 2020*, Feng, 2020; Yang et al., 2018	**4**	99.50%
21	Neural Network	Barlow et al., 2020; Stobbs et al., 2020; Sindhu et al., 2020; Ortiz Garces et al., 2019; Sirigineedi et al., 2020; Saha et al., 2020; Feng et al., 2018; Lakshmi et al., 2021*, Kitchenham et al., 2010	**9**	98.44%
22	Generative Adversarial Networks	AlEroud and Karabatis, 2020	**1**	–
23	Sequential Minimal Optimization	Rao and Pais, 2019; Sahingoz et al., 2019; Yadollahi et al., 2019; Kitchenham et al., 2010*	**4**	96.89%
24	Adaptive neuro-fuzzy inference system (ANFIS)	Barraclough et al., 2021*, Adebowale et al., 2019	**2**	99.00%

* Indicates the highest Accuracy obtained by the algorithm among the literature.

Forest was a mainly used algorithm and achieved the maximum Accuracy of 99.57%.

d. ***RQ4.*** *Which algorithm has the best Accuracy when it comes to detecting phishing attacks?*

After the evolution of the Convolution Neural Network technique, it gives the best accuracy results than traditional Machine Learning algorithms in any prediction analysis. The same is revealed in this comparison via the study of Wei et al. (Wei et al., 2020). They reported 99.98% accuracy in predicting the phishing attack, which is the highest among all other techniques. The second highest Accuracy among the other classifies was achieved by the LightGBM, which is 99.60%, as stated by Kasim (Kasim, 2021). As per the best-performing algorithms of traditional machine learning algorithms, Random Forest shows the best Accuracy with 99.57%, as reported by Gupta et al. (Gupta et al., 2021). Subsequent best algorithms, as per the performance of accuracy score in descending order, are RNN, J48, PART, Naive Bayes, KNN, and so on. Results are presented graphically in Fig. 16.

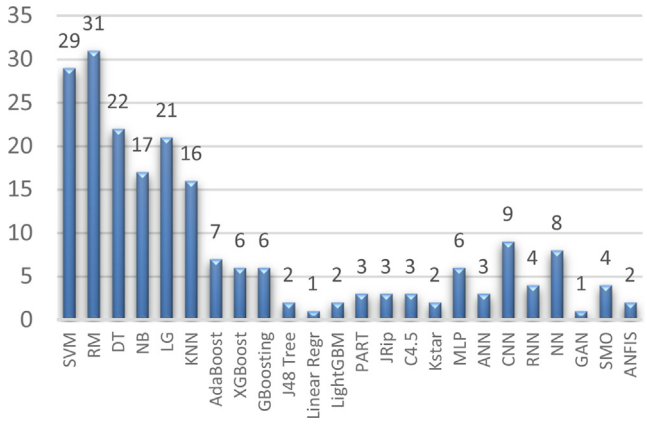

Fig. 15. Articles found by each algorithm.

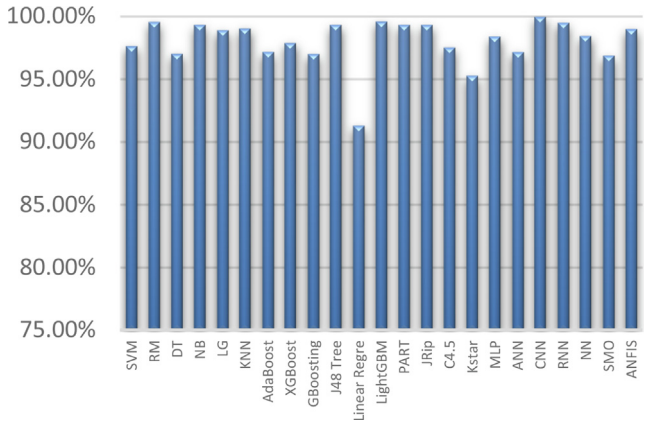

Fig. 16. Highest Accuracy produced by each algorithm.

5.3. Implication of research

This study will directly help scholars who want to know the detection technique of phishing attacks. The study will give platforms to scholars with essential information, like the dataset of phishing and legitimate websites. It is unique in terms of the collection of papers of the said period and methodology used (i.e., systematic literature review) which was not used earlier in this area of detecting phishing websites. This study will help various organizations and practitioners to be aware of the data set of faulty and legitimate websites, which will further help their end users to be safe while browsing and providing safe services. The study will help the policymakers or network administrator decide the technique or algorithm (along with their effectiveness) used for providing safe services.

5.4. Threats to validity

The primary threat to validity is that the authors had searched only mentioned five repositories mentioned in section 3. There might be several other publications at other resources. Nowadays, pre-prints are also gaining popularity and include several good publications, but due to the question of the work's authenticity, it was not included in the study. Further, the study follows the strict procedure of exclusion and inclusion of articles to include them in the study. For pre-verification of these criteria third person, an expert of a full-rank professor, was given with forms (Appendices A–D) to cross-check the criteria. In case of conflicting views, all three did a virtual meeting to resolve it. The study has considered those research articles that include the selected keywords. While excluding the articles, some good articles might have been skipped due to the fair inclusion and exclusion criteria.

6. Conclusion

The work done in this study involves the systematic literature survey of those studies which analyzed the performance of phishing website detection techniques. This study reports the dataset utilized and the algorithms used by the researchers in the previous five years in phishing website detection. A set of 537 research items from five electronic libraries were explored; after applying inclusion–exclusion criteria, the number of articles was reduced to 238. In the third exclusion criterion, it was reduced to 80 studies. A study of these 80 articles was performed by setting up research questions, and this was done to align the study in a direction. With the help of these research questions, this study will help to answer which technique, dataset, and algorithm were highly used in the literature and which algorithm or technique is performing best based on Accuracy.

In response to the first research question, based on the current survey, five phishing detection techniques are used mainly by the research community, as shown in Table 8. Among them, Machine Learning approaches have been used the most during the selected period. From 80 research items, 57 papers, or 71.25% of the studies, used Machine Learning approaches in their work. In addition, to answer the second research question, the survey revealed that mainly-two sources were used for analysis. For collecting phishing datasets, 53 or 66.25% of studies used the PhishTank website, whereas, for legitimate datasets, 29 or 36.25% of studies used the Alexa website. Twenty-five different datasets were used in these 80 studies. To answer the third and fourth research questions, the current study shows that authors used the Random Forest classifier, which is 38.75% out of 80 articles. Though the mostly used algorithm is Random Forest among the traditional Machine Learning algorithms, with the evolution of Convolution Neural Network (CNN), the Accuracy of the CNN algorithm is the best, i.e., 99.98% among all the studies included in this survey. It is regardless of the data set and features extracted for prediction analysis.

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.

Acknowledgements

I would like to express my gratitude to the Central University of Punjab, Bathinda for giving me the chance to participate as a research scholar there. We appreciate the independent reviewers for remarks and recommendations. He has given some great suggestions for achieving excellence in the work. We owe a debt of gratitude to our coworkers for their constant help.

Appendix A. Quality evaluation form 1

Did the systematic review related to phishing website detection techniques? yes or no Y N
Phishing website detection techniques are the subject of the study included in the systematic literature review. yes or no Y N
The articles, on the other hand, might be case study, a research article or an experimental study. yes or no Y N

If the reviewer answered Appendix A agreeably, then procced on to Appendix B.

Appendix B. Quality evaluation form 2

Was the goal of the paper to reveal various phishing website detection techniques? o Y o N
Did the work disclose various types of data utilized in phishing website detection? o Y o N
Did the SLR reveal different algorithms implemented in phishing website detection? o Y o N
Did the study disclose the highest accuracy archived by different algorithms? o Y o N

If the reviewer answered Appendix B agreeably, then procced on to Appendix C.

Appendix C. Quality evaluation form 3

Are the findings of the study stated clearly? o Y o N
Was the information provided are suitable for a comparative analysis? o Y o N

Appendix D. Data extraction form 4

Values	Full information
Date of data extraction	From January 2017 to February 2022
Bibliographic data	Paper title, Author, year, source of information
Set of databases	See Figs. 7 and 8
What method was used to do the comparison?	Based on the usage of every technique, data set used by the authors for phishing website detection, the algorithms used and the highest accuracy achieved by the algorithm.
Study findings	Figs. 9 and 10 highlight the major points from primary sources based on the keywords selected.

Appendix E. Data sources with relevant percentage 5

Databases	Papers found %
Scopus Indexed Journals	6.0%
ACM Digital Library	10.0%
IEEE eXplore	38.0%
Elsevier	21.0%
Springer	25.0%

Appendix F. Acronyms 6

Acronym	Full form
SLR	Systematic Literature Review
IC3	Internet Crime Complaint Center
TF-IDF	Term Frequency-Inverse Document Frequency
TWSVM	Twin Support Vector Machine Classifier
DOM	Document Object Model
FSS	Fuzzy Soft Set
URL	Uniform Resource Locator
HTML	Hypertext Markup Language
PSHCS	Phishing Sites Hosted on Compromised Servers
HOG	Histogram of Oriented Gradients
ABET	AdaBoost-Extra Tree
BET	Bagging-Extra Tree
RoFBET	Rotation Forest-Extra Tree
LBET	LogitBoost-Extra Tree
ANFIS	Adaptive Neuro-Fuzzy Inference System
ML	Machine Learning
NB	Naive Bayes
RF	Random Forest
MLP	Multilayer Perceptron
LR	Logistic Regression
DT	Decision Tree
SMO	Sequential Minimal Optimization
BN	Bayesian Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbor
NLP	Natural Language Processing
AAE	Adversarial Autoencoder
HEFS	Ensemble Feature Selection
GWO	Grey Wolf Optimizer
YAGGA	Yet Another Generating Genetic Algorithm
PHFBC	Phishing Hybrid Feature-Based Classifier
RFSSA	Recursive Feature Subset Selection Algorithm
ROC	Receiver-Operating Characteristic
CAE	Convolutional Autoencoder
RNN	Recurrent Neural Network
DL	Deep Learning
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
GAN	Generative Adversarial Networks

References

Abdelnabi, S., Krombholz, K., Fritz, M., 2020. VisualPhishNet: Zero-Day Phishing Website Detection by Visual Similarity. In: Proceedings of the ACM Conference on Computer and Communications Security, pp. 1681–1698. https://doi.org/10.1145/3372297.3417233.

Abedin, N.F., Bawm, R., Sarwar, T., Saifuddin, M., Rahman, M.A., Hossain, S., 2020. Phishing attack detection using machine learning classification techniques. In: Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020, pp. 1125–1130. https://doi.org/10.1109/ICISS49785.2020.9315895.

Abutair, H.Y.A., Belghith, A., 2017. Using case-based reasoning for phishing detection. Procedia Comput. Sci. 109, 281–288. https://doi.org/10.1016/j.procs.2017.05.352.

Adebowale, M.A., Lwin, K.T., Sánchez, E., Hossain, M.A., 2019. Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. Expert Syst. Appl. 115, 300–313. https://doi.org/10.1016/j.eswa.2018.07.067.

Al-Ahmadi, S., Alharbi, Y., 2020. A deep learning technique for web phishing detection combined Url features and visual similarity. Int. J. Comput. Netw. Commun. 12 (5), 41–54. https://doi.org/10.5121/ijcnc.2020.12503.
AlEroud, A., Karabatis, G., 2020. Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks. In: IWSPA 2020 - Proceedings of the 6th International Workshop on Security and Privacy Analytics, pp. 53–60. https://doi.org/10.1145/3375708.3380315.
Alkawaz, M.H., Steven, S.J., Hajamydeen, A.I., Ramli, R., 2021. A comprehensive survey on identification and analysis of phishing website based on machine learning methods. In: ISCAIE 2021 - IEEE 11th Symposium on Computer Applications and Industrial Electronics, pp. 82–87. https://doi.org/10.1109/ISCAIE51753.2021.9431794.
Almeida, R., Westphall, C., 2020. Heuristic Phishing Detection and URL Checking Methodology Based on Scraping and Web Crawling. In: Proceedings - 2020 IEEE International Conference on Intelligence and Security Informatics, ISI 2020, doi: 10.1109/ISI49825.2020.9280549.
Alsariera, Y.A., Adeyemo, V.E., Balogun, A.O., Alazzawi, A.K., 2020. AI meta-learners and extra-trees algorithm for the detection of phishing websites. IEEE Access 8, 142532–142542. https://doi.org/10.1109/ACCESS.2020.3013699.
Anupam, S., Kar, A.K., 2021. Phishing website detection using support vector machines and nature-inspired optimization algorithms. Telecommun. Syst. 76 (1), 17–32. https://doi.org/10.1007/s11235-020-00739-w.
Arshad, A, Rehman, A.U., Javaid, S., Ali, T.M., Sheikh, J.A., Azeem, M., 2021. A Systematic Literature Review on Phishing and Anti-Phishing Techniques. *arXiv*. https://doi.org/10.48550/arXiv.2104.01255.
Athulya, A.A., Praveen, K., 2020. Towards the Detection of Phishing Attacks. Proceedings of the 4th international Conference on Trends in Electronics and Informatics, ICOEI 2020, Icoei, pp. 337–343. https://doi.org/10.1109/ICOEI48184.2020.9142967.
Azeez, N.A., Misra, S., Margaret, I.A., Fernandez-Sanz, L., Abdulhamid, S.M., 2021. Adopting automated whitelist approach for detecting phishing attacks. Comput. Security 108,. https://doi.org/10.1016/j.cose.2021.102328 102328.
Babagoli, M., Aghababa, M.P., Solouk, V., 2019. Heuristic nonlinear regression strategy for detecting phishing websites. Soft Comput. 23 (12), 4315–4327. https://doi.org/10.1007/s00500-018-3084-2.
Bai, W., 2020. Phishing website detection based on machine learning algorithm. In: Proceedings - 2020 International Conference on Computing and Data Science, CDS 2020, pp. 293–298. https://doi.org/10.1109/CDS49703.2020.00064.
Barlow, L., Bendiab, G., Shiaeles, S., Savage, N., 2020. A Novel Approach to Detect Phishing Attacks using Binary Visualisation and Machine Learning. Proceedings - 2020 IEEE World Congress on Services SERVICES, 2020, pp. 177–182. https://doi.org/10.1109/SERVICES48979.2020.00046.
Barraclough, P.A., Fehringer, G., Woodward, J., 2021. Intelligent cyber-phishing detection for online. Comput. Security 104,. https://doi.org/10.1016/j.cose.2020.102123 102123.
Basit, A., Zafar, M., Liu, X., Javed, A.R., Jalil, Z., Kifayat, K., 2020. A comprehensive survey of AI-enabled phishing attacks detection techniques. Telecommun. Syst. 76 (1), 139–154. https://doi.org/10.1007/s11235-020-00733-2.
Basit, A., Zafar, M., Javed, A.R., Jalil, Z., 2020. A Novel Ensemble Machine Learning Method to Detect Phishing Attack. In: Proceedings - 2020 23rd IEEE International Multi-Topic Conference INMIC 2020. https://doi.org/10.1109/INMIC50486.2020.9318210.
Benavides, E., Fuertes, W., Sanchez, S., Sanchez, M., 2020. Classification of phishing attack solutions by employing deep learning techniques: a systematic literature review. In: Rocha, Á., Pereira, R. (eds) Developments and Advances in Defense and Security. Smart Innovation, Systems and Technologies, vol 152. Springer, Singapore. https://doi.org/10.1007/978-981-13-9155-2_5.
Bozkir, A.S., Aydos, M., 2020. LogoSENSE: A companion HOG based logo detection scheme for phishing web page and E-mail brand recognition. Comput. Security 95,. https://doi.org/10.1016/j.cose.2020.101855 101855.
Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M., Khalil, M., 2007. Lessons from applying the systematic literature review process within the software engineering domain. J. Syst. Softw. 80 (4), 571–583.
Bu, S.J., Cho, S.B., 2021. Deep character-level anomaly detection based on a convolutional autoencoder for zero-day phishing url detection. Electronics (Switzerland) 10 (12). https://doi.org/10.3390/electronics10121492.
Butnaru, A., Mylonas, A., Pitropakis, N., 2021. Towards lightweight url-based phishing detection. Future Internet 13 (6), 1–15. https://doi.org/10.3390/fi13060154.
Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., Shukla, S., 2022. Applications of Deep Learning for Phishing Detection: a Systematic Literature Review Knowl. Inf. Syst. 64 (6). https://doi.org/10.1007/s10115-022-01672-x.
Chiew, K.L., Tan, C.L., Wong, K.S., Yong, K.S.C., Tiong, W.K., 2019. a new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Inf. Sci. 484, 153–166. https://doi.org/10.1016/j.ins.2019.01.064.
Ding, Y., Luktarhan, N., Li, K., Slamu, W., 2019. A keyword-based combination approach for detecting phishing web pages. Comput. Security 84, 256–275. https://doi.org/10.1016/j.cose.2019.03.018.
Faris, H., Yazid, S., 2021. Phishing Web Page Detection Methods: URL and HTML Features Detection. In: IoTaIS 2020 - Proceedings: 2020 IEEE International Conference on Internet of Things and Intelligence Systems, pp. 167–171. https://doi.org/10.1109/IoTaIS50849.2021.9359694.
FBI, 2021. FBI Releases the Internet Crime Complaint Center 2020 Internet Crime Report, Including COVID-19 Scam Statistics. News, 2021, [Online]. Available: https://www.fbi.gov/news/pressrel/press-releases/fbi-releases-the-internet-crime-complaint-center-2020-internet-crime-report-including-covid-19-scam-statistics.
Feng, J., Yang Zou, L., Ye, O., Zhou Han, J., 2020. Web2Vec: Phishing webpage detection method based on multidimensional features driven by deep learning. IEEE Access 8. https://doi.org/10.1109/ACCESS.2020.3043188.
Feng, T., Yue, C., 2020. Visualizing and interpreting RNN Models in URL-based phishing detection. In: Proceedings of ACM Symposium on Access Control Models and Technologies, pp. 13–24. https://doi.org/10.1145/3381991.3395602.
Feng, F., Zhou, Q., Shen, Z., Yang, X., Han, L., Wang, J.Q., 2018. The application of a novel neural network in the detection of phishing websites. J. Ambient Intell. Hum. Comput., 1–15. https://doi.org/10.1007/s12652-018-0786-3.
Geng, G.G., Yan, Z.W., Zeng, Y., Jin, X.B., 2018. RRPhish: Anti-phishing via mining brand resources request. 2018 IEEE International Conference on Consumer Electronics, ICCE 2018, 2018-Janua, pp. 1–2. https://doi.org/10.1109/ICCE.2018.8326085.
Geyik, B., Erensoy, K., Kocyigit, E., 2021. Detection of Phishing Websites from URLs by using Classification Techniques on WEKA. In: Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, pp. 120–125. https://doi.org/10.1109/ICICT50816.2021.9358642.
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. Comput. Commun. 175 (April), 47–57. https://doi.org/10.1016/j.comcom.2021.04.023.
Hannousse, A., Yahiouche, S., 2021. Towards benchmark datasets for machine learning based website phishing detection: an experimental study. Eng. Applications Artif. Intell. 104, (April). https://doi.org/10.1016/j.engappai.2021.104347 104347.
Hidayat, R., Yanto, I.T.R., Ramli, A.A., Fudzee, M.F.M., 2021. Similarity measure fuzzy soft set for phishing detection. Int. J. Adv. Intell. Informatics 7 (1), 101–111. https://doi.org/10.26555/ijain.v7i1.605.
Hr, M.G., Mv, A., Gunesh Prasad, S., Vinay, S., 2020. Development of anti-phishing browser based on random forest and rule of extraction framework. Cybersecurity 3 (1), 1–14. https://doi.org/10.1186/s42400-020-00059-1.
Jain, A.K., Gupta, B.B., 2018. PHISH-SAFE: URL features-based phishing detection system using machine learning 729. https://doi.org/10.1007/978-981-10-8536-9_44.
Jain, A.K., Gupta, B.B., 2018. Two-level authentication approach to protect from phishing attacks in real time. J. Ambient Intell. Hum. Comput. 9 (6), 1783–1796. https://doi.org/10.1007/s12652-017-0616-z.
Jain, A.K., Gupta, B.B., 2019. A machine learning based approach for phishing detection using hyperlinks information. J. Ambient Intell. Hum. Comput. 10 (5), 2015–2028. https://doi.org/10.1007/s12652-018-0798-z.
Jain, A.K., Parashar, S., Katare, P., Sharma, I., 2020. PhishSKaPe: a content based approach to escape phishing attacks. Procedia Computer Sci. 171 (2019), 1102–1109. https://doi.org/10.1016/j.procs.2020.04.118.
Kasim, Ö., 2021. Automatic detection of phishing pages with event-based request processing, deep-hybrid feature extraction and light gradient boosted machine model. Telecommun. Syst. 78 (1), 103–115. https://doi.org/10.1007/s11235-021-00799-6.
Kathrine, G.J.W., Praise, P.M., Rose, A.A., Kalaivani, E.C., 2019. Variants of phishing attacks and their detection techniques. Proceedings of the international Conference on Trends in Electronics and Informatics, ICOEI 2019, Icoei, pp. 255–259. https://doi.org/10.1109/ICOEI.2019.8862697.
Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O.P., Turner, M., Niazi, M., Linkman, S., 2010. Systematic literature reviews in software engineering–a tertiary study. Inf. Softw. Technol. 52 (8), 792–805.
Korkmaz, M., Sahingoz, O.K., DIri, B., 2020. Detection of Phishing Websites by Using Machine Learning-Based URL Analysis. In: 2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020. https://doi.org/10.1109/ICCCNT49239.2020.9225561.
Korkmaz, M., Kocyigit, E., Sahingoz, O.K., Diri, B., 2021. Phishing Web Page Detection Using N-gram Features Extracted from URLs. In: HORA 2021 - 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings. https://doi.org/10.1109/HORA52670.2021.9461378.
Korkmaz, M., 2020. Feature Selections for the Classification of Web pages to Detect Phishing Attacks: A Survey. In: HORA 2020 - 2nd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings.
Kumar, S., Faizan, A., Viinikainen, A., Hamalainen, T., 2018. MLSPD - machine learning based spam and phishing detection, 11280 LNCS. Springer International Publishing. 10.1007/978-3-030-04648-4_43.
Kunju, M.V., Dainel, E., Anthony, H.C., Bhelwa, S., 2019. Evaluation of phishing techniques based on machine learning. 2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019, Iciccs, pp. 963–968. https://doi.org/10.1109/ICCS45141.2019.9065639.
Lakshmi, L., Reddy, M.P., Santhaiah, C., Reddy, U.J., 2021. Smart phishing detection in web pages using supervised deep learning classification and optimization technique ADAM. Wireless Pers. Commun. 118 (4), 3549–3564. https://doi.org/10.1007/s11277-021-08196-7.
Li, Y., Yang, Z., Chen, X., Yuan, H., Liu, W., 2019. A stacking model using URL and HTML features for phishing webpage detection. Future Gener. Comput. Syst. 94, 27–39. https://doi.org/10.1016/j.future.2018.11.004.
Li, J., Zhang, C., Yu, X., 2020. Webpage visual feature extraction and similarity algorithm. ACM Int. Conf. Proc. Ser., 80–85. https://doi.org/10.1145/3444370.3444552.
Liu, X., Fu, J., 2020. SPWalk: Similar Property Oriented Feature Learning for Phishing Detection. IEEE Access 8, 87031–87045. https://doi.org/10.1109/ACCESS.2020.2992381.
Liu, D.J., Geng, G.G., Jin, X.B., Wang, W., 2021. An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the

real web environment. Comput. Security 110,. https://doi.org/10.1016/j.cose.2021.102421 102421.
Mao, J. et al., 2018. Detecting phishing websites via aggregation analysis of page layouts. Procedia Computer Sci. 129, 224–230. https://doi.org/10.1016/j.procs.2018.03.053.
Maroofi, S., Korczynski, M., Hesselman, C., Ampeau, B., Duda, A., 2020. COMAR: Classification of Compromised versus Maliciously Registered Domains. In: Proceedings - 5th IEEE European Symposium on Security and Privacy, Euro S and P 2020, pp. 607–623, doi: 10.1109/EuroSP48549.2020.00045.
Munir Prince, M.S., Hasan, A., Muhammad Shah, F., 2021. A new ensemble model for phishing detection based on hybrid cumulative feature selection. In: ISCAIE 2021 - IEEE 11th Symposium on Computer Applications and Industrial Electronics, pp. 7–12. https://doi.org/10.1109/ISCAIE51753.2021.9431782.
Nakamura, A., Dobashi, F., 2019. Proactive phishing sites detection. In: Proceedings - 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI, pp. 443–448. https://doi.org/10.1145/3350546.3352565.
Nathezhtha, T., Sangeetha, D., Vaidehi, V., 2019. WC-PAD: Web crawling based phishing attack detection. In: Proceedings - International Carnahan Conference on Security Technology, vol. 2019-Octob, pp. 1–6, doi: 10.1109/CCST.2019.8888416.
Opara, C., Wei, B., Chen, Y., 2020. HTMLPhish: Enabling Phishing Web Page Detection by Applying Deep Learning Techniques on HTML Analysis. In: Proceedings of the International Joint Conference on Neural Networks. https://doi.org/10.1109/IJCNN48605.2020.9207707.
Ortiz Garces, I., Cazares, M.F., Andrade, R.O., 2019. Detection of phishing attacks with machine learning techniques in cognitive security architecture. In: Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019, pp. 366–370. https://doi.org/10.1109/CSCI49370.2019.00071.
Ozker, U., Sahingoz, O.K., 2020. Content Based Phishing Detection with Machine Learning. In: 2020 International Conference on Electrical Engineering, ICEE 2020. https://doi.org/10.1109/ICEE49691.2020.9249892.
Palaniappan, G., Sangeetha, S., Rajendran, B., Sanjay, S.G., Bindhumadhava, B.S., 2020. Malicious Domain Detection Using Machine Learning on Domain Name Features, Host-Based Features and Web-Based Features. Procedia Comput. Sci. 171 (2019), 654–661. https://doi.org/10.1016/j.procs.2020.04.071.
Paliath, S., Abu Qbeitah, M., Aldwairi, M., 2020. PhishOut: effective phishing detection using selected features. IEEE.
Parekh, S., Parikh, D., Kotak, S., Sankhe, S., 2018. A New Method for Detection of Phishing Websites: URL Detection. In: Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018, pp. 949–952. https://doi.org/10.1109/ICICCT.2018.8473085.
Patil, V., Thakkar, P., Shah, C., Bhat, T., Godse, S.P., 2018. Detection and Prevention of Phishing Websites Using Machine Learning Approach. In: Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018, pp. 1–5. https://doi.org/10.1109/ICCUBEA.2018.8697412.
Qabajeh, I., Thabtah, F., Chiclana, F., 2018. A recent review of conventional vs. automated cybersecurity anti-phishing techniques. Computer Sci. Rev. 29, 44–55. https://doi.org/10.1016/j.cosrev.2018.05.003.
Ramana, A.V., Rao, K.L., Rao, R.S., 2021. Stop-Phish: an intelligent phishing detection method using feature selection ensemble. Social Network Anal. Mining 11 (1), 1–9. https://doi.org/10.1007/s13278-021-00829-w.
Rao, R.S., Pais, A.R., 2019. Jail-Phish: An improved search engine based phishing detection system. Comput. Security 83, 246–267. https://doi.org/10.1016/j.cose.2019.02.011.
Rao, R.S., Pais, A.R., 2019. Detection of phishing websites using an efficient feature-based machine learning framework. Neural Comput. Appl. 31 (8), 3851–3873. https://doi.org/10.1007/s00521-017-3305-0.
Rao, R.S., Pais, A.R., 2020. Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach. J. Ambient Intell. Humanized Comput. 11 (9), 3853–3872. https://doi.org/10.1007/s12652-019-01637-z.
Rao, R.S., Vaishnavi, T., Pais, A.R., 2019. CatchPhish: detection of phishing websites by inspecting URLs. J. Ambient Intell. Humanized Comput. 11 (2), 813–825. https://doi.org/10.1007/s12652-019-01311-4.
Rao, R.S., Pais, A.R., Anand, P., 2020. A heuristic technique to detect phishing websites using TWSVM classifier. Neural Comput. Appl. 33 (11), 5733–5752. https://doi.org/10.1007/s00521-020-05354-z.
Rashid, J., Mahmood, T., Nisar, M.W., Nazir, T., 2020. Phishing Detection Using Machine Learning Technique. In: Proceedings - 2020 1st International Conference of Smart Systems and Emerging Technologies SMART-TECH 2020, pp. 43–46. https://doi.org/10.1109/SMART-TECH49988.2020.00026.
Saha, I., Sarma, D., Chakma, R.J., Alam, M.N., Sultana, A., Hossain, S., 2020. Phishing Attacks Detection using Machine Learning Approach. Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020, Icssit, pp. 1180–1185. https://doi.org/10.1109/ICSSIT48917.2020.9214132.
Saha, I., Sarma, D., Chakma, R.J., Alam, M.N., Sultana, A., Hossain, S., 2020. Phishing attacks detection using deep learning approach. Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020, Icssit, pp. 1180–1185. https://doi.org/10.1109/ICSSIT48917.2020.9214132.
Sahingoz, O.K., Buber, E., Demir, O., Diri, B., 2019. Machine learning based phishing detection from URLs. Expert Syst. Appl. 117, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029.
Sánchez-Paniagua, M., Fidalgo, E., González-Castro, V., Alegre, E., 1267. Impact of current phishing strategies in machine learning models for phishing detection. Adv. Intell. Syst. Comput. 1267 AISC (November 2020), 87–96. https://doi.org/10.1007/978-3-030-57805-3_9.
Shirazi, H., Bezawada, B., Ray, I., 2018, Know thy domain name: Unbiased phishing detection using domain name based features. In: Proceedings of ACM Symposium on Access Control Models and Technologies, SACMAT, pp. 69–75, doi: 10.1145/3205977.3205992.
Shirazi, H., Bezawada, B., Ray, I., Anderson, C., 2019. Adversarial sampling attacks against phishing detection. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11559 LNCS, no. June, pp. 83–101, doi: 10.1007/978-3-030-22479-0_5.
Shirazi, H., Muramudalige, S.R., Ray, I., Jayasumana, A.P., 2020. Improved Phishing Detection Algorithms using Adversarial Autoencoder Synthesized Data. Proc. – Conf. Local Computer Networks, LCN, 2020-Novem, pp. 24–32. https://doi.org/10.1109/LCN48667.2020.9314775.
Sindhu, S., Patil, S.P., Sreevalsan, A., Rahman, F., Saritha, A.N., 2020. Phishing detection using random forest, SVM and neural network with backpropagation. In: Proceedings of the International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2020, pp. 391–394. https://doi.org/10.1109/ICSTCEE49637.2020.9277256.
Singh, S., Beniwal, H., 2021. A survey on near-human conversational agents. J. King Saud Univ. – Comput. Inf. Sci. Volume 34 (10, Part A), 8852–8866. https://doi.org/10.1016/j.jksuci.2021.10.013.
Singh, S., Kaur, S., 2018. A systematic literature review: Refactoring for disclosing code smells in object oriented software. Ain Shams Eng. J. 9 (4), 2129–2151. https://doi.org/10.1016/j.asej.2017.03.002.
Singh, S., Singh, M.P., Pandey, R., 2020. Phishing detection from URLs using deep learning approach. In: Proceedings of the 2020 International Conference on Computing, Communication and Security, ICCCS 2020, pp. 16–19. https://doi.org/10.1109/ICCCS49678.2020.9277459.
Sirigineedi, S.S., Soni, J., Upadhyay, H., 2020. Learning-based models to detect runtime phishing activities using URLs. In: ACM International Conference Proceeding Series, pp. 102–106. https://doi.org/10.1145/3388142.3388170.
Somesha, M., Pais, A.R., Rao, R.S., Rathour, V.S., 2020. Efficient deep learning techniques for the detection of phishing websites. Sadhana – Acad. Proc. Eng. Sci. 45 (1). https://doi.org/10.1007/s12046-020-01392-4.
Sonowal, G., Kuppusamy, K.S., 2020. PhiDMA – A phishing detection model with multi-filter approach. J. King Saud Univ. – Comput. Inf. Sci. 32 (1), 99–112. https://doi.org/10.1016/j.jksuci.2017.07.005.
Stobbs, J., Issac, B., Jacob, S.M., 2020. Phishing web page detection using optimised machine learning. In: Proceedings - 2020 IEEE 19th international Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2020, pp. 483–490. https://doi.org/10.1109/TrustCom50675.2020.00072.
Suleman, M.T., Awan, S.M., 2019. Optimization of URL-Based Phishing Websites Detection through Genetic Algorithms. Automatic Control Comput. Sci. 53 (4), 333–341. https://doi.org/10.3103/S0146411619040102.
Tupsamudre, H., Singh, A.K., Lodha, S., 2019. Everything is in the name – a URL based approach for phishing detection, 11527 LNCS. Springer International Publishing. 10.1007/978-3-030-20951-3_21.
van Dooremaal, B., Burda, P., Allodi, L., Zannone, N., 2021. Combining text and visual features to improve the identification of cloned web pages for early phishing detection. ACM Int. Conf. Proc. Ser.. https://doi.org/10.1145/3465481.3470112.
Wang, Y., Liu, Y., Wu, T., Duncan, I., 2020. A Cost-Effective OCR Implementation to Prevent Phishing on Mobile Platforms," International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2020, doi: 10.1109/CyberSecurity49315.2020.9138873.
Wei, W., Ke, Q., Nowak, J., Korytkowski, M., Scherer, R., Woźniak, M., 2020. Accurate and fast URL phishing detector: a convolutional neural network approach. Comput. Netw. 178. https://doi.org/10.1016/j.comnet.2020.107275.
Wu, C.Y., Kuo, C.C., Yang, C.S., 2019. A Phishing Detection System based on Machine Learning. In: Proceedings - 2019 International Conference on Intelligent Computing and Its Emerging Applications, ICEA 2019, pp. 28–32. https://doi.org/10.1109/ICEA.2019.8858325.
Yadollahi, M.M., Shoeleh, F., Serkani, E., Madani, A., Gharaee, H., 2019. An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features. In: 2019 5th International Conference on Web Research, ICWR 2019, pp. 281–286. https://doi.org/10.1109/ICWR.2019.8765265.
Yang, L., Zhang, J., Wang, X., Li, Z., Li, Z., He, Y., 2021. An improved ELM-based and data preprocessing integrated approach for phishing detection considering comprehensive features. Expert Syst. Appl. 165 (July 2020). https://doi.org/10.1016/j.eswa.2020.113863.
Yang, P., Zhao, G., Zeng, P., 2018. Phishing website detection based on multidimensional features driven by deep learning. IEEE Access 7 (c), 15196–15209. https://doi.org/10.1109/ACCESS.2019.2892066.
Zabihimayvan, M., Doran, D., 2019. Fuzzy rough set feature selection to enhance phishing attack detection. IEEE Int. Conf. Fuzzy Syst. 2019, 1–6. https://doi.org/10.1109/FUZZ-IEEE.2019.8858884.
Zhu, E., Ju, Y., Chen, Z., Liu, F., Fang, X., 2020. DTOF-ANN: an artificial neural network phishing detection model based on decision tree and optimal features. Appl. Soft Comput. J. 95,. https://doi.org/10.1016/j.asoc.2020.106505 106505.
Zouina, M., Outtaj, B., 2017. A novel lightweight URL phishing detection system using SVM and similarity index. Human-Centric Comput. Inf. Sci. 7 (1), 1–13. https://doi.org/10.1186/s13673-017-0098-1.
Zuhair, H., Selamat, A., 2019. Phishing hybrid feature-based classifier by using recursive features subset selection and machine learning algorithms 843. https://doi.org/10.1007/978-3-319-99007-1_26.
Zuraiq, A.A., Alkasassbeh, M., 2019. Review: Phishing Detection Approaches. In: 2019 2nd international Conference on New Trends in Computing Sciences, pp. 1–6. https://doi.org/10.1109/ICTCS.2019.8923069.