Bachelor Degree Project

TECHNICAL IDENTIFIERS OF FRAUDULENT WEB PAGES, A SYSTEMATIC LITERATURE REVIEW

Bachelor Degree Project in Informatics with Specialization towards Network and System Administration
IT610G, G2E, 22.5 HP
Spring term 2020
Date of examination: 2020-06-15

Iker Orive Múgica
a19ikeor@student.his.se

Supervisor/Handledare: Joakim Kävrestad
Examiner/Examinator: Marcus Nohlberg

Acknowledgements

First of all, I want to thank my supervisor Joakim Kävrestad and the course examiner Marcus Nohlberg, who have guided me and provided the necessary feedback to achieve the results I was looking for.

Special appreciation is owed to the people who have helped me in the development of the project, regardless of the size of the contribution, because without them many aspects of the development would have been much more complicated and less bearable.

Getting through this project required more than academic support, that is why I want to thank the ones who always supported me: my family, friends, and my Animal Crossing villagers (except Brito a.k.a. Billy, who stole my birthday party and broke my heart). Jokes aside, I want to sincerely thank my family and friends as they have been a fundamental pillar not only during the development of the project but throughout my life.

Last but not least, I also want to thank the only person who has managed to get me up early for work, my landlord, who considered 8:00 a.m. an appropriate time to start drilling.

Abstract

Fraudulent pages are a danger to which all web users are exposed. These pages have illegitimate purposes such as the theft of sensitive user information. There are a lot of tools available on the market today that are aimed at detecting malicious pages, however, these are not reliable enough and that is why there is still a lot of room for future improvement. Therefore, further analysis of the malicious pages and their characteristics is a key element in protecting users and in the future in eradicating this type of malicious page.

A systematic review of the literature has been conducted to generate a list of features that can be used to detect malicious pages and that can ensure a high level of accuracy. During the development of the study, the different articles on the subject have been compared and analysed. For this process of analysis, thematic coding has been used, a qualitative method of analysis, which means that an in-depth understanding of ideas has been pursued. The document presents the already cited list of characteristics as well as offering suggestions and ideas that can be used in the development of future tools, by individuals using the Internet or by system administrators.

Keywords: Web, fraudulent, feature, tool, detection

Table of contents

List of Figures

List of Tables

1 Introduction

It is a common belief that humans are the weakest link in the security chain. According to Scheneier (2000) some people take advantage of that situation and the credulity of users; this has led to old scams being adapted to the web. Similarly, Oghaz et al. (2017) agree that the web has turned into a platform for fraudsters in which they can proliferate fraud and cybercrime. These fraudulent actions keep evolving and in recent years have continued to adapt to the newest technologies, consequently these frauds result on the loss of substantial amounts of money (Anderson et al., 2019). This is done with the purpose of tricking the user to reveal personal or sensitive information. These websites are a problem not just because of the fraud and the information theft but also because of the threat they are to the confidence of users when using the internet (Wuest & Ramzan, 2014) as users will feel less secure when using the Internet.

Most of the times, fraudulent websites will try to mistake users by making them believe that the non-legitimate website is the legitimate one by making these fraudulent websites look similar to their legitimate counterparts. According to Dawood et al. (2019) these websites can even trick well informed users. Taking this into account, the necessity of having a mechanism or tool to protect against fraudulent websites arises. Unfortunately, fraudulent sites such as phishing sites are constantly evolving and consequently countermeasures against them are not fully effective (Pienta et al., 2018).

According to Ushmani (2019) new preventive measures are needed to protect users against malicious websites, these measures could be for example effective data coding, legislative reforms, or fraud prevention strategies. This thesis will focus on a fraud prevention strategy, more specifically on technical identifiers as these can be used to develop tools for detecting malicious pages. Technical identifiers may be used in the detection of fraudulent websites as according to Oghaz et al. (2017) the fraudulent website detection methods depend on the components, contents and metadata of websites. Studies use components such as domain registration information, HTML (HyperText Markup Language) features and URLs (Uniform Resource Locator) to identify fraudulent websites.

In view of the large number of studies on the subject, the option of searching for new identifiers was ruled out and it was decided to generate a map of the most accurate identifiers already detected by the researchers. With this objective in mind, the systematic literature review was selected as a methodology because it is valued as an unbiased and reliable methodology to search and evaluate the already published literature. The methodology and its application will be further elaborated in chapter 4.

About the organization of the document, chapter 2 will begin by explaining the research already carried out and on which the thesis is based. Chapter will elaborate on the problem to be solved by explaining the objective of the project and its limitations. In chapter 4, as mentioned above, the methodology and its practical use

will be detailed. In chapters 5 and 6 the features found will be analyzed, synthesized and the importance of these will be assessed. The process, the validity of the results and the possible impacts of the project, both ethically and socially, will be analysed and discussed in chapter 7. Finally, in chapter 8 the future work and the conclusions drawn from the project will be presented.

2 Background

The aim of this chapter is to provide a background to the thesis which includes a general view of the fraudulent websites, and different ways these kinds of sites are handled. During the chapter, the scientific knowledge on which the thesis is based will also be discussed.

According to Abbasi et al. (2010), fraudulent websites can be divided into two categories: the ones attacking search engines and the ones whose targets are web users. The ones whose goal are search engines are called web spam. On the other hand, those that target users can be divided into two types: spoof sites and concocted sites. Web spam is defined by Najork (2017) as the systems to corrupt the "ranking algorithms of web search engines and cause them to rank search results higher than they would otherwise". Dinev (2006) states that spoof sites intend to trick users to extract their sensitive information by creating websites that look like legitimate web pages. Last, according to Abbasi et al. (2010) a concocted website is a false site that tries to look like legit e-commerce with malicious purposes.

As stated in the previous chapter, the world of fraudulent websites is a proliferating one and it is becoming more and more of a problem as these types of websites develop at a faster rate than the measures available to end or to avoid them.

One way we can resist the fraudulent pages is by using specialized tools against them. According to Amiri et al. (2015) these tools take different approaches and include classifiers such as hybrid classifiers, lookup classifiers, classifier, and ensemble systems.

Regarding lookup systems, Amiri et al. (2015) state that they depend on a list of fraudulent pages that is updated as new, non-legitimate pages are found. The client checks the blacklist and depending if the URL of the page is in it or not and it flags the page as unreliable or reliable depending on it. They also warn that these types of systems are more likely to fail because the malicious pages can succeed before they are listed.

In relation to classifier systems, Abbasi & Chen (2009) state that these systems "employ rule-based or similarity-based heuristics to content of website or domain registration information". With respect to rule-based systems the term is mostly used to refer to human-crafted systems (Mehrotra, 2020), so if the term is used in future sections of this thesis it will be to refer to these human-crafted systems which are represented in a simple "if-then" form (Kovarik, 2009). These types of tools do not depend on blacklist and therefore also they do not depend on when the web page is being visited, unlike lookup systems. Classifier systems are more susceptible of giving false positives and they take longer when classifying.

Regarding hybrid systems, these systems are the ones that bring together the lookup systems and the classifiers within the same tool (Amiri et al., 2015). In this way the errors that the two types of systems can give separately are minimized.

Finally in the classification of Amiri et al. (2015) are the ensemble systems, which are defined as systems that are based on the fusion of different models. In these systems the "individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples" (Dietterich, 2000).

Another protection method is individual monitoring of certain instructions. This is a first approach to the problem although it is not a fool proof method, as Dawood et al. (2019) state that even adequately prepared users can be mistaken by these sites. These rules can also be applied along with the previously mentioned tools. Governments and professionals in the field often provide some guidelines to protect users. For example, AT&T Inc. (How to Protect Yourself from Phishing and Fake Websites—DSL Internet Support, n.d.) provide some warning signs to look for and the following ones are the ones that apply when browsing the web:

- If the website you are entering uses a different URL than usual, it is likely to be fake.
- If a site asks to you confirm sensitive account information like your banking information, it may be a fraudulent web page.
- Spelling errors, especially on well-known websites, can be a sign that the page you are visiting is malicious.
- If the e-commerce you are entering is not a secure site (uses methods such as encryption to ensure that information remains safe) it is likely to be false.
- Low-quality sites (sites with low resolution images or texts, for example) are likely to be fake.

Professionals in the area also offer some guidance on how to avoid fake sites as users although in most of these studies these guidelines focus on the importance of being aware of the problem. According to Baral & Arachchilage (2019), users can navigate the web more safely when they begin to recognize factors that can expose pages as fraudulent, such as fake URLs. Similarly, in the study by Parulekar (2019) it is highlighted the importance of knowing the methods used by malicious sites along with their motivations in order to be more protected from them.

3 Problem Background

The subsequent chapter will deal with issues related to the definition of the problem such as the aim of the study, the limitations of the study and the context in which the problem lies.

Cybercrime is on the rise, IC3 (Internet Crime Complaint Center) stated that "2019 complaints and dollar losses were the highest since the center began tracking cybercrime statistics in 2000". This increase not only affects individuals or businesses that are victims of crime but can threaten the proper functioning of the economy (Gañán et al., 2017). Considering that these criminals attack individuals and companies alike, this problem is also considered very relevant from the system administrator's point of view.

The web is one of the areas that has the greatest capacity to accommodate new malicious activities, in in fields such as e-commerce for example (Ushmani, 2019). Ushmani (2019) states that preventive measures "such as legislative reforms, effective data coding and fraud prevention strategies" are needed in these new fraudulent areas.

Although a very wide range of detection tools and methods are available on the market, there are some research state that these methods are showing limitations (Minhaz Uddin et al., 2019). Therefore, there is a need for a detection method that can ensure a high level of accuracy in detecting malicious pages.

Regarding the methods of detecting malicious pages, Herzberg & Jbara (2008) underline the importance of indicators. After experimentation, they concluded that "improved indicators can significantly improve the detection rates", even if it is old this assessment can still be considered relevant today due to the numerous studies that try to find the most relevant features. Accordingly, and as will be expanded in more detail in the next subsection, this will be the area to address in the investigation.

3.1 Aim and Purpose

The thesis has been planned as part of a larger project, which in this case is the development of a tool for the detection of malicious pages. As part of this larger project, the objective of this thesis is the creation of a map or list with the technical identifiers of the pages with fraudulent or malicious character. Apart from laying the foundations for the future project, the thesis also aims to create greater awareness of the problem. With this in mind, the research question is:

What are the features that can help detect malicious pages with high accuracy?

Due to the foundations that are intended to be laid for the development of a tool, the project will also include some guidelines that may be helpful in it apart from the list with the most relevant features. These guidelines will be proposed as long as the

information presented can support them and will include aspects such as threshold values or directives on how to detect some malicious elements. Although the study will be oriented towards the future tool that will be developed, individuals and system administrators can also use some guidelines as it will help them distinguish between malicious and authentic pages more easily.

Apart from this, the thesis will also try to bring together the features that can provide a high accuracy in detection. The thesis will be based on the most relevant features of the literature, and these will be classified according to whether they can or cannot claim to provide a high accuracy.

The effect to be achieved is limited to medium-term time ranges as this is a reasonable time for the development of the hypothetical detection tool but it is not a long enough time for the pages to evolve so much that they make the results of the study outdated.

In conclusion, the aim of the project is to create a list of features that can be trusted to detect malicious pages. Apart from this, where possible, some guidelines will be given that may be helpful in the future development of a detection tool. These conclusions will be obtained from the research that has already been carried out, and for this purpose the research will be carried out in the form of a systematic literature review.

3.2 Limitations

Regarding the limitations of the project, one of the major limitations of the study is the already mentioned boundary of the area. As said before, the study is limited to the fraud in the web pages which leaves aside web frauds like email phishing. Another aspect which may limit the result of the research is the constant evolution of the field. Because the research is a systematic literature review the basis of the study is already carried out studies, this creates the possibility that recent improvements or aspects relevant to the study may not be included because of their newness.

There are also limitations on the number of people and the time available. Despite the fact that the available resources have been optimized to have the greatest possibilities of getting relevant results, it is undeniable that with more resources, more extensive research could be carried out.

Finally, it is also important to mention that the study will only list the features that are considered relevant. Therefore, in case the information about a feature is not sufficient, it will be clearly stated, but no research will be done to complement the existing information.

3.3 State of Art

For the detection of malicious pages there is a wide range of methods, and that is why in this chapter concerning the state of the classification made by Eshete (2013) will be

used as a starting point as it is the classification found that provides more data and explanations. Even so, more importance to the detection systems related to features will be given since it is the area to address in this research. Eshete (2013) developed the following list of approaches to malicious site detection:

- Blacklist approach
- Dynamic analysis approaches
- Heuristics based approaches
- Static analysis approaches

Blacklisting is one of the key approaches to malicious site detection. This method consists in putting the URLs considered as fraudulent in a list and checking against it the URL of the page that is being visited. Bell & Komisarczuk (2020) state that the effectiveness of this approach is dependent on characteristics such as the size, the scope, and the accuracy. In the study Google Safe Browsing, PhisTank, and OpenPhish phishing blacklists are regarded as key.

The dynamic analysis checks the execution of the web pages, among this type of analysis tools such as honey clients can be catalogued. Honey clients are detection tools that are intended to be compromised by the malicious pages. According to Bell & Komisarczuk (2020) honey clients are an effective technique in detection and are effective against the sites that try to exploit the browsers vulnerabilities. Yet studies such as Eshete (2013) conclude that while effective, these methods have a high computational cost and may not always be the best option.

Heuristic based approaches are the ones that rely on patterns of already detected malicious sites. According to Le et al. (2011) these patterns can be extracted from detection systems or antivirus applications. This approach method heavily depends on the features or patterns extracted as the judgement will be based on these.

Static approaches are the ones that analyze the features of the web without the need of executing it. Approaches such as feature analysis fall into this category, for example.

Machine learning approaches extract features and depending on them and the training received it categorizes sites as malicious or legitimate. Bell & Komisarczuk (2020) conclude that various approaches can be taken relative to the algorithms and extract the next ones as the most relevant from the literature:

- Support vector machine
- Logic regression
- Naïve Bayes
- Decision tree

Feature analysis uses the extracted features to categorize the websites as fraudulent or legitimate (Gupta et al., 2017). Fraudulent sites are detected with this method as they often have some features that differ from the ones of the legitimate sites (Eshete,

2013). The key to these static analysis approaches is the feature selection as the results will be based on them in its entirety.

There is a large number of studies related to the detection of malicious pages using features such as the ones that can be found on URLs or source code. The tools and studies carried out in these are categorized in the two approaches referred to the mentioned features (heuristic approach and static analysis) and some examples of these are presented below.

Li et al. (2020) for example develop a detection system based on URL features using linear and nonlinear space transformation methods intending to improve the detection of malicious sites based on URL features.

Other studies such as Li et al. (2019) for example do not limit the study to just one feature set and in this case, they rely on the URL and HTML features for detection. Zhang et al. (2014) similarly inspect the URL and the content of the sites to detect malicious sites, although this time the detection is focused on Chinese phishing online shops.

Ding et al. (2019) use Search & Heuristic Rule & Logistic Regression (SHLR) to detect sites that try to bypass detection methods with obfuscation techniques. In this study the detection is carried by using features as diverse as the title tag feature and heuristic character features.

4 Methodology

In this chapter the method used to conduct the research is explained as well as the reasoning behind the decision of the method selection. The aim of this work is to develop requirements for the development of a tool that flags potentially fraudulent web pages as no tool is accurate enough when doing this job.

To fulfil this objective, the selected methodology has been the systematic literature review, the process of conducting the systematic literature review is discussed in subchapter 4.1. Systematic literature review has been chosen as it is one of the methodologies that ensures an unbiased and reliable study while analysing and synthetizing research in the area to address. By definition, a literature review is a systematic review of the literature about one topic in which the research is analysed, evaluated, and synthesized (Efron & Ravid, 2019). According to Kitchenham (2004) one of the main reasons to perform a systematic literature review is to provide a background for future researches or developments, that is why this methodology is the one that best suited the project (synthesizing previous work by researches to get a list of technical identifiers with a high accuracy when detecting fraudulent websites). Kitchenham (2004) also highlight the usefulness of systematic literature reviews in the detection of gaps in research, aspect which may be relevant to propose future works in the area.

4.1 Systematic Literature Review

In a systematic literature review information collected from the sources is rigorous and unbiased. Pre-defined methods are also followed in order to assess the conclusion of studies (Efron & Ravid, 2019). As reported by Kitchenham (2004) regular literature reviews are of little scientific value as they can easily be biased and unfair. On the other hand, systematic literature reviews are trustworthy as they are performed following a predefined search strategy and can easily be repeated by other research getting consistent results. Although a more objective and fair research is achieved with systematic reviews, this is at the cost of a greater workload for the researcher as they require considerably more effort than regular reviews, declares Kitchenham (2004).

Kitchenman (2004), divided the development of a systematic literature review in three major stages:

1. Planning
2. Conducting
3. Reporting

The presented steps, due to their general and unspecific nature, can be easily divided into minor more precise points. One more specific division is the one done by Paré and Kitsiou (2017), as according to them a structured literature review can be divided in the following 7 steps:

1. Formulate a research question or aim
2. Perform literature searches
3. Apply inclusion and exclusion criteria
4. Perform quality assessment
5. Extract data
6. Analyse data

In the following figure (Figure 1) the procedure of developing the analysis is illustrated while bringing together the two presented divisions for greater clarity of the process. As well as showing the above-mentioned divisions the subdivisions made in some phases in the document will also be included in the figure for a better understanding of the report.

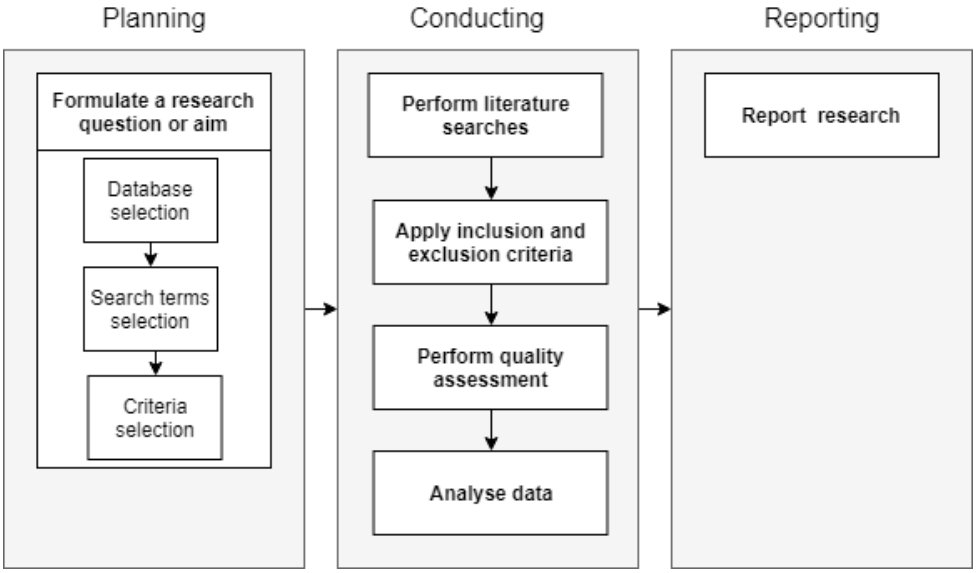


Figure 1: Phases of a systematic literature review (author's own)

To ensure the legitimacy and to avoid researcher bias it is important to pre-define a protocol which will be used to undertake the review. The most relevant of the steps for the research and the protocol decisions are furtherly discussed in following sections.

4.1.1 Databases

Another study on the topic by Fink (2019) emphasizes the importance of finding suitable information sources and of the precision on searches. Brereton et al. (2007) suggests IEEExplore, ACM Digital Library, ScienceDirect and CiteSeerX to develop research in the area of computer science. The proposed databases were analysed, and the relevance of the search results was checked regarding the topic of the research. After the study, CiteSeerX was discarded as the showed results were mainly not pertinent for the topic of the research.

To sum up, the following information sources where selected for the project:

- IEEExplore
- ACM Digital Library
- ScienceDirect

Another database that was evaluated was the use of Google scholar, this had to be finally discarded due to two main reasons, the difficulty of discerning between peer reviewed articles and the impossibility of automating processes such as downloading several documents. The fact of downloading one by one all the articles considered as relevant was considered a key factor to discard the database since it was an overwhelming job for only one person.

4.1.2 Search Terms

In order to optimize results some keywords where selected, this searching method is encouraged by Jesson et al. (2011) and Ridley (2012) as they agree on the use of keywords to get relevant records. According to Jesson et al. (2011) the use of boolean operators on search statements improves the search results. According to Wohlin et al. (2012) when making searches there is a trade-off between finding all pertinent studies and not getting an excessive number of false positives (incorrect result that is taken as positive when it should be false).

Taking these aspects into account and trying to find a balance between getting relevant results and not getting an overwhelming number of false positives, the coming ones were the selected search keywords:

- fraudulent AND website AND detection
- fake AND website AND detection
- identify AND fraudulent AND website
- identify AND fake AND website

As a starting point, the most descriptive words of the research (fraudulent, website, and detection) were selected and combined using boolean operators. After that, initial test searches were carried out to find equivalent words used in the research. Finally, the found synonyms and equivalent words, such as fake instead of fraudulent and identify instead of detection were introduced in the searches in order to cover the widest possible range of options that would provide relevant results.

4.1.3 Article Selection Criteria

For article selection a mixture of inclusion and exclusion criteria was used, if one of the inclusion criteria was not fulfilled the article was automatically discarded. On the other hand, if any of the exclusion criteria matched with the document it was also discarded from the research. According to Wohlin et al. (2012) the inclusion and exclusion criteria should be identified before starting the research process although it can be updated during the development of it, the starting exclusion and inclusion criteria can be checked in the next list:

- Starting inclusion criteria:
  - Peer reviewed
  - Written in English
  - Publication date between 2015 and 2020 (last 5 years)
  - Relevant to the topic
- Starting exclusion criteria:
  - Not meeting inclusion criteria
  - Need to pay or login
  - Duplicated articles

Finally, the criteria selected at the beginning has not been changed during the project development process as no reasons have been seen to alter them. Table 1 provides an overview of the selected criteria and the reasoning behind their choice.

Table 1: Inclusion and exclusion criteria

Criteria type	**Criterion**	**Reasoning**
Inclusion	Peer reviewed	In order to get the highest quality research as peer reviewed articles had been approved by experts in the area.

Criteria type	**Criterion**	**Reasoning**
	Written in English	In order to ensure proper understanding and transfer of information of the studies, the searched studies will be in the same language the thesis is going to be written.
	Publication date between 2015 and 2020 (last 5 years)	Since the world of the pages is constantly evolving, relatively current studies are sought in order not to rely on already outdated studies. It should also be noted that English is the most widely used language in research and is therefore the language that offers the widest range of articles. This limit of 5 years was decided looking for similar projects in the computer science area, although some variations were found in general the time limit used was approximately of 5 years
	Relevant to the topic	Searched articles must be useful for the study.
Exclusion	Not meeting inclusion criteria	Not meeting the inclusion criteria means that the paper is not relevant or suitable for the research.
	Need to pay or login	The search pool will be limited to open access papers or accessible through the databases with access provided by the university. Access provided by either of the two universities in which the author is enrolled, Skövde University or Mondragon Unibertsitatea.
	Identical to other found articles	Identical articles will be discarded in order to not falsify stats by increasing the importance of the data appearing in those duplicated articles.

4.1.4 Reverse Snowballing

The method of reverse snowballing, according to Wohlin (2014), is a method in which the reference list of papers is used to find new relevant articles. Although reverse snowballing can be used as an individual research method, in this case it is going to be used as a complementary method. The addition of this method as a complement is done with the aim of completing the results of the searches to achieve a wider and more exhaustive range of applicable studies for research.

To include the method within the study it has been decided to implement reverse snowballing on top of the database search. For this purpose, after getting the results of the database searches the reference list of the approved ones is examined in order to find further results. The above-mentioned studies are checked against the inclusion and exclusion criteria so that they are discarded or accepted for the research depending on their features. After that, the accepted studies are processed in the same way the articles found in the database searches.

Because the development of the thesis is carried out by only one person and because of the time limitations, it has been decided to carry out a single cycle of the process. This means that only accepted articles from database searches will be checked for new references and that this process will not be applied to the results from the reverse snowballing. It was decided to do a single iteration as the number of items would grow more and more with each iteration, making it very difficult to handle that large number of items by one person. It was also considered that due to the short range of time used as criteria for each iteration performed, it would be more difficult to find articles that met that requirement, making most of the work performed pointless.

4.1.5 Method of Analysis

Regarding the analysis of the accepted studies, the chosen method is thematic coding, a form of qualitative analysis. According to Gibbs (2007), coding is a method for ordering and arranging content to set up a “framework of thematic ideas” about it. Although oriented for psychology, the guidelines provided by Braun & Clarke (2006) can be applied to different fields such as computer science. The six-phase analysis they present is the one which is going to be followed in the development of the project, these are the steps:

1. Familiarizing with data
2. Generate initial codes
3. Look for themes
4. Inspect themes
5. Define and designate themes

6. Write paper

In accordance with Braun & Clarke (2006), the process starts with the reading of the paper the necessary times in order to note down the initial thoughts about the study. Second, data features are coded in a systematic way over the entire data set, collecting data relevant to each code. Then, we get all possible themes and assemble all important data linked to them. After that, a "thematic 'map' of the analysis" is generated. In the next stage, a defined explanation and naming are formed for each theme. Finally, in the last chance to analyse, the report is produced. During the whole development of the analysis thematic maps are generated so that data connections can be visualized. Thanks to these maps main themes can be easily identified.

4.1.6 Ethical Considerations

In order to avoid future conflicts regarding the thesis the ethical aspects of it had been attended. Although systematic literature reviews use publicly available information and do not collect delicate or classified data (Suri, 2020) there are still some elements to consider in the area.

Suri (2020) summarizes his work in the area in some bullet points to which we must prioritize and take care of with special emphasis. The following ones are the main given points:

- To represent the views and perspectives of the authors of the research in a credible manner so that nuances or aspects of the research are not lost.
- The applicability domain of the studies so that the results of the studies are not extrapolated to other research or contexts where they are not suitable.
- Ensure transparency of the research "to maximise an ethical impact of the review findings".

These ethical aspects and other problems that may appear such as incorrect references and citations are going to be carefully taken care of and will be thoroughly analysed so that no ethical problems arise.

Other doubt that may arise about the study is the malicious use that can be given to the results. Considering that aspects that can expose fraudulent pages are going to be analysed, is possible that this information could be used to further blur the line between fraudulent and authentic pages. Even so, after weighing up the pros and cons that the study may entail, it is argued that well-intentioned purposes prevail over malicious ones and therefore it has been decided to continue with the study.

4.1.7 Threats to Validity

Zhou et al. (2016) categorize assessing threats to validity as something critical when trying to provide quality research on Systematic literature reviews in the area of Software Engineering. Regarding from when these aspects should start to be assessed

Wohlin et al. (2012) stress the importance of assessing these threats from the early stages of study in order to get adequate validity.

Although there may be lots of different threats to the validity of the research Wohlin et al. (2012) and Zhou et al. (2016) agree on the categorizations they made depending on their features, these categorizations resulted in the following four sections:

- *Construct validity (CT).* Validity is related to the generalization of the outcome to the principles on which the research is based
- *Internal validity (IT).* Validity is related to "influences that can affect the independent variable with respect to causality, without the researcher's knowledge" (Wohlin et al., 2012).
- *External validity (EX).* Validity is related to the conditions that influence the generalisation of results in an industrial environment and limit the ability to do so.
- *Conclusion validity (CN).* Validity is related to getting the correct outcome from the connections between the result and the approach to the research.

These presented categories are not exclusive between them and more than one can apply to each threat. Zhou et al. (2016) proposed a list of threats which were divided in three major phases: planning phase, conducting phase, and reporting phase. In the following table (Table 2) we can see the threats from the proposed list which have been considered most important by the author for this study in addition to how these threats will be addressed in a mixture of the ideas proposed by Zhou et al. (2016) and the author's own.

Table 2: Threats to validity of the research

Phase	**Validity**	**Threat**	**Way to address**
Planning	CT, IN	The given details for the literature reviews are not enough (T1)	Strong criteria and protocols are going to be used in the development
	CT, IN	Inadequate search terms (T2)	All rules such as search terms are going to be checked multiple times.
	IN	Deficient number of studies (T3)	Multiple trustful databases are going to be used.
		Wrong inclusion and exclusion criteria (T4)	Criteria is going to be checked multiple times
Conducting	IN, CN	Bias when including or excluding studies (T5)	Inclusion and exclusion criteria will be followed closely.
	IN, CN	Duplicated studies (T6)	All studies will be examined, and duplicated ones will be removed
	IN	Inclination to positive results (T7)	Only the already defined criteria are followed without giving rise to opinions or other external factors

	IN	Subjective quality estimation (T8)	Strong inclusion and exclusion criteria are going to be used to get the higher quality sources.
Reporting	ET	Focus in a limited area (T9)	Decisions such as search terms and criteria are going to be made in order to broaden the area and have high primary study generalizability

Zhou et al. (2016) also show the problems that can arise if threats are not satisfactorily addressed as it is shown in the figure below (Figure 2). Each of the threats has a code assigned (T1, T2 etc.) as shown in the table above (Table 2).

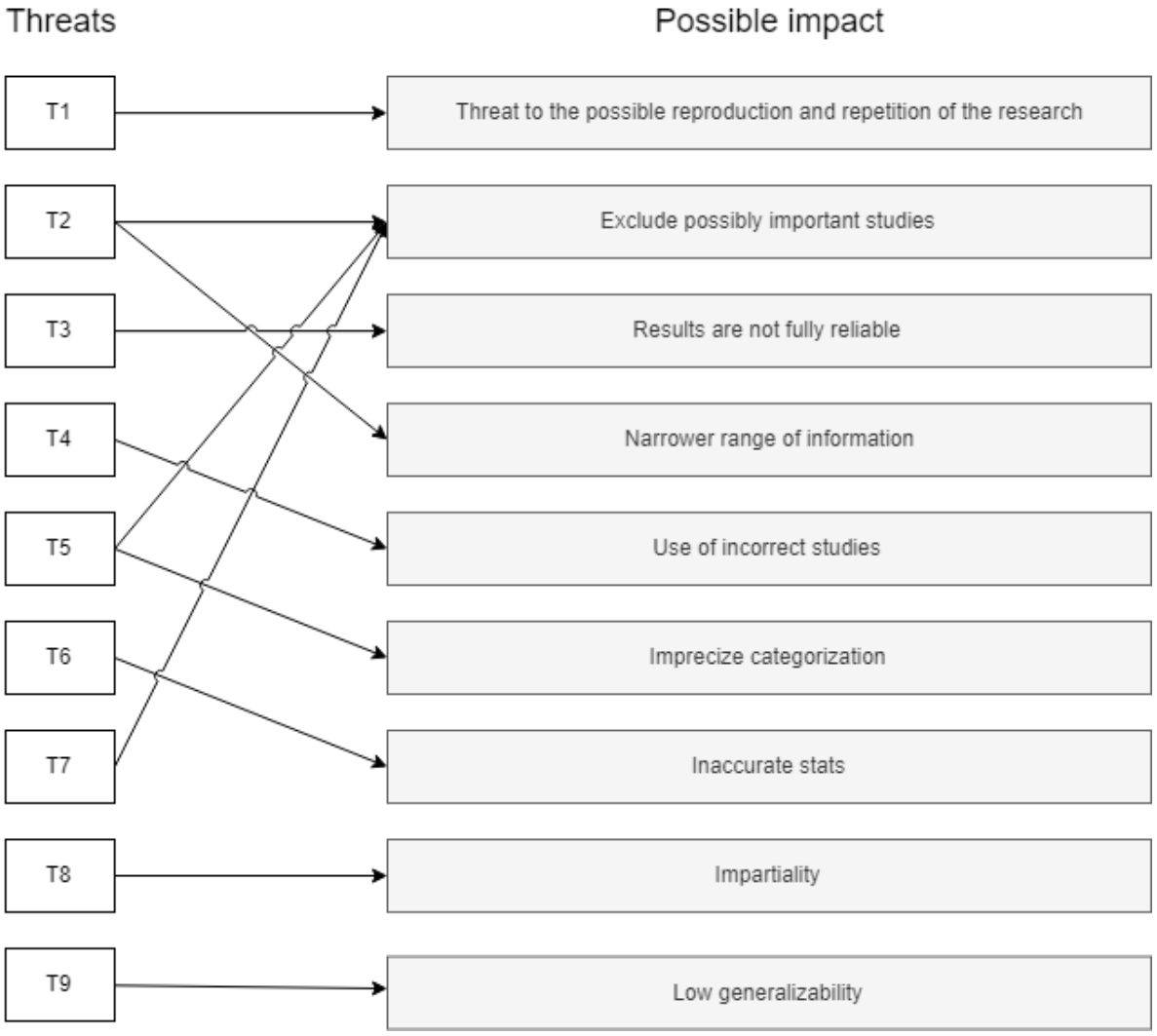


Figure 2: Possible impact of threats to validity (model by Zhou et al. (2016) and adapted by the author)

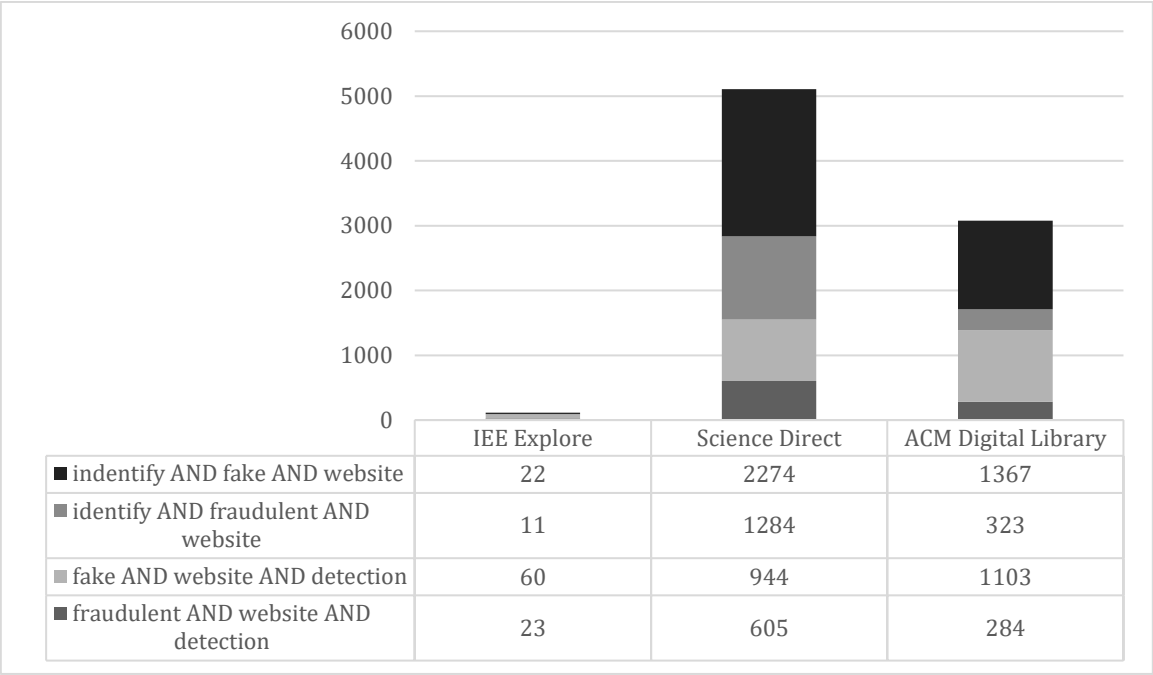
4.2 Practical Procedures

This chapter will explain how the guidelines already shown in the previous chapter (4.1 Systematic Literature Review) have been followed in practice. The chapter will focus on two of the main phases of the project, the article selection procedure, and the article analysis phase. It is important to be transparent about the process and to map out how these two phases have been carried out for possible replication of the studies and full understanding of the project and its results.

4.2.1 Article Selection Practical Procedure

Article selection is one of the critical phases of the process as it lays the foundation for the study. This chapter and its subchapters detail how the guidelines given in the chapters on databases (4.1.1), search terms (4.1.2), article selection criteria (4.1.3), and backward snowballing (4.1.4) have been put into practice. It should also be noted that all the processes described in this section have not been handled manually, for this purpose the Zotero tool has been used. Zotero is a free and open-source reference management software which can be used to import articles, manage the citations or to manage the bibliography. In this thesis development Zotero has been mainly used to automate citations and to manage the bibliography by automating processes such as the management of duplicates or the download of imported articles.

The article selection process begins with the searches for articles in the databases, when these were carried out the first basic criteria for inclusion and exclusion were applied. The criteria applied were those which the databases allowed to be entered when carrying out the searches, these are for example the date of publication, the language or free access with university accounts to the document. The following figure (Figure 3) graphically represents the number of articles that were accepted during this first phase of the selection process divided by database and search term.



	IEE Explore	Science Direct	ACM Digital Library
indentify AND fake AND website	22	2274	1367
identify AND fraudulent AND website	11	1284	323
fake AND website AND detection	60	944	1103
fraudulent AND website AND detection	23	605	284

Figure 3: Results of searches in databases (author's own)

As can be seen in the image Science Direct and ACM Digital Library have contributed a much larger share of studies to the project, yet the results of IEE Explore have turned out, in percentage of accepted studies in relation to the overall results of a database, to be more relevant to the project than those of the other two databases. After this first step duplicate studies are eliminated, resulting in the following number of articles to proceed with the exclusion of articles using the remaining criteria:

- IEE Explore: 83 articles.
- Science Direct: 3317 articles.
- ACM Digital Library: 1618 articles.

After applying the rest of the criteria, a total of 30 relevant articles were obtained. This may not seem much in relation to the total number of articles, but it should be noted that most of the articles were discarded because they were not related to the study in any way. In this phase, articles were eliminated starting first with the title, then the abstract was analyzed, and finally the articles that could not be accepted or discarded in relation to the title or the abstract were subjected to further analysis.

The number of discarded articles can be considered very high and therefore it can be considered that the search terms could have been improved. Despite this it was decided to continue with the initial search terms as it was preferred to have to discard more articles rather than consider the possibility of losing any relevant articles. Furthermore, it should be noted that the discarding of non-relevant articles is generally a simpler task than finding new articles that can be considered relevant.

The last step in item selection is backward snowballing. The bibliographies of the accepted articles have been analyzed, and the year and title filters have been applied to these studies first, resulting in a total of 41 articles. After applying the rest of the criteria, a total of 10 items are extracted from the snowballing process. In conclusion, it has been compiled a bibliography of 40 articles. The resulting articles have met all the criteria and are considered as potentially significant for the research and a list of them can be found on Appendix A. In the following chart (Figure 4) the studies contributed by each database and by the snowballing to the research can be seen graphically.

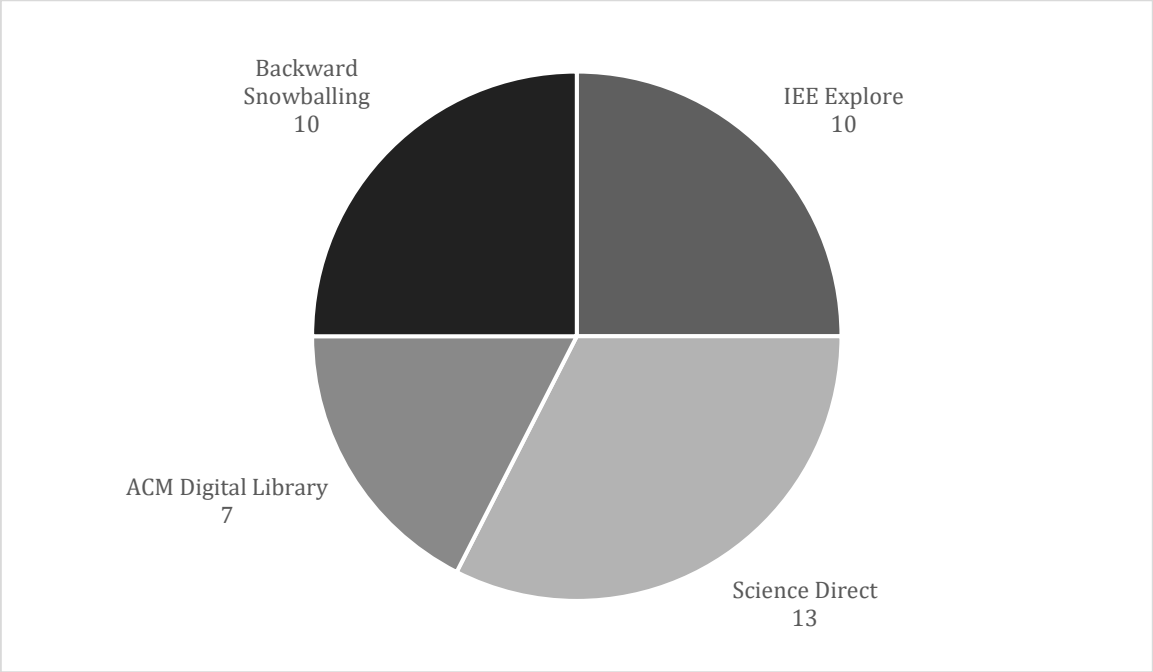


Figure 4: Studies provided by each database and by the backward snowballing (author's own)

4.2.2 Analysis Practical Procedure

The subsequent section details the implementation of the analysis process presented in the chapter on the method of analysis (4.1.5 Method of Analysis). The guidelines of Braun & Clarke (2006) have been used in the development.

During the first process of the phase, the familiarization with the data, several readings of the articles were made with the objective of finding patterns, definitions, and other relevant aspects of the article. The first of these readings was a quick read to create a mind map of the document. After this, some more in-depth readings were done in which some codes were generated, code generation which would be continued in the second phase.

During the second phase, generation of initial codes, new codes continue to be created concerning the matter under analysis. For the coding process the free tool QDA Miner

Lite, a qualitative data analysis software was used. The tool facilitates the creation of codes and their subsequent management as it allows actions such as merging codes, deleting or refactoring them.

After that, categories are created from the extracted codes by sorting the codes into themes to which they may potentially belong such as the length of the URL within the search bar features. One theme called “miscellaneous” was also created to fit all the codes that do not fit the main themes as Braun & Clarke (2006) point out. Later themes are refined and some of them are deleted if there is not enough data to support them, Baidu index for example was deleted as the information regarding it was insufficient. Finally, the generated themes and codes are refined and are renamed to their final name. During this whole process of coding and refining, thematic maps of the themes and codes were generated at each step to graphically represent the relationship between the found data. The latest version of the thematic map can be seen in the figure below.

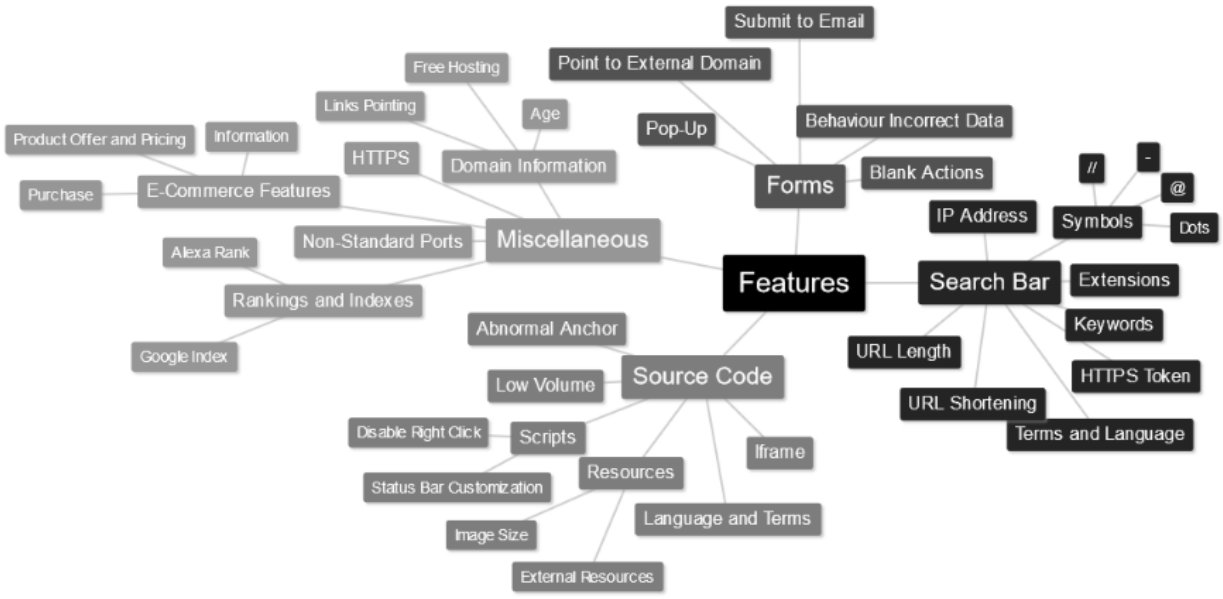


Figure 5: Final thematic map (author’s own)

5 Analysis

In the following section the analysis of the accepted articles is done following the practices specified in the chapter on analysis (4.1.5 Method of Analysis). For this purpose, the accepted articles are examined, and the most relevant features and text fragments are extracted from them. Some of the features cited in some articles will not appear in the chapter, this is due to the fact that the information about these features did not meet some minimums of relevance to the problem being addressed.

Although the number of articles will not be relevant when categorizing a feature as relevant, it is important to note that the more articles that talk about a specific feature, the more information and comments about this feature are likely to be available. It should also be considered that the more detailed the analysis of a certain feature, the more complete it will be its analysis. In several sections of this chapter it is mentioned the number of articles that develop on a feature. This is not done with the intention of drawing conclusions about features but with the intention of putting in context how accepted the use of a feature can be among scholars.

Finally, the use of labels to cite articles has been employed instead of the usual citations in order to simplify the reading and comprehension of the sections. In appendix A is available the table of labels in which each individual label is assigned to an article and the title, authors, and year of publication of each study are specified.

5.1 Search Bar Features

In the subsequent chapter the features related to the search bar will be analysed. In this group are the features related to the website address, for example.

5.1.1 Extensions

Including extensions in the URL and active download links has been assessed by A13 and A15 studies. According to A13, a ".exe" extension in the URL of the web page means that the site may try to run malware in the background. A15, on the other hand, warns about the danger of pages with active download links with extensions such as ".aaa", ".abc", ".exx" or other extensions formed by random characters. According to A15, sites with such extensions will likely try to perpetrate crime-ware based attacks.

5.1.2 IP Address

When surfing the web, that a page has the IP address of the machine hosting the web instead of the domain name in the URL most likely means that one is visiting a malicious website as this is an unlikely possibility on legitimate sites. More than half of studies from the bibliography agree on classifying this feature as important in detecting fraudulent pages since it is used in their research for this purpose. Some

selected articles such as A3 and A8 even use this feature in rule-based systems as one of the decisive factors. It is such a common characteristic of malicious pages that even well-known applications like WhatsApp do not allow URL hyperlinks if they use IP addresses according to A17.

Regarding the usefulness of using an IP address instead of the usual domain name, some articles such as A6, A13, A15 and A27 explain it as a mechanism to hide domain information from the user. A27 also values the option that it is because of not wanting to spend money on buying a domain for a fake website since the life span of this type of websites is more limited. A17, on the other hand, sees the use of the IP address as a method to bypass DNS based URL-blacklist engines.

A15 and A17 studies warn about the possibility of fraudulent pages using hexadecimal characters to represent the digits of the IP address so that they can avoid some detection systems. Usually the hexadecimal characters of URLs start with the "%" symbol according to A15.

5.1.3 Use of Symbols

The use of symbols in the address of a page can also be used for the detection of malicious pages. In the following sections the information collected about the most relevant symbols is further detailed and analysed.

5.1.3.1 Use of "-" Symbol

The use of a symbol such as the hyphen is very rare in legitimate sites and websites with addresses that use them are regarded as a potential hazard by 19 of the selected studies of the literature.

Main vision about the usage of the hyphen is that criminals try to mistake users by adding prefixes or suffixes that might trick them to believe the site is legitimate. The previous vision is shared by studies such as A3, A13, A18, A28 and A33. Another hypothesis is presented by A6 in which it is stated that hyphen is used to mistake users because of its similarity with the underscore ("_"), a common character in legitimate sites.

The last view about the issue is presented by A12 in which it is stated that cyber criminals try to mistake users by taking advantage of internationalized domain names (IDN), domains that contain labels displayed in language specific alphabets. A12 exemplifies this theory with "xn--facbook-ts4c.com" address that is displayed as "facebook.com" in the search bar.

5.1.3.2 Use of "//" Symbol

Using double "/" symbol is normal in legitimate sites after HTTP (Hypertext Transfer Protocol) or HTTPS (HyperText Transfer Protocol Secure) protocols but having them in an unusual place is an indicator that a page may be fraudulent according to studies such as A3, A7, A8, A13, A15, A20, A22, A28, A39, and A40.

Multiple studies of the literature (A7, A13, and A40) agree on that the symbol is used to redirect users to web sites without the user being aware of this. They achieve this as the part of the URL before the symbols is ignored. To detect these symbols in unusual places, two methods are presented by the selected research. Studies such as A3 and A8 examine that the symbols are not in a place after the seventh character, a position in which they would be in a page that has the HTTPS protocol. To label the pages as fraudulent, A15 and A28 keep track of how many times the symbols appear and if they are found more than once the website is classified as potentially malicious.

5.1.3.3 Use of "@" Symbol

The usage of @ symbol is also regarded as an indicative of website fraud by 20 of the 40 studies of the bibliography. Research that analysed the reasoning of the usage of this symbol (A3, A6, A13, A15, A16, A33, and A35) unanimously state that it is used to redirect the user inadvertently to malicious pages. This is achieved because the browser ignores everything that precedes the @ symbol. In this way, cybercriminals write the address of a legitimate web page, the @ symbol and then the actual address of the malicious page they want to redirect the user to.

5.1.3.4 Use of Dots

Dots are used in all legitimate URLs but the excessive use of them may show a site is fraudulent, 23 of the selected 40 research use this feature as a potential indicative of malicious sites. A13 states that malicious sites often count with more domains than legitimate ones. Having multiple domains is reasoned as wanting to confuse the user into believing that the page is legitimate and secure when it is not.

Studies such as A13, A20, A25, A26, A27, A28, A31, and A33 agree on counting the number of dots and using this value as a feature to detect malicious sites. However, none of the mentioned studies indicate a threshold value to classify the number of dots as excessive as most of them use this feature as a numerical value in artificial intelligence-based detection methods.

5.1.4 Terms and Language

Words or word alterations used in page addresses can be a sign that a page is malicious. The most relevant features related to the language in the URLs are detailed below.

5.1.4.1 HTTPS Token in URL

Fake sites try to mimic legitimate sites intending to deceive users to think they are legitimate. One trick these websites use is including protocols such as HTTPS in non-protocol parts of the URL. This is done intending to make users think a protocol they trust is included on the website when that is not the case. Using these tricks is valued by studies A8, A14, A20, A22, and A39 and is assessed as a suspicious practice by them.

5.1.4.2 Keywords

Use of certain words has been also assessed by 15 of the 40 selected research. A10, A13, A15, A16, A20, A24, A25, A26, A28, and A36 studies agree on stating that the use of certain keywords on URLs such as brand names, legitimate sites names or sensitive words such as banking or login is a feature to analyse. Using these words is regarded as a SEO (Search Engine Optimization) technique by research such as A10. Other studies such as A14, A17, and A26 add to the equation the typosquated words i.e. words written incorrectly or with characters exchanged for similar ones. Typosquated words try to take advantage of unsuspecting users who may not realize that the word they are reading does not correspond to the genuine one.

5.1.5 Use of URL Shortening Services

URL shortening services are the ones which make URLs friendlier to users by shortening them but still redirecting to same page. Usage of these services has been regarded as a potential hazard by research such as A5, A15, A17, A18, A22, A37, and A40. Although it is not the common direction that has been taken in studies, A5 for example uses this feature in a rule-based system.

The most common belief among studies is shared by A15, that this type of service is used to hide the proper domain and thus confuse users. As an alternative point of view, A17 states that the aim of the usage of shortening services is to trick detection tools as different hash codes are created from the URLs generated by the services.

5.1.6 Length of URL

Unusually long URLs may also be an important factor on malicious webpage detection. This opinion is shared by 23 of the selected 40 articles as they use this feature in their research. The reasoning of having long URLs on malicious sites seems to be the attempt to hide suspicious parts of the URL from the users in order to confuse them and make them believe that they are on a legitimate and secure page. The above opinion is shared for example by the investigations A3, A5, A6, A27, A28, and A30.

Regarding where the limit is set to classify a URL as too long and suspicious, there is a difference of opinion. The most accepted and shared view seems to put the limit on 54 characters, this opinion is shared by studies A3, A5, and A28 for example. Although A5 shares the opinion of categorizing URLs of less than 54 characters as safe it also does not classify the ones that surpass this threshold as malicious but contemplates a range between 54 and 75 in which sites are classified as suspicious. A15, on the other hand, sets the limit to the 35 characters. Alternatively, and unlike previous studies, A27 recommends dividing the URL into different sections and calculating the threshold for each of these separately.

5.2 Forms

Into this subcategory fall the features related to forms such as login information. The following one is a critical category as it is via forms the way phishers or other cyber criminals get the sensitive information through the users most of the times.

5.2.1 Blank Action

When filling a form, provided information should be treated in a way or another by the site. If there is no action performed with the information given, it may be a sign that the page is fraudulent. Doing nothing with the user data is regarded as an indicator that a site is fraudulent by several articles of the bibliography such as A6, A8, A20 and A22.

Regarding the detection of this feature, A6 and A22 agree that it should be checked if the action attribute of the form has an empty string or an “about:blank”. A20 states that inclusion of “#” or “javascript:true” should be checked besides the two attributes mentioned above.

5.2.2 Form submits to Email

This feature has been considered by many of the selected studies (A6, A8, A20, A22, and A40). According to studies such as A6 forms should only be uploaded to secure and legitimate servers as they handle personal information most times. The studies agree that in case the form sends the information to an email it is likely that it is a malicious page and that the creator of the form is trying to send the information to his own email.

Forms can be submitted to emails in two different ways according to A8, the use of “mailto:” function or the “mail()” PHP function. The use of “mailto:” can be seen in the following example in which the data is sent to the fraudulent@email.com address.

```
        <form action="mailto:fradulent@email.com" method="get"
enctype="text/plain">
```

5.2.3 Behaviour when Incorrect Login Information is Introduced

The behaviour of the site when inputting incorrect login information has been regarded as a critical feature by A1 and A9. A1 claims that this feature is the biggest weakness of the fake pages for the difference in behaviour of these pages compared to the authentic ones. In contrast, and although the feature is used in A9 study, it is stated that a percentage of the fraudulent pages imitate the behaviour of the real pages and that consequently it is not effective every time.

According to A1, fake sites accept all types of login inputs regardless of their correctness. Legitimate sites, on the other hand, show an incorrect password or username message or they redirect the user to another login page which often include the need of additional inputs such as CAPTCHA tests. It is also declared that URL should change when invalid data is introduced. With that in mind, A1 proposes the comparison of the hash values of the original page and the redirected one as the values should differ in legit sites.

A9 states that the new redirected site should not include a password field in fake sites as they accept fake login information. A9 also warns about the possibility that the fake page will redirect to its authentic counterpart and therefore recommends checking that the initial page and the redirected page are in the same domain.

5.2.4 Point to External Domain

Legitimate sites rarely rely on external domains to handle the data, and having the form point to an external domain might mean that the webpage is fraudulent. Articles such as A8, A15, A18, A20 and A34 agree on the fact that forms pointing to addresses outside the domain might be an indicator of fake sites.

Although it can be useful to detect fake sites most of the articles use this feature besides other features on artificial intelligence systems as it may not be enough to judge a site just with this feature. Also, according to A8 this feature on its own is not able of tipping the scales against a website but is used to label it as suspicious.

5.2.5 Use of Pop-Up Windows

Various studies (A8, A15, A18, A20, A21, A30, A32, A33, and A40) of the bibliography select the use of pop-up windows to gather information from the user as a feature when trying to detect fake sites. While this is a feature noted in several studies, some of these (A8, A18 and A33) are cautious and report that the use of pop-up forms is not exclusive of fake sites and that although it is rare some legitimate sites use them.

As far as motivation is concerned, A15 is the only study of the bibliography that addresses it. It claims that the use of pop-up windows is due to the need to avoid

discrepancies between the domain of the web page and the active form field address which may be a detail that gives away the page as fraudulent as it can be seen on chapter 5.2.4 (5.2.4 Point to External Domain).

5.3 Source Code

Source code analysis is one of the most discussed sections in the accepted literature. In the following sections the most relevant source code related features of the literature will be detailed.

5.3.1 Abnormal Anchor

Anchor links, <a> tag in HTML, are links that allow the user to navigate within the same website, making it easier to browse. Several articles of the bibliography (A6, A8, A9, A14, A16, A18, A19, A20, A21, A22, A30, A32, A33, A35, and A40) value this feature as potentially relevant in detecting fake pages since it is used in the developed tools or conducted studies.

The selected articles highlight two main abnormalities in the anchor link, that it is blank or that they point to other domains. Some studies (A6, A14, and A32) conclude that having blank links is conclusive proof to label a page as fraudulent. Unlike previous studies, A19 only takes as convincing to have blank links in the footer of the website. It bases this result on experiments in which no blank anchors were found in the footers of the web pages but were found instead in other parts of the page such as the website logo. Other studies such as A20 and A30 accept the option that legitimate pages may have blank links and propose the use of a ratio between blank and non-blank links when detecting malicious sites.

The studies value a wide range of values in the "href" HTML attribute which cause the link to be considered blank, these can be summarized in the following three points:

- Values starting with "#" such as "#skip", "#content", and "#" symbol on its own
- Empty strings
- "JavaScript:void(0)"

In addition to this, A9 claims that several malicious pages try to bypass detection methods based on blank anchors by putting the address of the page itself instead of a null link.

When analysing the reasoning of blank anchors A35 and A19 agree that this is because sites want the user to remain on the same page as long as possible until he or she enters the personal data the site is looking for. Besides that, A32 values another option about the motivation, according to it fraudulent sites have blank anchors to

pretend they have lots of hyperlinks, and therefore, be as similar as possible to a legitimate page.

While A6 and A14 are the only one studies which claims anchors are not supposed to point to another domain. Other articles (A20 and A35), although they accept it as an abnormal feature, propose the use of a ratio of anchors pointing to the same domain to anchors pointing to the domain out of the site. Similarly, A8, A20, and A30 studies use the ratio of abnormal to no abnormal anchors. In these, they count anchors with either of the above two abnormalities (blank anchors and anchors pointing to external domains) as abnormal.

A16 states that discrepancies between the value of the href attribute and the URL showed may be a good indicator of malicious sites. An example of the discrepancy can be seen below, in this example the malicious site is trying to confuse the user by making him believe that he will go to the legitimate address "https://www.legitimate.com" while in reality he will be redirected to "http://www.fake.com".

```
<a href= "http://www.fake.com" > https://www.legitimate.com </a>
```

Lastly, it applies to note the warning from A6. According to this, criminals can encode the alphabets with their corresponding ASCII code. In this way criminals can bypass systems that may look for specific URLs from blacklists.

5.3.2 Low Content Volume

The volume of content is another feature to consider when detecting fake sites and is assessed by studies A2, A9, A12, A20, A35 and A38. As far as reasoning is concerned, A39 states that malicious pages often have an inconsiderable amount of text to avoid text-based detectors.

A2 and A20 agree on that that malicious sites often have low volumes of content and mostly consist of images. Images which according to A2 will try to attract users to submit information. Furthermore, A35 and A12 state that some fraudulent websites may try to trick users by embedding a screenshot of the legitimate sites they are trying to mimic. According to A35 these web pages will also have several input fields following the design of the original sites so that users are tricked to enter their personal information.

To detect this feature A2 uses the ratio of the volume of text in links to the total volume of the text in the page. Alternatively, in A20 it is checked if there is no text at all and the page consists only of images. A9 and A35 agree on checking the absence of anchors in the body of the site, a characteristic that is very unusual in legitimate web sites.

5.3.3 Iframe Tag

The HTML iframe tag allows to create an inline frame which can be used to insert documents or other websites inside a web page. This tag is studied by several of the selected research (A2, A4, A5, A8, A9, A18, A20, A22, A38, and A40) in which it is considered as a significant feature for detecting malicious sites.

A9 concludes that the adoption of this tag is due to wanting to circumvent the source-code based detection systems as the content within the iframe is hidden from the source code of the site.

As far as the use of this tag is concerned, there are different comments in the literature. A2 states that this tag is used to embed videos or popups to show advertisements in survey scams. A4 declares that it is notable the use of this tag on drive-by downloads and on clickjacking attacks. A9 and A18 agree on saying that it is often used to display alternative websites to mistake users. According to A18 the frame border is eliminated to maximize the impression of legitimacy of the site.

5.3.4 Scripts

The use of scripts is one of the methods that allow a wider range of options for criminals to deceive users. In the following sections the two most relevant methods in the accepted articles are analysed, disabling the right click and changing the status bar.

5.3.4.1 Disabling Right Click

Disabling the right click is seen as a threat and a useful feature for detecting malicious pages by several studies (A5, A8, A18, A20, A21, A22, and A30) of the literature. According to A18 this is done intending to hide malicious activities from the users. Disabling the right click causes the user to be unable to open the context menu and therefore incapable to perform actions such as inspecting the page source.

This operation is done through JavaScript commands according to A20. A22 extends this by stating that you should look for the command “event.button==2” in the source code to detect these practices. This command allows cybercriminals to trigger events when the secondary button, usually the right one, is pressed.

5.3.4.2 Status Bar Customization

Customizing the address bar is a straightforward attempt to trick users. This is used for the detection of malicious pages in research such as A5, A8, A18, A20, A21, A22, A20, A33, and A40.

A8, A18, A20 and A22 studies agree on stating that JavaScript is used to show different URLs to the genuine one to mistake users and make them believe they are in a legitimate site. To achieve this effect, there is also an agreement on stating that "OnMouseOver" JavaScript event is used. This event is triggered once the mouse pointer is moved onto a certain element, when this happens the cybercriminal can execute actions that can for example change the displayed URL.

5.3.5 Language and Term Usage

The usage of language can also be a feature to keep in mind when detecting malicious sites. Studies such as A2, A10, A12, A21, A24, and A38 use different language related features with this aim.

The most repeated tendency in these studies is the use of certain keywords or sequence of terms, articles such as A2, A10, A24, and A38 use them as detection features to mark potentially malicious sites. According to A10 and A24 words or terms such as "guaranteed satisfaction", brand names, "replica" or "copy" are used to improve the positioning of pages in search engines.

Another feature that may be relevant is the one highlighted by A12, the text obfuscation. String obfuscation consists of hiding the above-mentioned terms and sequences in the HTML code. This is done with the objective of avoiding text-matching detection mechanisms. Text obfuscation works by swapping characters of words for characters that look similar such as the upper case "I" for the lower case "L".

5.3.6 Resources

In the coming sections, the use of external resources and the size of the image will be analysed in relation to the detection of malicious pages. These are features related to resources, one of the most basic aspects of web pages and some characteristics that can be very useful in detecting fake pages.

5.3.6.1 Use of External Resources

Resources like images, videos, or scripts for example are one fundamental part of the websites and differences between legitimate and fake sites arise in relation to them.

Various studies (A8, A18, A20, A22, A32, A35, A36, A38, and A40) assess the use of external resources in their research as it is uncommon for legitimate sites to rely on external domains.

A8, A18, A22 and A40 emphasize that the use of the external favicon is an exposure of fraudulent pages, since it is more rarely loaded from external domains compared to other types of resources.

Regarding other type of resources, the most proposed method is using a ratio of resources loaded from domain in the address bar to resources loaded from external domains as it is done by A8, A18, and A20. Alternatively, A22 and A32 use of a count of external resources which is later compared to a threshold. A35 proposes to take the most repeated domain in the source code and then compare it with the domain of the web page to decide accordingly.

5.3.6.2 Image Size

The quality of images is another feature to check according to A2, A18, and A24. According to these studies the images used in fake sites are smaller and consequently of poorer quality compared to the legitimate sites. A2 states the smaller size of images is a result of most of the pictures being in its majority logos. In the research they extracted the value of 2kB as the average image size on fake sites.

5.4 Miscellaneous

In the following sections features that do not fall into the previous groups are going to be analysed.

5.4.1 HyperText Transfer Protocol Secure (HTTPS)

Having a protocol such as HTTPS that ensures the security of communications often indicates to users that they can trust a website and confirm the identity of the site. 18 of the selected 40 studies use the inclusion or the absence of this protocol as a feature when trying to detect malicious sites.

Although this protocol may be a clear indicative of a page being secure A6, A14, A15, A18, A22, and A27 advise that this certificate can be self-assigned or assigned by untrustworthy third parties. With that in mind, these articles recommend checking if the certificate has been assigned by trustworthy sources such as GeoTrust or Verisign. Studies as A18 and A22 recommend checking how long the certificate has been issued as well.

5.4.2 Domain Information

Domain information is information that can be very useful in detecting malicious pages since several studies in the literature use this information for that purpose. Aspects such as the lack of information about the owner of the website or information about DNS records are very useful for detection. In the selected studies there seems to be unanimity in using the WHOIS protocol to extract this information. The subsequent sections present in more detail the most relevant characteristics.

5.4.2.1 Domain Age

The time a domain has been registered is used as a feature for detection by 16 studies of the literature. Articles such as A6 and A10 agree on stating that malicious websites tend to have shorter lives as they are continuously suspended. Although there is an agreement on the fact that fraudulent sites live less than the legitimate ones there is not such an agreement on the threshold value to select to categorize a site as fake. While studies such as A5 put the limit on 6 months of life other investigations such as the one of A13 put the line at one year of age.

Some studies like A6 and A25 also recommend observing the time it is left before the domain expires apart from its age. According to A6, legitimate sites usually pay several years in advance while fake ones, due to their ephemeral character, pay for shorter periods of time.

5.4.2.2 Free Hosting

Using a free hosting service is another feature that can trigger alarms and studies such as A13, A14, and A22 use this characteristic to find malicious sites. According to A14 mature providers act rapidly against malicious sites, this means that cybercriminals need to look for less proven hosting services. A13 states that criminals as phishers use free hosting services like “000webhost.com” as any user can create a website easily after registering.

5.4.2.3 Links Pointing to Website

The number of links pointing to a page may also be relevant when detecting fraudulent sites, according to studies such as A6 and A33 the more links pointing to a site the most likely it is to be relevant. This feature has been assessed by A6, A8, A18, A22, A23, A25, A33, and A40. A6 supports his theory with experimentation, more than half of the fake pages in its dataset had no link pointing to them. Although no other study highlights this issue, A23 claims that this feature is especially notable in fake online shops.

5.4.3 Use of Non-Standard Ports

When trying to detect fake sites the deviation from the rules and standards can be a warning sign. Thus, used ports are no exception and studies such as A8, A13, A15, A17, A18, A22, and A40 assess the impact of this feature on detection of fake sites. The ports that the studies analysed inspected for deviations from the standard were ports 80 and 443 which belong to the HTTP and HTTPS services, respectively.

A17 states that malicious sites use specific ports to bypass blocklists, these ports can be modified periodically to avoid the previously mentioned blocklists. A18 also recommends looking at the ports that are open as having numerous open ports means that criminals can run practically all services. This means that information of the user can be seriously threatened.

5.4.4 Rankings and Indexes

Sometimes it can be convenient to rely on third party tools. In the following chapters for example the features related to third party rankings and indexes will be analysed.

5.4.4.1 Alexa Rank

Legitimate sites are much more popular compared to fake ones; this favours the usage of rankings based on the popularity of the pages to detect fake pages. Studies such as A6, A9, A11, A22, A25, A26, A29, A36, A37, A38, and A39 use the Alexa ranking and scores to distinguish between legitimate and fraudulent sites.

Alexa ranking assigns a numerical value to web pages based on the number of user visits, page views and a data history for the last three months. The smaller the value assigned, the more popular the page is and therefore, according to A6, the more likely it is to be legitimate as malicious sites rarely have an Alexa ranking or have it with very high values. Studies such as A26, A29 and A37 agree that we could categorize a page as legitimate if it were in the top 1 million of Alexa ranking.

5.4.4.2 Google Index

Being indexed by Google to appear on the search engine is another third-party feature that may be useful on fake site detection according to A5, A7, A8, A18, A22, A25, A39, and A40. A39 states that due to the brief life period of fake sites they are often not indexed by Google, in this way not being indexed can be helpful in detecting a fake page.

5.4.5 E-Commerce Features

The features on the following sections are characteristics that unlike the previous ones are not applicable to the rest of the fraudulent pages since they are inherently related to e-commerce and can help to detect fake shops more easily.

5.4.5.1 E-Commerce Information

Legitimate web shops try to ensure the user the security of making transactions on their website, for this, they clearly offer detailed information of the company with

data such as a telephone number, physical address or links to social networks of the store. A23 and A10 studies agree on stating that the lack of this information or having false information is a clear sign of fraudulence on commerce.

A10 also draws attention to the lack of user attention on malicious pages. According to it, legitimate businesses usually have email addresses, forms, or some other means of contact for users to resolve their doubts.

5.4.5.2 Product Offer and Pricing

Fake commerce often attracts users with substantial offers and discounts. A10 states that, because of this, fake commerce often emphasizes prices and discounts on products. The disbelief that these low prices produce is stressed by the large number of products that are usually available on these pages. The catalogue of these is usually more extensive than the one of the legitimate pages according to A10 and A23. A10 notes that showed prices are usually in different currencies as, unlike in proper shops, there is just one site to serve all countries.

5.4.5.3 Product Purchase

Differences between malicious and legitimate sites can also arise during the product purchase process. According to A23, some fake shops do not count with the shopping cart function on their sites while almost every legitimate shop has them. Regarding the payment, A23 and A10 agree to affirm that malicious websites allow fewer methods of payment than authentic ones. A10 states that malicious sites usually just allow methods that permit criminals to remain anonymous such as Western Union.

6 Synthesis

In the following chapter the information extracted from the articles in the analysis chapter (5 Analysis) will be synthesized. In this section, not only will the analysis be synthesized, but also the features will be classified as relevant or irrelevant. It is important to mention that the features that are searched are the ones that have a high accuracy when it comes to detecting malicious pages, this is why only those high accuracy features will be considered as relevant.

In case the available information is not enough to classify the feature in a reliable way it will not be categorized, this will be presented in a clear way so that there is no room for uncertainty. Finally, the section will propose areas to investigate in case there are gaps or diverging voices in the features considered relevant.

6.1 Search Bar Features

The results that have been extracted in relation to the search bar features analysed are detailed below. The following table (Table 3) summarizes the classification made of the features.

Table 3: Results of search bar features

	Feature	**Classification**
	Extensions	Inconclusive
	IP Address	Relevant
Use of Symbols	Use of "-" Symbol	Relevant
	Use of "//" Symbol	Relevant
	Use of "@" Symbol	Relevant
	Use of Dots	Relevant
Terms and Language	HTTPS Token in URL	Relevant
	Keywords	Relevant
	Use of URL Shortening Services	Inconclusive
	Length of URL	Relevant

6.1.1 Extensions

Extensions on the URL of the site was the first feature that was analysed. These can be used in legitimate sites, but some studies warn us about malicious sites using them to run malware on the background (Aung & Yamana, 2019) or to automatically download files (Orunsolu et al., 2019). Although having these extensions may not be common in legitimate sites, the information provided by the studies is not enough to categorize this feature as relevant for detection. Because of the inconsiderable amount of relevant studies and the information given on those more research should be done to get conclusive results about the relevance of the feature.

6.1.2 IP Address

Although the used method is qualitative and not a quantitative one, it seems relevant to note that this feature has been used in a large part of the selected studies. This feature is seen as a hazard as they regard it as a mechanism of mistaking users by articles such as Gajera et al. (2019) or a blacklist bypassing mechanism by Silva et al. (2020). It is also important to note the possibility of the IP being written in hexadecimal code as pointed out by Orunsolu et al. (2019) and Silva et al. (2020).

Studies such as the ones by Ahmed & Abdullah (2016) and Rajab (2018) are confident enough to include this feature in rule-based systems and other artificial intelligence-based studies as Zhang et al. (2016) classify this feature as critical for detection. Having the mentioned facts into account, this factor should be acknowledged as a good detection feature.

6.1.3 Use of Symbols

In the following sections are synthesized and evaluated the features related to the use of the symbols in the URL address.

6.1.3.1 Use of "-" Symbol

The usage of "-" symbol to mistake users is accepted by the articles that deal with the subject, even if there is a disparity of opinion about how the user is being misled. Studies such as Ahmed & Abdullah (2016) think that criminals try to mistake user by adding prefixes and suffixes. Gajera et al. (2019) on the other hand thinks criminals try to take advantage of the similarity of the hyphen with the underscore, a common character in real web pages. Tian et al. (2018) are of the opinion that criminals try to use IDNs (Internationalized Domain Name) to mislead users. Although most articles categorize it as relevant, Zhang et al. (2016) characterize the concerning feature as non-critical. All in all, and despite some conclusions to the contrary, it is considered that the information background is sufficient, and the use of this feature is believed to be relevant in detecting malicious pages.

6.1.3.2 Use of "//" Symbol

The selected studies overwhelmingly affirm that the use of the symbol "//" in an unusual site means that a page is fraudulent since it tries to redirect the user without his or her knowledge. Considering the information analysed and the amount of studies that agree with this, it can be concluded that it is a relevant feature in the detection of fake sites.

6.1.3.3 Use of "@" Symbol

Using the @ symbol to redirect the user without him knowing it is considered as relevant by most of the selected literature. It is raised as a discordant voice the study by Zhang et al. (2016) in which because of its experiments this feature has been categorized as non-critical. In a similar way to the section related to the use of the "-" symbol (6.1.3.1 Use of "-" Symbol), it is considered that studies considering this feature as critical provide the necessary background to contradict those who do not. In this way, it is concluded that this feature is relevant in the detection of fake web pages.

6.1.3.4 Use of Dots

When it comes to the use of dots, there is a consensus among the selected studies. As an example, it is stated by Aung & Yamana (2019) that malicious sites "use more subdomains than legitimate websites, or unnecessary "." symbols in the URL path". It is considered a relevant feature for detection considering all the information provided by the selected studies and their conclusions, e.g. Zhang et al. (2016) concludes that this is a critical feature. When working with this feature, most studies have chosen to use it in artificial intelligence systems or to calculate a threshold in rule-base systems.

6.1.4 Terms and Language

The analysis regarding the term usage in the web address will be synthesized and evaluated below.

6.1.4.1 HTTPS Token in URL

Regarding the inclusion of the HTTPS token in the URL from a logical point of view it seems a clear attempt to deceive users. Moreover, the hypothesis that it is a clear symptom of a fraudulent website is well supported and the selected studies give sufficient information about it. In conclusion, the use of this feature is considered relevant in detecting fake pages.

6.1.4.2 Keywords

Use of certain words or alteration of words in the address of sites can indicate that a website is not legitimate. Fake sites tend to use certain words or brand names to mistake users and to optimize their positioning in search engines. All in all, it is concluded that the use of keywords is a relevant feature when detecting fake sites although it is advised that more research should be done to get a reliable word list.

6.1.5 Use of URL Shortening Services

Studies that analyse the use of shortening services agree on stating that malicious sites tend to use them as a way of hiding information from the user (Orunsolu et al., 2019) or a way to trick detection tools (Silva et al., 2020). Although the information provided seems to be relevant it seems to be too little to categorize this feature. In conclusion, a proper assessment cannot be made as there is not enough information available, so more research is needed to decide.

6.1.6 Length of URL

URLs that are longer than usual may be a method to hide information about the site to the users. Articles that do analyse the utility of this feature for criminals unanimously agree with the previous hypothesis as it is one of the most discussed and supported theories in the literature. Although there is agreement that it is a feature to be considered this does not happen when setting the limit to classify a URL as too long or when selecting how to measure the URL. Finally, it is considered that the information provided is sufficient to consider this feature important. Yet it is believed that the feature should be further investigated to make the most appropriate decision on all points of discussion.

6.2 Forms

In the following sections, the results obtained from the analysis of the features related to the forms will be synthesized and evaluated. Table 4 shows the classification made of the features depending on the resulting relevance

Table 4: Results of form features

Feature	**Classification**
Blank Action	Relevant
Form Submits to Email	Relevant
Behaviour When Incorrect Login Information is Introduced	Relevant
Point to External Domain	Inconclusive
Use of Pop-Up Window	Not Relevant

6.2.1 Blank Action

Not doing anything with the data provided by the user is considered as fraudulent by several articles in the bibliography. For example, Gajera et al. (2019) states that data should be sent to a legitimate server to process the information. Apart from this, from an external point of view it seems a clear attempt to confuse users by making them believe that these data provided have some utility or purpose. It is considered that the

information provided is sufficient to categorize the use of this feature as suspicious and therefore useful for the detection of malicious pages.

6.2.2 Form Submits to Email

Similarly to the previous section (6.2.1 Blank Action), the sending of data to an email is not considered a secure treatment of the user data. Studies on the subject that discuss the hazards of the feature unanimously affirm that the probabilities that the page is malicious and that the author is trying to steal information from the user are very high. For example, Gajera et al. (2019) and Adebowale et al. (2019) agree on stating that information should be uploaded to a server to perform with it the corresponding actions. Considering the data and the unanimity among the studies, this feature is considered relevant in the detection of web pages of a malicious nature.

6.2.3 Behaviour when Incorrect Login Information is Introduced

Articles discussing this feature regard it as crucial and as the biggest weakness of malicious sites (Ndibwile et al., 2017). Although Rao & Pais (2017) state that malicious sites are able to copy the legitimate pages in their behaviour avoiding this kind of abnormalities, they still regard the feature as critical as sites with these irregularities are still very likely to be malicious.

In this case, are classified as abnormal behaviours redirecting to different domains or accepting incorrect data. It is concluded that although not many articles deal with the subject, those that do, provide enough background and are confident enough to classify this feature as important in detection.

6.2.4 Point to External Domain

Several articles of the literature discuss the hazard of forms pointing to external domains. Articles such as the one by Orunsolu et al. (2019) state that forms pointing to external sites are likely to be have malicious intentions. Other articles such as Rajab (2018) are not confident enough with this feature although they use it to categorize sites as suspicious. All in all, it seems that the information provided by the research are not complete enough and considering that among these they are not confident enough either, the usefulness of this property is considered as inconclusive. More research would be necessary to be able to affirm that this property is relevant.

6.2.5 Use of Pop-Up Window

Use of pop-up windows to show forms it is an anomaly on web sites and is assessed by a large amount of research of the literature. Although it is assessed by many articles

Rajab (2018), Adebowale et al. (2019), and Zhang et al. (2016) note that pop-ups are also used in legitimate sites and that it cannot be taken as a critical feature for detection. As a result, use of pop-ups to show forms is not regarded as an important feature as the provided background in support of the theory is not considered to be sufficient. It should also be noted that information regarding it as a non-relevant feature is more complete than the ones regarding it otherwise.

6.3 Source Code

This section will summarize the results of the analysis of the features that can be detected by looking at the source code of the web page. Below is a table (Table 5) showing the categories derived from this evaluation.

Table 5: Results of source code features

	Feature	**Classification**
	Abnormal Anchor	Relevant
	Low Content Volume	Relevant
	Iframe Tag	Relevant
Scripts	Disabling Right Click	Relevant
	Status Bar Customization	Relevant
	Language and Term Usage	Relevant
Resources	Use of External Resources	Relevant
	Image Size	Inconclusive

6.3.1 Abnormal Anchor

Anything out of the rule can be used for detection of false pages, in this case the abnormalities of the anchor are discussed by a wide range of studies in the literature. In these studies, are considered as the main abnormalities pointing to external domains, blank anchors, and disparities between the direction to be redirected and the URL shown to the user. Many articles expect abnormalities to some extent in legitimate sites like for example Rao & Ali (2015) with logos. It is concluded that the information provided is complete enough to categorize this feature as relevant. Although it is regarded as important this feature should be used with ratios or counts of abnormalities and not the mere presence of one of them. It is also relevant to note that further research should be done to calculate threshold values if needed.

6.3.2 Low Content Volume

The low content volume is analysed by numerous articles in the literature and some hypotheses arise in them. Orunsolu et al. (2019) and Rao & Pais (2019) for example state that some fake sites are mainly composed by screenshots. Although there is

enough evidence to say the concerning feature is relevant in detection further research needs to be done to draw a line between enough and too little content.

6.3.3 Iframe Tag

Several studies assess the use of the iframe tag, but there are discrepancies between the reason of using it. The most accepted hypothesis is presented by Rao & Pais (2017) and Adebowale et al. (2019), these two state that is used to display alternative sites. All in all, and despite the discrepancies between the reasoning studies agree that it is a relevant feature and there is enough evidence to state that it is a relevant feature to check.

6.3.4 Scripts

Scripts are often used to trick users into thinking they are on a legitimate website. The following is a summary of the results of the analysis of the most relevant scripts in the subject.

6.3.4.1 Disabling Right Click

Disabling the right click is a method to hide information of users or make the process of getting the information harder. This theory is extended by articles as the one by Chiew et al. (2019) in which it is indicated that JavaScript is used to achieve this effect. Having into account that articles in the bibliography support this concept and provide the information to back it up, this feature is classified as effective when detecting fraudulent pages.

6.3.4.2 Status Bar Customization

Changing the address shown in the status bar is not only a clear attempt to confuse the user but is also a feature analysed by several studies. The theory that it is a useful feature for detection is widely accepted by the selected studies and there are no apparent discrepancies with it. With this in mind, it can be concluded that this feature is potentially useful for detecting malicious pages.

6.3.5 Language and Term Usage

There is enough evidence to say that some words or terms are more likely to appear on fake sites. Some methods such as the obfuscation, presented by Tian et al. (2018), have also a clear intention to mistake users. Although it is a complex feature to detect as a reliable list of words and terms is needed it is regarded as important. Similarly to

6.1.4.2 Keywords, a reliable list is needed and further research should be done to get it.

6.3.6 Resources

Resources are one of the fundamental pillars on websites today, which is why they can be useful in detecting malicious pages. The results of the features related to the resources are summarized below.

6.3.6.1 Use of External Resources

Use of external resources is a feature checked by many studies of the bibliography as it is somewhat uncommon on legitimate sites to rely on third parties to get them. One of the strongest opinions presented in the studies is that the external favicon may be an important feature on its own. This opinion is shared by research such as Rajab (2018) or Adebowale et al. (2019). Regarding the other resources we can conclude from the studies that a threshold needs to be calculated to get reliable results as the possibility of using some external resources in legitimate sites is accepted. Although the use of ratios or counts is widely advised by the studies another alternative is presented by Rao & Pais (2019), in this study it is proposed the comparison of the most repeated source with the site domain. It is concluded that although different methods are presented to find abnormalities, the evidence presented is sufficient to categorize the use of external resources as an important feature for the project.

6.3.6.2 Image size

Image size has also been discussed by Kharraz et al. (2018), Rajab (2018), and Kim et al. (2015) studies. According to these studies, malicious pages have a smaller average image size than legitimate ones and Kharraz et al. (2018) even sets 2kB as the average size on malicious pages. The information and insights provided by these studies are judged to be too scarce, which is why the feature cannot be categorized as relevant. More studies should be conducted on the subject to draw conclusive results.

6.4 Miscellaneous

The following section will synthesize the features that could not be classified within any of the main categories above. Table 6 summarizes the categorization of these features.

Table 6: Results of rest of features

Feature	**Classification**

	Feature	Classification
HyperText Transfer Protocol Secure (HTTPS)		Relevant
Domain Information	Domain Age	Relevant
	Free Hosting	Inconclusive
Use of Non-Standard Ports		Relevant
Rankings and Indexes	Alexa Rank	Relevant
	Google Index	Inconclusive
E-Commerce Features	E-Commerce Information	Relevant
	Product Offer and Pricing	Inconclusive
	Product Purchase	Inconclusive

6.4.1 HyperText Transfer Protocol Secure (HTTPS)

Protocols such as HTTPS try to secure user data and in a way are also indicative for users to believe a website or not. Although many documents in the bibliography use this feature to know if a page is legitimate or not, some of these studies warn us about more aspects to check. Articles such as the one by Gajera et al. (2019) warn users about the probability of having self-assigned protocols, to avoid being tricked by these it is advised to check the protocol sources and only rely on sites with protocols from trustworthy sources. Adebowale et al. (2019) and AlShboul et al. (2018) extend the above and recommend checking the time taken by the assigned protocol besides the source of the protocol. Hence, this feature is considered significant in detecting malicious pages as long as the relevant aspects of the protocol are analysed.

6.4.2 Domain Information

Data concerning the domain such as the age of the domain and the use of free hosting are synthesized and categorized below.

6.4.2.1 Domain Age

The age of the domain can be useful to discover a false page since these usually have a shorter life than the legitimate pages. This is a hypothesis supported by most of the selected studies but some of these (Gajera et al., 2019) go even further and recommend also to see how much is left to expire the domain since the legitimate ones are usually paid with years of anticipation. It is concluded that the information provided by the studies is sufficient to categorize this feature as significant. Although the feature is relevant, it is also believed that a trustworthy threshold should be calculated since the studies do not reach an agreement, even so everything seems to point out that this should be located between 6 and 12 months.

6.4.2.2 Free Hosting

The use of free hosting is used to detect fake pages since according to studies such as the one by Ding et al. (2019) the most important providers act quickly eliminating the malicious pages. Although there exists an obvious correlation between more trustworthy sites and proven hosting services, there is no evidence enough to claim that free hosting is a relevant feature in identifying fake websites. More extensive work needs to be done to draw reliable conclusions.

6.4.3 Use of Non-Standard Ports

As stated above, any deviation from the norm is a data to be analysed when investigating the malicious pages and the use of ports is no exception. According to Silva et al. (2020) ports are used to bypass blocklists and Adebowale et al. (2019) warns about the possibility of criminals running multiple services when all ports are open. Although there is not a high amount of information about it, it is considered that the information provided is convincing enough to categorize this feature as important for detecting malicious pages.

6.4.4 Rankings and Indexes

Third-party ratings and classifications are also used in the detection of malicious pages, and therefore the most relevant rankings and indexes in the literature are analysed below.

6.4.4.1 Alexa Rank

Third party rankings are also used in the detection of fraudulent pages as fake pages enjoy less popularity than legitimate ones. The most relevant between all the rankings is the Alexa rank as it is used by numerous research. Studies such as the one by Gajera et al. (2019) state that the sites with lower value are more likely to be legitimate. There also seems to be an agreement on categorizing sites in the top 1 million as legitimate, this view is shared by studies such as the ones by Sahingoz et al. (2019), Kanei et al. (2019), and Arshad et al. (2017). In general, the Alexa range is considered valuable for identification, as it appears to be widely accepted and studies provide sufficient background to support the measure.

6.4.4.2 Google Index

The Google index is also used by several of the chosen studies, studies such as K.p & Damodaram (2016) claim that malicious pages are rarely indexed by Google. Although from a logical point of view it is likely that a large company like Google will not index the malicious pages, it is considered that the information provided by the studies is

too scarce and of little significance. Because of this, the results cannot be conclusive and Google index cannot be categorized as important in detecting fake pages.

6.4.5 E-Commerce Features

Finally, the analysis of the features that can be less extrapolated to other types of web pages, those related to e-commerce, will be synthesized. These features are the ones that are believed to help detect fake stores among other types of fraudulent commerce.

6.4.5.1 E-Commerce Information

Information about the shops such as telephone, email contact or social networks are analysed by Mostard et al. (2019) and Carpineto & Romano (2017). The few studies that deal with this issue agree that not having information or having false information is a clear indication of the falseness of the website. The information from these studies is assessed as convincing and complete, and it can be said that this feature is potentially important in order to detect malicious websites.

6.4.5.2 Product Offer and Pricing

Characteristics related to the greater number of products that have fraudulent pages and the price of these products are also analyzed. Although it may have indications of being real, the information provided does not allow a clear conclusion to be drawn, it is concluded that more research needs to be done to correctly categorize this feature.

6.4.5.3 Product Purchase

Differences regarding the purchase of the product such as payment methods or having or not having a shopping cart on the site are assessed by Mostard et al. (2019) and Carpineto & Romano (2017). They also do not support this theory every time as they cannot claim that every legitimate site follows the same rules. Hence, it cannot be concluded that these features are of any relevance in detecting fake pages.

7 Discussion

This chapter will discuss relevant factors in the development of the project such as the review process. Apart from this, the validity of the research results will be discussed, and the impact of the research will be evaluated from different angles.

7.1 Reviewing Process

During the development of the project, the process already presented in chapter 4 on methodology has been followed.

The first phase proved to be straightforward, although decisions had to be secured several times to lay a proper foundation for the next project stages. This stage of the process is composed of the choice of databases, terms, and criteria. With regard to the selection of databases, the recommendations of Brereton et al. (2007) have been followed, although one of the recommended databases (CiteSeerX) had to be discarded as it was not believed that the results it provided matched the search carried out. In order to find the right search terms the most basic terms were used at first, a first article search was performed in test mode and new terms or synonyms used by scholars in reference to the problem to address were found in the resulting articles. Finally, the inclusion and exclusion criteria were defined, and a first draft of criteria based on other research and guidelines was presented. This first approach was finally refined together with the supervisor in order to achieve the greatest number of articles relevant to the study.

The second phase of the project (conducting) on the other hand has been more complex than the first one, especially the analysis of the studies. The part of carrying out the searches turned out to be a direct and uncomplicated job. On the other hand, applying the selected criteria and doing the quality assessment was more complex. While some studies were simply eliminated, in others it was necessary to analyse them more thoroughly, in some the line between accepting them or eliminating them was even very thin. The most laborious part of the phase turned out to be the analysis as it included not only reading and coding but also refining the codes and selecting which parts are relevant to the project. Finally, the backward snowballing and another cycle of the conducting process were performed. The snowballing process resulted in less accepted items than expected, which is possibly due to the date range used. Many articles that would otherwise have been accepted had to be removed as the publication year criterion was not met. This criterion has still been key in the realization of the document since if the range had been extended the resulting number of articles would have been unreachable due to the number of people carrying out the project and the time available.

The last step, the reporting has been a heavy work since all the relevant aspects of the research and analysis have had to be reflected in a trustworthy form.

7.2 Validity of Results

As far as the validity of the study is concerned, certain decisions have been taken to ensure it. The first of these decisions is the method itself. Systematic literature review is regarded as an "unbiased, systematic, rigorous, and replicable" (Efron & Ravid, 2019) method as it explicitly explains the process followed and all development is guided by pre-selected guidelines. In this way, the bias to which the project could be subjected is minimized as much as possible.

To ensure the validity of the results, the classification of the features on which the information was deemed insufficient has not been carried out. These features have been left in limbo pending further research to properly classify them.

Finally, the guidelines presented in Table 2 have been followed in order to avoid the threats shown. As a conclusion, it is believed that all the dangers that could arise concerning the validity of the study have successfully been addressed.

7.3 Ethical and Research Ethical Impact

Malicious websites are an evil to which all internet users can be exposed. The results of this thesis intend to lay the foundations for the development of user protection tools. On the other hand, it has also been considered the probability that the study could be helpful for cybercriminals since they can use the list generated to directly address what could be considered their greatest weaknesses. With respect to the ethics of the study, it is concluded that the possible good that it can do is more relevant than the possible bad use that can be given to it.

Doubts may also arise regarding the ethics of research. The most relevant points in this matter are presented in chapter 4.1.6 and are extracted from the study by Suri (2020).

Firstly, an exhaustive study of the accepted articles was carried out and special care was taken to transfer the ideas and passages from the original documents so that no nuances or perspectives were lost. Second, the different criteria were also selected, so that the accepted studies would be applicable to the research and that the results of these would be extrapolated to the thesis. Finally, the transparency of the study has been ensured, not only by explaining the process followed to allow easy replication but also by trying to be as clear as possible during the analysis and conclusions of the project. In this way it has been tried to be as clear as possible presenting the different points of view, and the disparities of opinion when necessary. it is also noteworthy that documentation of the process can also be used for analysing the document and, among other things, assessing the ethical impact of the thesis.

To summarize, it can be said that during the whole process of the development of the thesis, starting from planning aspects and ending with the documentation, aspects related to ethics have been taken care of and addressed.

7.4 Societal Impact

Technology continues to advance providing economic growth but also causes the evolution of cybercrime, which is increasingly present in our daily lives (Mihaela-Sorina & Mihaela-Emilia, 2019). The large number of articles and studies on this subject are a clear sign that the malicious pages are a problem present in our daily lives.

A detection tool can not only be useful for the protection of the user in real time but can also help users lose the fear of performing actions such as shopping online. Because of this, the study tries to lay the foundations on which to develop a tool for detecting malicious pages. The extracted features can not only be useful for professionals or scholars who develop a tool, but any internet user can use them as a guide of points to check when entering a web page that is susceptible of being malicious.

Although it is impossible that the project can put an end to the malicious pages it is hoped that it can lay the foundation for future tools and research since it has tried to bring together the most relevant views of the research community. Finally, the community of researchers is encouraged to continue research in this field that can complete the results of this thesis and find new relevant features as the malicious pages do not stop evolving.

8 Conclusion

The objective of the study is to produce a list of features that can provide high accuracy in detecting fraudulent pages. This is made for the purpose of laying the foundations of a tool for detecting malicious pages and providing an aid to detect this type of fraudulent pages. An aid from which developers creating the detection tools, individuals using the web and system administrators of businesses can benefit.

In order to fulfill the purpose set for the project, a systematic study of the literature has been carried out. In this study, the most relevant features found in the literature have been extracted and categorized according to the available information. Throughout this process, the guidelines and methodologies set out above have been followed to ensure the validity of the process and research and to allow easy replication and assessment of the study.

After analysing the data from the accepted studies and extracting the relevant conclusions from them, the objective of the project has been achieved, to obtain a list of features that allow a high accuracy to detect the malicious pages. The list of features considered relevant to the research question can be found in the table below (Table 7) along with a summary of what each of the features is about.

Table 7: Relevant features

Feature Group	***Feature***	**Summary**
Search Bar Features	*IP Address*	The website has an IP address instead of a domain name in the address.
	Use of "-" Symbol	The website has a hyphen in the address.
	Use of "//" Symbol	The website address has "//" symbols in an anormal position.
	Use of "@" Symbol	The website has an at sign in the address.
	Use of Dots	The website has an excessive number of dots in the address.
	HTTPS Token in URL	The website has HTTPS written in a non-protocol part of the address.
	Keywords	The website has keywords regarded as suspicious in the address.
	Length of URL	The URL of the website is abnormally large.
Forms	*Blank Action*	The information provided through the form is not treated in any way
	Form Submits to Email	The information provided through the form is sent to an email account.
	Behaviour when Incorrect Login Information is introduced	The page behaves abnormally when entering incorrect data during login

Feature Group	*Feature*	**Summary**
Source Code	*Abnormal Anchor*	A considerable number of anomalous anchors are found
	Low Content Volume	The site has a low volume of content.
	Iframe Tag	The website uses the iframe tag to embed a third website in the page or to perform drive-by downloads.
	Disabling Right Click	Right Click of the is disabled or altered.
	Status Bar Customization	Address bar is altered to show false information.
	Language and Term Usage	Terms regarded as suspicious are used on the website.
	Use of External Resources	The website uses an anomalous amount of resources extracted from outside domains.
Miscellaneous	*HyperText Transfer Protocol Secure (HTTPS)*	The website does not have the HTTPS protocol, or it has it from an untrustworthy provider.
	Domain Age	The domain has a life span that can be considered suspicious.
	Use of Non-Standard Ports	Ports are used in an anomalous way such as having services in non-standardized ports or having all ports open.
	Alexa Rank	The website is in a bad position in Alexa ranking or does not even appear in it.
	E-Commerce Information	The site has no or false business information.

It has also been observed during the development that all studies employ the features in conjunction with others, so it can be concluded that there is no fail-proof feature for detection and that this identification procedure depends on many factors. Therefore, the use of several features in the detection systems is recommended. Moreover, the use of features could be complemented with other types of detection systems such as third-party blacklisting.

Although it has been possible to extract the list of the most relevant features, some of them could not be categorized. Gaps have been found in the research regarding some features and even some types of fraudulent pages. For this reason, in the chapter on future work (9 Future Work), these features and found gaps that can be studied in greater depth will be presented.

8.1 Future Work

From the thesis, several points have been extracted in which more research is needed. The first proposal is the development of a tool to detect malicious pages that would be based on the results presented in this thesis. Along with the development of this type of tool, there is a need to analyze certain aspects of the features that have been considered relevant. Table 8 shows the areas that are recommended for further study.

Table 8: Advised further research in relevant features

Feature	**Aspects to address**
Keywords on the URL	Development of a list of suspicious words or word alterations.
Length of URL	Decide which parts of the URL should be measured.
	Calculate a threshold value to classify an URL as too long.
Abnormal anchor	Calculate the approximate value that the ratio of abnormal to normal anchors should be.
Low content volume	Calculate a limit value to classify the volume level as low.
Language and term usage	Develop a list of suspicious terms.
Domain Age	Calculate the time range in which most malicious sites are located.

In the future, it is also recommended to analyse the next features in more depth, since with the current information, it has not been possible to get conclusive evidence to categorize them as either relevant or irrelevant:

- Extensions on the URL.
- User of URL shortening services.
- Forms pointing to external domains.
- Image size.
- Free Hosting.
- Google index.
- Product offer and pricing abnormalities.
- Product purchase abnormalities.

It is also remarkable the focus of studies on phishing, leaving aside other types of fraudulent pages such as fake online stores. It is highly recommended to study these other types of malicious websites since a large gap in information about them has been detected during the development of the study.

Finally, studies on malicious sites should be conducted on a regular basis. This is due to the rapid evolution of these pages since the features that can be useful now to detect them will most likely not be as significant in the future.

References

Abbasi, A., & Chen, H. (2009). A Comparison of Tools for Detecting Fake Websites. *Computer*, *42*(10), 78–86. https://doi.org/10.1109/MC.2009.306

Abbasi, Zhang, Zimbra, Chen, & Nunamaker. (2010). Detecting Fake Websites: The Contribution of Statistical Learning Theory. *MIS Quarterly*, *34*(3), 435. https://doi.org/10.2307/25750686

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 Systems with Applications*, *115*, 300–313. https://doi.org/10.1016/j.eswa.2018.07.067

Ahmed, A. A., & Abdullah, N. A. (2016). Real time detection of phishing websites. *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 1–6. https://doi.org/10.1109/IEMCON.2016.7746247

Ali Ahmed, A. (2018). Malicious Website Detection: A Review. *Journal of Forensic Sciences & Criminal Investigation*, *7*(3). https://doi.org/10.19080/JFSCI.2018.07.555712

Ali, W. (2017). Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection. *International Journal of Advanced Computer Science and Applications (Ijacsa)*, *8*(9), Article 9. https://doi.org/10.14569/IJACSA.2017.080910

AlShboul, R., Thabtah, F., Abdelhamid, N., & Al-diabat, M. (2018). A visualization cybersecurity method based on features' dissimilarity. *Computers & Security*, *77*, 289–303. https://doi.org/10.1016/j.cose.2018.04.007

Amiri, I. S., Akanbi, O. A., & Fazeldehkordi, E. (2015). *A Machine-Learning Approach to Phishing Detection and Defense* (1st ed.). Syngress Publishing.

Amrutkar, C., Kim, Y. S., & Traynor, P. (2017). Detecting Mobile Malicious Webpages in Real Time. *IEEE Transactions on Mobile Computing*, *16*(8), 2184–2197. https://doi.org/10.1109/TMC.2016.2575828

Anderson, R., Barton, C., Boehme, R., Clayton, R., Ganan, C., Grasso, T., Levi, M., Moore, T., Savage, S., & Vasek, M. (2019). *Measuring the Changing Cost of Cybercrime*. 32.

Arshad, S., Kharraz, A., & Robertson, W. (2017). Include Me Out: In-Browser Detection of Malicious Third-Party Content Inclusions. *ArXiv:1811.00926 [Cs]*, *9603*, 441–459. https://doi.org/10.1007/978-3-662-54970-4_26

Aung, E. S., & Yamana, H. (2019). URL-Based Phishing Detection Using the Entropy of Non-Alphanumeric Characters. *Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services*, 385–392. https://doi.org/10.1145/3366030.3366064

Bell, S., & Komisarczuk, P. (2020). An Analysis of Phishing Blacklists: Google Safe Browsing, OpenPhish, and PhishTank. *Proceedings of the Australasian Computer Science Week Multiconference*. https://doi.org/10.1145/3373017.3373020

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

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. *Journal of Systems and Software*, *80*(4), 571–583. https://doi.org/10.1016/j.jss.2006.07.009

Buber, E., Demir, Ö., & Sahingoz, O. K. (2017). Feature selections for the machine learning based detection of phishing websites. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–5. https://doi.org/10.1109/IDAP.2017.8090317

Carpineto, C., & Romano, G. (2017). Learning to Detect and Measure Fake Ecommerce Websites in Search-Engine Results. *Proceedings of the International Conference on Web Intelligence*, 403–410. https://doi.org/10.1145/3106426.3106441

Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S. C., & Tiong, W. K. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection

system. *Information Sciences*, *484*, 153–166. https://doi.org/10.1016/j.ins.2019.01.064

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In G. Goos, J. Hartmanis, & J. van Leeuwen (Eds.), *Multiple Classifier Systems* (Vol. 1857, pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Dinev, T. (2006). Why spoofing is serious internet fraud. *Communications of the ACM*, *49*(10), 76–82. https://doi.org/10.1145/1164394.1164398

Ding, Y., Luktarhan, N., Li, K., & Slamu, W. (2019). A keyword-based combination approach for detecting phishing webpages. *Computers & Security*, *84*, 256–275. https://doi.org/10.1016/j.cose.2019.03.018

Efron, S. E., & Ravid, R. (2019). *Writing the literature review: A practical guide*. The Guilford Press.

Eshete, B. (2013). Effective analysis, characterization, and detection of malicious web pages. *Proceedings of the 22nd International Conference on World Wide Web*, 355–360. https://doi.org/10.1145/2487788.2487942

Facts + Statistics: Identity theft and cybercrime | III. (n.d.). Retrieved 7 June 2020, from https://www.iii.org/fact-statistic/facts-statistics-identity-theft-and-cybercrime

Fink, A. G. (2019). *Conducting Research Literature Reviews: From the Internet to Paper* (5th ed.). SAGE Publications.

Gajera, K., Jangid, M., Mehta, P., & Mittal, J. (2019). A Novel Approach to Detect Phishing Attack Using Artificial Neural Networks Combined with Pharming Detection. *2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 196–200. https://doi.org/10.1109/ICECA.2019.8822053

Gañán, C. H., Ciere, M., & van Eeten, M. (2017). Beyond the pretty penny: The Economic Impact of Cybercrime. *Proceedings of the 2017 New Security Paradigms Workshop*, 35–45. https://doi.org/10.1145/3171533.3171535

Gibbs, G. R. (2007). *Analyzing Qualitative Data*. SAGE Publications.

Gupta, B. B., Tewari, A., Jain, A. K., & Agrawal, D. P. (2017). Fighting against phishing attacks: State of the art and future challenges. *Neural Computing and Applications*, *28*(12), 3629–3654. https://doi.org/10.1007/s00521-016-2275-y

Herzberg, A., & Jbara, A. (2008). Security and identification indicators for browsers against spoofing and phishing attacks. *ACM Transactions on Internet Technology*, *8*(4), 16:1–16:36. https://doi.org/10.1145/1391949.1391950

How to Protect Yourself from Phishing and Fake Websites—DSL Internet Support. (n.d.). Retrieved 18 March 2020, from https://www.att.com/support/article/dsl-high-speed/KM1010136

Jesson, J., Lacey, F. M., & Matheson, L. (2011). *Doing Your Literature Review: Traditional and Systematic Techniques* (1st ed.). SAGE Publications.

Kanei, F., Chiba, D., Hato, K., & Akiyama, M. (2019). Precise and Robust Detection of Advertising Fraud. *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, *1*, 776–785. https://doi.org/10.1109/COMPSAC.2019.00115

Kharraz, A., Robertson, W., & Kirda, E. (2018). Surveylance: Automatically Detecting Online Survey Scams. *2018 IEEE Symposium on Security and Privacy (SP)*, 70–86. https://doi.org/10.1109/SP.2018.00044

Kim, D. W., Yan, P., & Zhang, J. (2015). Detecting fake anti-virus software distribution webpages. *Computers & Security*, *49*, 95–106. https://doi.org/10.1016/j.cose.2014.11.008

Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews*. 33.

Kovarik, V. J. (2009). Chapter 12 - Cognitive Research: Knowledge Representation and Learning. In B. A. Fette (Ed.), *Cognitive Radio Technology (Second Edition)* (pp. 367–399). Academic Press. https://doi.org/10.1016/B978-0-12-374535-4.00012-6

K.p, S. B., & Damodaram, D. R. (2016). Phishing Detection in Websites Using Neural Networks and Firefly. *International Journal of Engineering and Computer Science*, *5*(9), Article 9. http://www.ijecs.in/index.php/ijecs/article/view/2552

Le, V. L., Welch, I., Gao, X., & Komisarczuk, P. (2011). Identification of potential malicious web pages. *Proceedings of the Ninth Australasian Information Security Conference - Volume 116*, 33–40.

Li, T., Kou, G., & Peng, Y. (2020). Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods. *Information Systems*, *91*, 101494. https://doi.org/10.1016/j.is.2020.101494

Li, Y., Yang, Z., Chen, X., Yuan, H., & Liu, W. (2019). A stacking model using URL and HTML features for phishing webpage detection. *Future Generation Computer Systems*, *94*, 27–39. https://doi.org/10.1016/j.future.2018.11.004

Liang, B., Su, M., You, W., Shi, W., & Yang, G. (2016). Cracking Classifiers for Evasion: A Case Study on the Google's Phishing Pages Filter. *Proceedings of the 25th International Conference on World Wide Web*, 345–356. https://doi.org/10.1145/2872427.2883060

Machado, L., & Gadge, J. (2017). Phishing Sites Detection Based on C4.5 Decision Tree Algorithm. *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 1–5. https://doi.org/10.1109/ICCUBEA.2017.8463818

Maktabdar Oghaz, M., Zainal, A., Maarof, M., & Kassim, M. (2017, December 13). *Content based Fraudulent Website Detection Using Supervised Machine Learning Techniques*.

Marchal, S., Saari, K., Singh, N., & Asokan, N. (2016). Know Your Phish: Novel Techniques for Detecting Phishing Sites and their Targets. *ArXiv:1510.06501 [Cs]*. http://arxiv.org/abs/1510.06501

Mehrotra, D. D. (2020). *AI - Artificial Intelligence Basics For School Students (Class IX): As per the latest CBSE curriculum (Code No. 417)*. Notion Press.

Minhaz Uddin, S., Shihabuz Zaman, Zul Kawsar, Ashaduzzaman, & Pritom, A. I. (2019). *Phishing Website Detection Using Effective Classifiers and Feature Selection Techniques*. https://doi.org/10.13140/RG.2.2.24043.08483

Moghimi, M., & Varjani, A. Y. (2016). New rule-based phishing detection method. *Expert Systems with Applications*, *53*, 231–242. https://doi.org/10.1016/j.eswa.2016.01.028

Mostard, W., Zijlema, B., & Wiering, M. (2019). Combining Visual and Contextual Information for Fraudulent Online Store CIassification. *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 84–90.

Najork, M. (2017). Web Spam Detection. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 1–5). Springer. https://doi.org/10.1007/978-1-4899-7993-3_465-3

Ndibwile, J. D., Kadobayashi, Y., & Fall, D. (2017). UnPhishMe: Phishing Attack Detection by Deceptive Login Simulation through an Android Mobile App. *2017 12th Asia Joint Conference on Information Security (AsiaJCIS)*, 38–47. https://doi.org/10.1109/AsiaJCIS.2017.19

Orunsolu, A. A., Sodiya, A. S., & Akinwale, A. T. (2019). A predictive model for phishing detection. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2019.12.005

Paré, G., & Kitsiou, S. (2017). Chapter 9 Methods for Literature Reviews. In *Handbook of eHealth Evaluation: An Evidence-based Approach [Internet]*. University of Victoria. https://www.ncbi.nlm.nih.gov/books/NBK481583/

Parekh, S., Parikh, D., Kotak, S., & Sankhe, S. (2018). A New Method for Detection of Phishing Websites: URL Detection. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 949–952. https://doi.org/10.1109/ICICCT.2018.8473085

Parulekar, C. (2019). *Minimize Phishing Attacks: Securing Spear attacks*. *06*(06), 5.

Perner, P. (2010). *Advances in Data Mining. Applications and Theoretical Aspects*. Springer.

Pienta, D., Thatcher, J. B., & Johnston, A. C. (2018). *A Taxonomy of Phishing: Attack Types Spanning Economic, Temporal, Breadth, and Target Boundaries*. 18.

Qabajeh, I., Thabtah, F., & Chiclana, F. (2018). A recent review of conventional vs. Automated cybersecurity anti-phishing techniques. *Computer Science Review*, *29*, 44–55. https://doi.org/10.1016/j.cosrev.2018.05.003

Rajab, M. (2018). An Anti-Phishing Method Based on Feature Analysis. *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, 133–139. https://doi.org/10.1145/3184066.3184082

Rao, R., & Pais, A. R. (2017). Detecting Phishing Websites Using Automation of Human Behavior. *Proceedings of the 3rd ACM Workshop on Cyber-Physical System Security*, 33–42. https://doi.org/10.1145/3055186.3055188

Rao, R. S., & Ali, S. T. (2015). PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach. *Eleventh International Conference on Communication Networks, ICCN 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Data Mining and Warehousing, ICDMW 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Image and Signal Processing, ICISP 2015, August 21-23, 2015, Bangalore, India*, *54*, 147–156. https://doi.org/10.1016/j.procs.2015.06.017

Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications*, *31*(8), 3851–3873. https://doi.org/10.1007/s00521-017-3305-0

Ridley, D. (2012). *The Literature Review: A Step-by-Step Guide for Students* (2nd ed.). SAGE Publications.

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029

Schneier, B. (2000). *Inside risks: Semantic network attacks*. Association for Computing Machinery. https://doi.org/10.1145/355112.355131

Shekokar, Narendra. M., Shah, C., Mahajan, M., & Rachh, S. (2015). An Ideal Approach for Detection and Prevention of Phishing Attacks. *Proceedings of 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)*, *49*, 82–91. https://doi.org/10.1016/j.procs.2015.04.230

Shirsat, S. D. (2018). *Demonstrating Different Phishing Attacks Using Fuzzy Logic—IEEE Conference Publication*. https://ieeexplore-ieee-org.libraryproxy.his.se/document/8473309

Silva, C. M. R. da, Feitosa, E. L., & Garcia, V. C. (2020). Heuristic-based strategy for Phishing prediction: A survey of URL-based approach. *Computers & Security*, *88*, 101613. https://doi.org/10.1016/j.cose.2019.101613

Sonowal, G., & Kuppusamy, K. S. (2016). MASPHID: A Model to Assist Screen Reader Users for Detecting Phishing Sites Using Aural and Visual Similarity Measures. *Proceedings of the International Conference on Informatics and Analytics*. https://doi.org/10.1145/2980258.2980443

Suri, H. (2020). Ethical Considerations of Conducting Systematic Reviews in Educational Research. In O. Zawacki-Richter, M. Kerres, S. Bedenlier, M. Bond, & K. Buntins (Eds.), *Systematic Reviews in Educational Research: Methodology, Perspectives and Application* (pp. 41–54). Springer Fachmedien. https://doi.org/10.1007/978-3-658-27602-7_3

Tian, K., Jan, S. T. K., Hu, H., Yao, D., & Wang, G. (2018). Needle in a Haystack: Tracking Down Elite Phishing Domains in the Wild. *Proceedings of the Internet Measurement Conference 2018*, 429–442. https://doi.org/10.1145/3278532.3278569

Ushmani, A. (2019a). *(PDF) Internet Fraud Analysis*. ResearchGate. https://www.researchgate.net/publication/331481708_Internet_Fraud_Analysis

Ushmani, A. (2019b, March 3). *Internet Fraud Analysis*.

Vrbančič, G., Fister, I., & Podgorelec, V. (2018). Swarm Intelligence Approaches for Parameter Setting of Deep Learning Neural Network: Case Study on Phishing Websites Classification. *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. https://doi.org/10.1145/3227609.3227655

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14*, 1–10. https://doi.org/10.1145/2601248.2601268

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer-Verlag. https://doi.org/10.1007/978-3-642-29044-2

Wuest, C., & Ramzan, Z. (2014). *Methods and systems for identifying fraudulent websites*. ResearchGate. https://www.researchgate.net/publication/302787823_Methods_and_systems_for_identifying_fraudulent_websites

Wüest, C., & Ramzan, Z. (2014). *Methods and systems for identifying fraudulent websites* (United States Patent No. US8856937B1). https://patents.google.com/patent/US8856937B1/en

Zhang, D., Yan, Z., Jiang, H., & Kim, T. (2014). A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. *Information & Management*, *51*(7), 845–853. https://doi.org/10.1016/j.im.2014.08.003

Zhang, W., Ren, H., & Jiang, Q. (2016). Application of Feature Engineering for Phishing Detection. *IEICE Transactions on Information and Systems*, *E99.D*(4), 1062–1070. https://doi.org/10.1587/transinf.2015CYP0005

Zhou, X., Jin, Y., Zhang, H., Li, S., & Huang, X. (2016). A Map of Threats to Validity of Systematic Literature Reviews in Software Engineering. *2016 23rd Asia-Pacific*

Software Engineering Conference (APSEC), 153–160. https://doi.org/10.1109/APSEC.2016.031

Zouina, M., & Outtaj, B. (2017). A novel lightweight URL phishing detection system using SVM and similarity index. *Human-Centric Computing and Information Sciences*, *7*(1), 17. https://doi.org/10.1186/s13673-017-0098-1

APPENDIX A – Accepted Articles

Label	Title	Author and Publication Year
A1	*UnPhishMe: Phishing Attack Detection by Deceptive Login Simulation through an Android Mobile App*	Ndibwile, J. D., Kadobayashi, Y., & Fall, D. (2017)
A2	*Surveylance: Automatically Detecting Online Survey Scams*	Kharraz, A., Robertson, W., & Kirda, E. (2018)
A3	*Real time detection of phishing websites*	Ahmed, A. A., & Abdullah, N. A. (2016)
A4	*Detecting Mobile Malicious Webpages in Real Time*	Amrutkar, C., Kim, Y. S., & Traynor, P. (2017)
A5	*Demonstrating Different Phishing Attacks Using Fuzzy Logic*	Shirsat, S. D. (2018)
A6	*A Novel Approach to Detect Phishing Attack Using Artificial Neural Networks Combined with Pharming Detection*	Gajera, K., Jangid, M., Mehta, P., & Mittal, J. (2019)
A7	*A New Method for Detection of Phishing Websites: URL Detection*	Parekh, S., Parikh, D., Kotak, S., & Sankhe, S. (2018)
A8	*An Anti-Phishing Method Based on Feature Analysis*	Rajab, M. (2018)
A9	*Detecting Phishing Websites Using Automation of Human Behavior*	Rao, R., & Pais, A. R. (2017)
A10	*Learning to Detect and Measure Fake Ecommerce Websites in Search-Engine Results*	Carpineto, C., & Romano, G. (2017)
A11	*MASPHID: A Model to Assist Screen Reader Users for Detecting Phishing Sites Using Aural and Visual Similarity Measures*	Sonowal, G., & Kuppusamy, K. S. (2016)
A12	*Needle in a Haystack: Tracking Down Elite Phishing Domains in the Wild*	Tian, K., Jan, S. T. K., Hu, H., Yao, D., & Wang, G. (2018)
A13	*URL-Based Phishing Detection Using the Entropy of Non-Alphanumeric Characters*	Aung, E. S., & Yamana, H. (2019)
A14	*A keyword-based combination approach for detecting phishing webpages*	Ding, Y., Luktarhan, N., Li, K., & Slamu, W. (2019)
A15	*A predictive model for phishing detection*	Orunsolu, A. A., Sodiya, A. S., & Akinwale, A. T. (2019)
A16	*An Ideal Approach for Detection and Prevention of Phishing Attacks*	Shekokar, Narendra. M., Shah, C., Mahajan, M., & Rachh, S. (2015)
A17	*Heuristic-based strategy for Phishing prediction: A*	Silva, C. M. R. da, Feitosa,

Label	Title	Author and Publication Year
	survey of URL-based approach	E. L., & Garcia, V. C. (2020)
A18	*Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text*	Adebowale, M. A., Lwin, K. T., Sánchez, E., & Hossain, M. A. (2019)
A19	*PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach*	Rao, R. S., & Ali, S. T. (2015)
A20	*A new hybrid ensemble feature selection framework for machine learning-based phishing detection system*	Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S. C., & Tiong, W. K. (2019)
A21	*A recent review of conventional vs. Automated cybersecurity anti-phishing techniques*	Qabajeh, I., Thabtah, F., & Chiclana, F. (2018)
A22	*A visualization cybersecurity method based on features' dissimilarity*	AlShboul, R., Thabtah, F., Abdelhamid, N., & Al-diabat, M. (2018)
A23	*Combining Visual and Contextual Information for Fraudulent Online Store CIassification*	Mostard, W., Zijlema, B., & Wiering, M. (2019)
A24	*Detecting fake anti-virus software distribution webpages*	Kim, D. W., Yan, P., & Zhang, J. (2015)
A25	*Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods*	Li, T., Kou, G., & Peng, Y. (2020)
A26	*Machine learning based phishing detection from URLs*	Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019)
A27	*New rule-based phishing detection method*	Moghimi, M., & Varjani, A. Y. (2016)
A28	*Phishing Sites Detection Based on C4.5 Decision Tree Algorithm*	Machado, L., & Gadge, J. (2017)
A29	*Precise and Robust Detection of Advertising Fraud*	Kanei, F., Chiba, D., Hato, K., & Akiyama, M. (2019)
A30	*Swarm Intelligence Approaches for Parameter Setting of Deep Learning Neural Network: Case Study on Phishing Websites Classification*	Vrbančič, G., Fister, I., & Podgorelec, V. (2018)
A31	*A novel lightweight URL phishing detection system using SVM and similarity index*	Zouina, M., & Outtaj, B. (2017)
A32	*A stacking model using URL and HTML features for phishing webpage detection*	Li, Y., Yang, Z., Chen, X., Yuan, H., & Liu, W. (2019)
A33	*Application of Feature Engineering for Phishing Detection*	Zhang, W., Ren, H., & Jiang, Q. (2016)
A34	*Cracking Classifiers for Evasion: A Case Study on the Google's Phishing Pages Filter*	Liang, B., Su, M., You, W., Shi, W., & Yang, G. (2016)
A35	*Detection of phishing websites using an efficient feature-based machine learning framework*	Rao, R. S., & Pais, A. R. (2019)

Label	Title	Author and Publication Year
A36	*Feature selections for the machine learning based detection of phishing websites*	Buber, E., Demir, Ö., & Sahingoz, O. K. (2017)
A37	*Include Me Out: In-Browser Detection of Malicious Third-Party Content Inclusions*	Arshad, S., Kharraz, A., & Robertson, W. (2017)
A38	*Know Your Phish: Novel Techniques for Detecting Phishing Sites and their Targets*	Marchal, S., Saari, K., Singh, N., & Asokan, N. (2016)
A39	*Phishing Detection in Websites Using Neural Networks and Firefly*	K.p, S. B., & Damodaram, D. R. (2016)
A40	*Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection*	Ali, W. (2017)