Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment

Dong-Jie Liu[a,b], *Guang-Gang Geng*[c,*], *Xiao-Bo Jin*[d], *Wei Wang*[a]

[a] *Computer Network Information Center, Chinese Academy of Sciences, Beijing, China*
[b] *University of Chinese Academy of Sciences, Beijing, China*
[c] *College of Cyber Security, Jinan University, Guangzhou 510632, China*
[d] *Department of Intelligent Science, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China*

ARTICLE INFO

Article history:
Received 5 January 2021
Revised 11 May 2021
Accepted 26 July 2021
Available online 3 August 2021

Keywords:
Phishing detection
CASE feature framework
Multistage model
Machine learning
Real web environment

ABSTRACT

Phishing has become a favorite method of hackers for committing data theft and continues to evolve. As long as phishing websites continue to operate, many more people and companies will suffer privacy leaks or financial losses. Therefore, the demand for fast and accurate phishing website detection grows stronger. However, the existing phishing detection methods do not fully analyze the features of phishing, and the performance and efficiency of the models only apply to certain limited datasets and need to be improved to be applied to the real web environment. This paper fully considers the social engineering principles of phishing, proposes a comprehensive and interpretable CASE feature framework and designs a multistage phishing detection model to effectively detect phishing sites, especially in the real web environment, where high efficiency and performance and extremely low false alarm rates are required. To fully verify the proposed method, two kinds of data experiments were carried out. One was the comparative experiments among different features and different detection models on CASE, which covers both classic machine learning and deep learning algorithms based on a constructed complex dataset. The other was a one-year phishing discovery experiment in the real web environment. The proposed method achieves better detection results under the premise of significantly shortening the execution time and works well in real phishing discovery, which proves its high practicability in reality.

1. Introduction

Phishing is a typical social engineering attack. Attackers use the instincts, curiosity, trust, fear and greed of users to commit crimes. Phishing attacks have seen an impressive 350% increase during the COVID-19 quarantine Phishing attacks increase 350 percent amid covid-19 quarantine and seriously threaten the privacy and property of web users. According to cybersecurity research reports, phishing is a hacker's favorite method of data theft Popular phishing techniques used by hackers. It is estimated that the cost of phishing is now 1/4 of the cost of traditional cyber-attacks, but the income is

* Corresponding author.
E-mail address: guanggang.geng@gmail.com (G.-G. Geng).
https://doi.org/10.1016/j.cose.2021.102421

twice as high as it was in the past. Medium-sized companies paid an average of $1.6 million to deal with phishing attacks Phishing statistics: What every business needs to know. A bigger problem is that companies can lose customers faster than they can get new customers. Deloitte reports that if a company encounters a network security breach, 1/3 of consumers will lose interest in the company. Even if they did not suffer economic losses due to vulnerabilities, the figure still applies Consumer data under attack.

Phishing attacks have many forms and usually involve a variety of communication channels, such as email, instant messages and social media. Regardless of the channel used, attackers often counterfeit well-known banks, credit card companies or famous e-commerce websites to intimidate or urge users to log in to the phishing website to do things that the customer will regret later. For example, a user might receive an instant message indicating a problem with their bank account and are directed to a web link that is very similar to a link used by the bank. Without hesitation, the user enters their username and password in the fields provided. Criminals record this information and then use it to access the user's legitimate accounts.

At present, statistical learning-based online phishing site detection is the mainstream antiphishing method, but its robustness and efficiency in complex web environments need to be improved (Basit et al., 2020; Chiew et al., 2015; Geng et al., 2013; 2015; Moghimi and Varjani, 2016; Sahingoz et al., 2019). The main problems of antiphishing methods based on machine learning are summarized as follows:

- An increasing number of features are extracted by antiphishing methods, but why these features are extracted is not clear. The existing features do not well reflect the nature of phishing, which steals sensitive information through spoofing. This leads to a result in which the features are only valid in a few limited and specific scenarios, such as for specified datasets or a browser plug-in (Chen et al., 2018; Chiew et al., 2015; Gupta, 2016; Liang et al., 2020; Likarish et al., 2008).
- The existing algorithms treat all websites the same, which leads to the inefficiency of the model (Gupta, 2016; Liang et al., 2020; Moghimi and Varjani, 2016; Sahingoz et al., 2019). In other words, the models are not suitable for the real web environment, which contains a large number of complex webpages.
- Most datasets do not contain enough samples, and the sample diversity is not considered; moreover, the ratio of positive and negative samples is not realistic (Chiew et al., 2015; Jain and Gupta, 2018; Moghimi and Varjani, 2016; Rao and Pais, 2018). In general, models based on such datasets experience heavy overfitting, and the robustness of the models needs further improvement.

Based on the discussion above and aiming at the real web environment, this paper designs a robust and efficient large-scale phishing detection method based on statistical machine learning algorithms and proves its efficiency with two kinds of experiments. The contributions of this work are summarized as follows:

- In the aspect of antiphishing statistic feature extraction, through an in-depth analysis of the pattern of phishing attacks, this paper extracts comprehensive and interpretable quaternary features, called CASE, which includes "Counterfeiting", "Affiliation", "Stealing" and "Evaluation" features. "Counterfeiting" and "Stealing" features reflect the social engineering characteristics of phishing attacks. "Affiliation" and "Evaluation" features reflect the relevance and quality of the web contents. The CASE feature framework covers the feature space that reflects the spoofing nature of phishing, ensures the discrimination and generalization of features, and provides feature-level support for effective phishing detection.
- In terms of the detection model design, considering the extremely unbalanced reality of legitimate and phishing sites, a multistage detection model is proposed. The core idea of this multistage model is "fast filtering + accurate recognition". As the name implies, most legitimate websites are excluded during the rapid filtering stage; then, an accurate supervised recognition model is designed by learning specific positive and negative samples in a smaller range. This model design ensures high performance under the premise of a shorter detection time, which is more applicable to the real web environment.
- In terms of dataset construction, to make the dataset as similar as possible to those in the real web environment, this paper tries to construct a dataset that contains websites obtained from different sources with different languages, content qualities and brands. In addition, considering that phishing detection is a class imbalance problem, this paper used a large ratio between positive and negative samples. In addition, this dataset contains multiple confused samples that are very hard to detect. All these features of the dataset increase the detection difficulty but, more importantly, prove that the proposed detection method is effective and practical in a real web environment.

2. Related work

Machine learning-based phishing website detection is the current mainstream antiphishing method (Basnet et al., 2012; Chen et al., 2018; Hiransha et al., 2018; Liang et al., 2020; Moghimi and Varjani, 2016; Tajaddodianfar et al., 2020; Wei et al., 2019). In recent years, the media of phishing communications (e.g., email, instant messaging, and social media) has become more and more diverse. Although there are still some studies that try to distinguish phishing from other media, such as e-mail (Moradpoor et al., 2017; Subasi et al., 2017), the focus of antiphishing research is still mainly on the detection of phishing websites. Therefore, to lay a good foundation for the following research, this paper mainly summarizes related work of statistical phishing website detection, including

statistical feature extraction, detection models and dataset construction for machine learning-based phishing detection.

2.1. *Statistical feature extraction*

It is crucial that the extracted features can effectively reflect the pattern of the detection object. In the existing research, extracted antiphishing features include URLs, titles, hyperlinks, login boxes, copyright information, sensitive terms, and search engine information (Arachchilage et al., 2016; Bahnsen et al., 2017; Basnet et al., 2012; Bilge et al., 2011; Chiew et al., 2015; Feroz and Mengel, 2015; Garera et al., 2007; Geng et al., 2015; Jain and Gupta, 2017; Likarish et al., 2008; Ma et al., 2009; Moghimi and Varjani, 2016; Oliveira et al., 2017; Sahingoz et al., 2019; Xiang et al., 2011; Xiang and Hong, 2009). These statistical features show the characteristics of different aspects of phishing websites. It has been shown that they can be used to identify phishing websites. In addition, visual spoofing features and evaluation features have received more attention in recent years (Chiew et al., 2015; Dhamija and Tygar, 2005; Geng et al., 2013; 2015; Maurer and Herzner, 2012; Wang et al., 2011; Xiang et al., 2011; Xiang and Hong, 2009).

To deceive users and achieve the purpose of falsification, phishing websites are often visually highly similar to the corresponding brand websites. We call this "visual spoofing". At present, a few studies on phishing recognition are based on visual analysis (Chiew et al., 2015; Dhamija and Tygar, 2005; Geng et al., 2013; 2015; Maurer and Herzner, 2012; Wang et al., 2011). Dhamija et al. proposed a dynamic skin approach (Dhamija and Tygar, 2005) that focuses on the authentication of image entities. However, the article only includes a theoretical analysis and does not give an experimental evaluation. Maurer et al. used the visual similarity of websites to detect the frameworks of suspicious phishing websites (Maurer and Herzner, 2012), and this study also lacks experimental verification. The phishing detection method based on logo recognition mines important brand elements that characterize website identities (Chiew et al., 2015; Wang et al., 2011). However, it is not sufficiently robust to only use the logo to determine whether a website is a phishing website. For example, if a website contains a brand logo on the page, it does not mean that the website is a phishing website; it may be a news website, a subbrand website or has been authorized to use the logo. Chiew et al. (2015) used the Google search engine to identify the attribution of a logo; this method has a certain effect but relies too much on search engine resources, which makes the method lack practicality. We proposed a detection method based on favicon recognition in our previous work (Geng et al., 2013), which has good detection performance on phishing scams with favicon counterfeiting; later, in another work, we further integrated logo and copyright recognition, and the performance was further improved (Geng et al., 2015), but it encountered a similar problem in which other features were ignored and the detection performance could be better.

Third-party service features were extracted to reduce the false positive rate (Xiang et al., 2011; Xiang and Hong, 2009). The extracted third-party features include the search engine index, the PageRank value, the domain name age and other features. We analyzed the importance of "brand authorization recognition" in our early work and extracted several features, which proved to be effective in reducing the false detection rate (Geng et al., 2015).

The APWG statistics show that "more than 98% of phishing websites use fake domain names Global phishing survey, Apwg phishing attack trends reports list" to achieve the purpose of "name spoofing". Existing research does not pay enough attention to this fact. URL string information is used by most researchers (Le et al., 2011), but mining the underlying information behind the domain name, such as the domain registration and resolution, is also very important (Ali et al., 2019; Pandey and Singh, 2019). This information can often indicate whether a domain name has the right to provide related brand services.

Based on the discussion above, the current phishing feature extraction seems subjective and incomplete and lacks effective analyses of social engineering factors. Another example, Cantina+(Xiang et al., 2011), is characterized by rich features and includes features such as URL string features, web page structure features and search service features. The dataset (with positive and negative sample equalization) only achieved a 92% TPR (true positive rate), and the performance was far from satisfactory.

In summary, as for feature extraction, questions, such as "why extract these features, why do these features have a certain effect, what other features should be further extracted, and how can multi-scale features be comprehensively utilized?", are lacking in sufficient analysis. Therefore, extracting effective features is the primary research task of this paper. We analyze social engineering attacks and propose a comprehensive and interpretable feature framework that not only covers all aspects of phishing attacks but also covers web content quality and relevance.

2.2. *Detection model*

Machine learning-based phishing detection models include communication channel-based models (Aggarwal et al., 2014; Akinyelu and Adewumi, 2014), website identifier-based models (Bilge et al., 2011; Le et al., 2011; Ma et al., 2009) and website content-based models (Basnet et al., 2012; Dhamija and Tygar, 2005; Geng et al., 2013; 2015; Likarish et al., 2008; Maurer and Herzner, 2012; Wang et al., 2011; Xiang et al., 2011). Phishing sites are often the last and most important stage in social engineering attacks, and most phishing scams eventually mislead users to visit their prebuilt spoof sites. At present, website content-based phishing detection is the most important countermeasure against phishing attacks (Basnet et al., 2012; Castao et al., 2021; Geng et al., 2013; 2015; Maurer and Herzner, 2012; Wang et al., 2011; Xiang et al., 2011; Zhang et al., 2021).

The existing classification algorithms used in phishing website detection mainly include naive Bayes, decision trees, random forest, AdaBoost, and support vector machines. Basnet et al. (2012); Bilge et al. (2011); Garera et al. (2007); Likarish et al. (2008); Ma et al. (2009); Moghimi and Varjani (2016); Xiang et al. (2011). These learning algorithms are widely used in the field of pattern recognition, such as document classification and biometric recognition, and have achieved good performance. What is more, deep learning algorithms, such as CNN and LSTM, are currently used in phishing detection (Chen et al., 2018; Hiransha et al., 2018; Liang

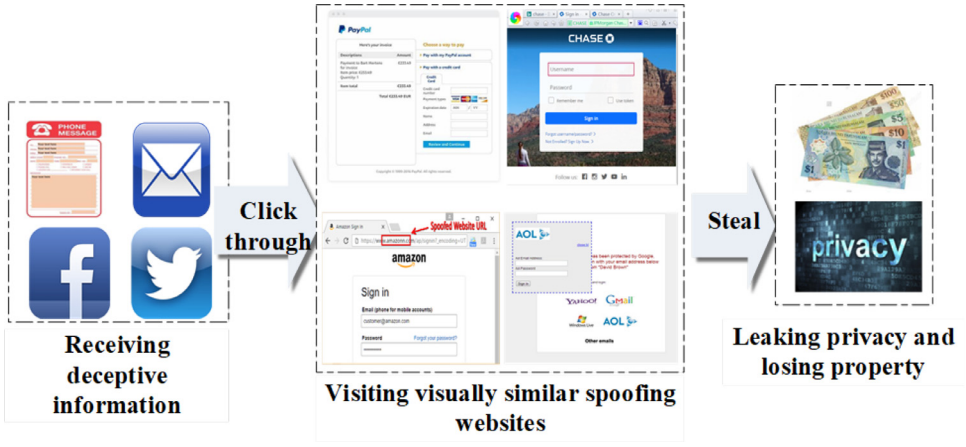

Fig. 1 – Typical Phishing Process.

et al., 2020; Tajaddodianfar et al., 2020; Wei et al., 2019). If these models are to achieve good results in the real web environment, they need training samples that cover a sufficiently large range of various web pages. However, most of the existing antiphishing research experiments are based on relatively small sample sets, so the generalization ability and practicality of the models are questionable. Taking a step back, even if the training sample set is large enough to cover all kinds of samples from the real web environment, phishing website detection has an extreme class imbalance problem (there are more than one billion websites, but there are only approximately 100,000 phishing websites per year Apwg phishing attack trends reports list). Therefore, it is difficult to obtain good detection performance by directly employing the abovementioned classification models.

2.3. *Dataset*

Different from the common pattern recognition problem, the number of targeted brands is large, the website uptime is extremely short (less than 30 hours on average), and brand spoofing technology is constantly evolving (such as picture-in-picture without any text). These characteristics make the construction of a practical and authoritative phishing dataset very difficult. Moreover, there is some research that only evaluates the proposed algorithm on a few dozen samples, and thus, the algorithm generalization is greatly reduced.

In 2015, Mohammad et al. published a phishing dataset (Mohammad et al., 2015), which includes a publicly available phishing dataset on the Internet. This dataset has the following shortcomings: no webpage data are provided; only 30-dimensional prefetched features are provided, and users cannot verify whether the features are accurate or not; the number of samples is small, and the number of positive and negative samples in the dataset is balanced, which is not the case in reality; no webpage screenshots are provided, so vision-based phishing detection experiments cannot be supported; and without webpage data and screenshot data, extra feature extraction is impossible for the phishing sites that have expired. Another dataset by Vrbančič et al. also has the above problems(Vrbančič et al., 2020). The pre-extracted features of the data set are only extracted from the URL; hence, its scope is narrower, and the application of the dataset is limited. Based on these limitations, this paper builds a practical dataset that contains multiple obfuscated samples to support the validation of the effectiveness of the proposed features and model.

3. Proposed CASE feature framework

The typical process of a phishing attack is shown in Fig. 1. The user receives information containing social engineering content and fraudulent links and is then misled to visit a phishing site, which is visually similar to a brand site. If the user inputs sensitive information, the information will be stolen. Currently, the spread of phishing websites is not limited to e-mail but is further extended to instant messaging and social networks. The diversity in the means of communication has made it increasingly unrealistic to combat phishing from the source of the information communication. Moreover, online banks, payment services, etc. cannot identify the real identity of visitors, for instance, whether visitors are phishers or legitimate users. These services only recognize username/card numbers and passwords visitors entered. At present, the most reliable and practical antiphishing method is phishing website detection, and the phishing recognition ability of models relies on whether the extracted features are effective(Basit et al., 2020; Basnet et al., 2012; Bilge et al., 2011; Geng et al., 2013; 2015; Jain and Gupta, 2018; Kang et al., 2015; Moghimi and Varjani, 2016; Xiang et al., 2011).

Phishing websites look similar to real brand websites and steal accounts, passwords or other private information submitted by victims. From the basic definition of phishing websites, it is obvious that there are two key types of features that need to be extracted to identify phishing websites. One is features related to webpage similarity, and the other is features that implement the theft function. Features extracted by most research belong to these two types. For instance, Geng et al. extracted favicon, logo, copyright notice redirection, incoming links and domain name system (DNS) information (Geng et al., 2015); Kang et al. used logo images (Kang et al., 2015); Sahingoz et al. chose URLs for classification (Sahingoz et al., 2019);

Table 1 – Possible Features under Each Feature Space.

Feature Space	Features
Counterfeiting features	brand-related strings (e.g.,"PayPal" and "Apple ID"), sensitive words(e.g.,"login" and "bank"), visual elements (e.g., favicon and logo), brand resource request (e.g., behavior of calling image and CSS)
Stealing features	"form feature", "submission feature", "password feature", "https feature", etc.
Affiliation features	domain name registrant, name server, IP address, etc.
Evaluation features	PageRank value, Alexa ranking, etc.

Jain and Gupta especially extracted visually similar features (Jain and Gupta, 2017); and Xiang and Hong extracted login forms, such as FORM tags and INPUT tags, and named entities (Xiang and Hong, 2009).

Since all the phishing sites have these features, utilizing these two types of features helps to ensure a high recall rate. However, well-known brands usually have multiple domain names and many subbrands and even authorize third parties to use their brand elements; misclassification of these sites can cause serious losses and widespread influence. Avoiding misjudging these kinds of websites and ensuring a high precision rate are equally important. Some researchers have similar considerations, and they extract websites' affiliations or third-party evaluation information for phishing detection. For example, in addition to the two types of features mentioned above, Rao and Pais used age of domain and search engine results (Rao and Pais, 2018); Xiang and Hong used search engine results to compare domains (Xiang and Hong, 2009); Xiang et al. adopted page-in-top search results, the PageRank, etc. (Xiang et al., 2011); and Sunil and Sardana used the PageRank, domain age, etc. (Naga Venkata Sunil and Sardana, 2012).

However, the problem is that most existing works do not really analyze the extracted features and classify them into certain types, and the result is that the extracted features are arbitrary, scattered and incomplete. In other words, these works apply some of these features, hoping to have a good classification performance on their own datasets. However, the reasons why the features are selected over others, if these features are enough, what may happen if other features are also used, or whether the performance with these features on a different dataset will still be satisfactory were not considered, especially in the real web environment.

Based on the analysis above, this paper tries to construct an interpretable feature framework that can reflect the characteristics of phishing websites as fully as possible and lays a solid foundation for phishing detection aimed at the real web environment. The features related to webpage similarity are collectively referred to as counterfeiting features. Features that implement the theft function are classified as stealing features. To ensure the precision rate, the features used to identify legitimate brand websites with multiple domain names, subbrand websites and authorized websites, that is, to identify their affiliations, are called affiliation features. To further reduce the rate of false positives, features used to exclude a very small number of high-credit websites, which are often obtained from third-party website evaluation platforms, are categorized as evaluation features. To facilitate memorization, four keywords are selected, namely, "Counterfeiting", "Stealing", "Affiliation" and "Evaluation". Their first letters are selected, and the "CASE" framework is built. The CASE framework is shown in Fig. 2.

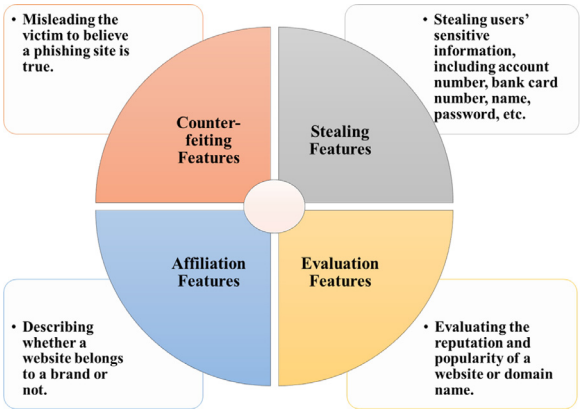

Fig. 2 – Antiphishing CASE Feature Framework.

Counterfeiting features and stealing features ensure high recall rates, avoid missing detection as much as possible and eliminate most legitimate sites. Multiscale affiliation features and evaluation features ensure high accuracy and low false detection rates. The CASE features are presented in Table 1 and described in detail as follows:

Counterfeiting features: To achieve the effect of falsification, phishing websites often contain the same or similar brand elements, similar to trusted third-party brand websites. This paper extracts a series of brand counterfeiting features, including brand-related strings, such as "PayPal", "Apple ID" and "Bank of America"; sensitive words, such as "registration", "login" and "bank"; visual elements, such as favicon and logo; and brand resource request features, such as the behavior of a calling image, CSS (cascading style sheets), and JS (JavaScript) files of the brand website through the hyperlink.

Stealing features: Stealing features refer to features associated with stealing sensitive information, such as account numbers and passwords. Stealing features include "form feature", "submission feature", "password feature", "https feature", etc. Taking the "form feature" as an example, the values are 0 and 1, indicating whether the <form> tag is included in the webpage.

Affiliation features: The abovementioned counterfeiting and stealing features provide high recall rates. However, well-known brands usually have multiple domain names, multiple subbrands, and even authorize third parties to use their brand elements. The features that distinguish such sites are called affiliation features. Affiliation features include a domain name registration time, a name server that characterizes

who provides the DNS resolution, an IP address that characterizes where the website comes from, and an autonomous system that reflects the attribution of IP address. These features can be extracted by services such as Whois and DIG.

Evaluation features: Evaluation features are mainly used to exclude a very small number of high-credit websites, aiming to further reduce the rate of false positives. Considering the adaption to the real web environment, which requires very fast detection speeds, features that need search engine results are excluded, so evaluation features refer to PageRank value, Alexa ranking, and other features that can be retrieved in advance and saved locally; that is, local data can be read directly during phishing detection.

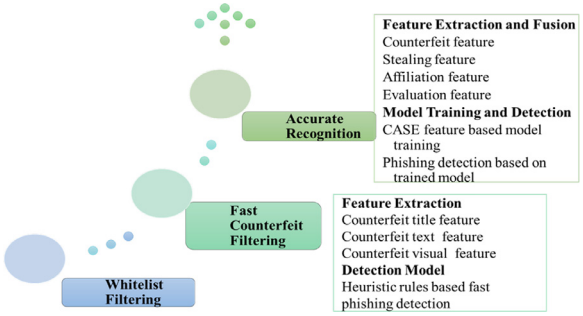

Fig. 3 – A multistage model for large-scale phishing detection scenes.

4. Proposed multistage phishing detection model

Based on the CASE framework, to adapt to large-scale phishing detection scenarios, the designed model should be able to detect phishing both quickly and accurately. In other words, the designed model needs to ensure a low false detection rate, high accuracy and a high recall rate under the premise of a short detection time. However, most research directly adopted existing classification models, such as naive Bayes, decision trees, random forest and deep learning algorithms. To be applicable to large-scale detection scenarios, they need to train a large amount of various data and ensure a fast detection speed; however, most research cannot meet this requirement (Chen et al., 2018; Geng et al., 2015; Hiransha et al., 2018; Liang et al., 2020; Wei et al., 2019; Xiang et al., 2011). Even assuming that research can collect enough and comprehensive samples, phishing is a seriously imbalanced classification problem; that is, the number of legitimate websites is much larger than the number of phishing websites. Therefore, if more legitimate websites can be filtered before machine learning algorithms are adopted, the datasets can become more balanced, and the accuracy and speed of training and detection can be improved.

Therefore, a multistage model is designed in this paper. By analyzing the traffic of the DNS recursive server (1.2.4.8) for a week[1], statistics show that the top 1,000 websites have more than 52% of the total page visits on the Internet; the top 10,000 have a higher page visit rate of 87.9%. Therefore, whitelist filtering is especially suitable for large-scale actual network environments, such as browsers, client applications, or phishing detection scenarios for large-scale user access log analysis. Therefore, the top ranked websites are designed to be excluded quickly with whitelist filtering at the first stage, which is referred to as the whitelist filtering stage.

Then, to further filter the legitimate websites quickly, lightweight features, such as sensitive words, brand-related strings, copyright notices, favicons, and brand resource calls, are used. The reason for choosing these features for rapid filtering is that during an initial analysis of a large number of samples, almost all phishing websites, especially high-quality phishing websites, were found to contain at least one or more of these features; additionally, these features can be easily extracted, which ensures the efficiency of the model. These features all belong to the counterfeiting features, so we call this stage the fast counterfeit filtering stage.

Finally, after as many legitimate websites are filtered as possible, the class imbalance problem of positive and negative samples on the actual original Internet sample set is effectively alleviated. Then, to ensure high accuracy and a low false positive rate, the classifiers learned with the multiscale CASE features. This is the third stage: the accurate recognition stage.

The multistage detection model for large-scale phishing detection scenarios is shown in Fig. 3, and more details are introduced in the following subsections.

4.1. *Whitelist filtering stage*

The whitelist data include but are not limited to the phishing targeted brand website domain name. Since, as mentioned before, the top 10,000 websites have a higher page visit rate of 87.9% and when the size of the whitelist is 10,000, this step can improve the overall performance of the method by at least 8 times. This step is the first step to ensure the efficiency of the model. The whitelist uses a hash query with a time complexity of O(1). Further expansion of the whitelist can further increase efficiency, and in practice, a balance between efficiency and missed detection is sought.

4.2. *Fast counterfeit filtering stage*

The fast counterfeit filtering stage uses only part of the counterfeiting information of the web page itself, excluding the logo data that occupy storage and bandwidth and third-party resource data, to ensure efficient filtering. Websites that do not have any counterfeiting behavior at the text, visual, and resource levels are quickly excluded. This stage ensures the recall rate of suspected phishing decisions.

It is worth mentioning that although the favicon is a picture, it is easy to extract and has a small volume; it is generally a 16x16 picture. To further improve the filtering speed, this paper adopts favicon picture matching based on the gray histogram Girod. Considering the favicon image to be processed in this paper, to make the calculation process clearer, formula

[1] The recursive DNS server has nearly 100,000 users, and the average query quantity per day is approximately 50 million.

1 is created:

$$compare(x) = \min_{k=1}^{M} \sqrt{\frac{1}{N}\sum_{i=0}^{N}(hg(x[i]) - hg(y_k[i]))^2}, \quad (1)$$

where x is the favicon to be detected, $hg(x)$ is the histogram function, y_k is the brand favicon, $N = 255$, and M is the number of brands. Based on formula 1, the favicon feature value is computed based on formula 2:

$$f(x) = \begin{cases} 0 & compare(x) \le T \\ 1 & compare(x) > T. \end{cases} \quad (2)$$

The value of T is a very small positive number close to zero. In our experiment, the value of T was chosen to be 0.0001.

In the real web environment, this stage achieves rapid filtering, effectively eliminating sites that are not phishing, and only a small percentage of "suspected phishing websites" enter the next stage. This step ensures the recall rate of suspected phishing decisions through a comprehensive analysis of the multiscale counterfeiting features.

4.3. *Accurate recognition stage*

This stage further classifies the "suspected phishing websites" with a detection model based on multiscale CASE features. This stage aims to remove the websites, including subbrand websites of phishing target brands, cobranded websites, promotional websites, and high-reputation websites with potential cooperation offline, that are prone to misjudgment as suspected phishing websites.

The model at this stage comprehensively utilizes multiscale CASE features; that is, in addition to the above lightweight features, the multiscale features further include logo-like picture features, stealing features, evaluation features, and affiliation features. The detection model used in the accurate recognition stage is not limited to a specific algorithm. AdaBoost (Rätsch et al., 2001), sequential minimal optimization (SMO) (Keerthi et al., 2014) and random forest (Liaw et al., 2002) are used in the experiments. The Adaboost algorithm is implemented by changing the data distribution, which determines the weight of each sample based on whether it is correctly classified in each training set and the accuracy of the last overall classification. The modified new dataset is sent to the next round of classifiers for training, and finally, the classifiers obtained by each training set are merged as the final decision classifier. The basic idea of SMO is to optimize only two variables in one iteration and fix the remaining variables, that is, to decompose a large optimization problem into several small optimization problems, which are often easy to solve and save time. Random forest refers to a classifier that uses multiple trees to train and predict samples. The training set used by each tree is back-sampled from the total training set, which means that some samples in the total training set may appear multiple times in a tree's training set, or they may never appear in the training set of a tree.

The multistage antiphishing model proposed in this paper significantly improves the detection efficiency compared with the traditional single-stage model. Considering that the supervised model of the accurate recognition stage is learned from more representative positive and negative samples, the detection performance is also effectively improved.

5. Comparative dataset-based experiments

To fully evaluate the effectiveness of the CASE features and the proposed multistage model, a complex dataset was first constructed. Based on the dataset, the detection ability of different feature spaces under the CASE framework is evaluated. Then, the proposed multistage phishing detection model is implemented, and comparative experiments are conducted on the constructed dataset.

The algorithms are implemented on a MacBook Pro with 12.3 GHz Double-Core Intel Core i5 CPU and 8 GB 2133 MHz LPDDR3 RAM (13-inch, 2017, Two Thunderbolt 3 ports). The CASE feature extraction and the proposed multistage model are programmed in the Java language, and Weka is used for the Adaboost, random forest and SMO implementations (Russell and Markov, 2017). The comparison experiments with a CNN and LSTM are implemented in Python with the Keras API (Gulli and Pal, 2017).

5.1. *Dataset construction*

To fully evaluate the proposed CASE features and the multistage model, a complex dataset was first constructed. It is necessary to cover as many legitimate and phishing websites as possible and include websites in different languages and with different content qualities because the quality of the dataset determines whether the model being trained can work well in the real web environment. The phishing samples were from social reports received by PhishTank (https://www.phishtank.com/) and APAC (http://www.apac.cn/); legitimate samples were obtained from DMOZ (https://dmoz-odp.org/), randomly selected DNS resolution requests, and web search results of brand names and sensitive words. The reason for selecting samples from the search engine and DNS log is to increase the confusion of the samples so that it can better reflect an actual complex network environment, increase the detection difficulty and more objectively reflect the robustness and practicability of the proposed method. In addition, considering that phishing detection is a class imbalance problem, the proportion of positive and negative samples in the dataset needs to be large enough, which further increases the difficulty of detection. Based on the discussion, a complex dataset is constructed (see Table 2), and more details about the data imbalance, sources and obfuscation are described below.

The dataset contains 3972 legitimate samples and 195 phishing samples. Generally, a dataset is balanced when each label contains the the same number of samples. There is currently no clear definition for a class imbalance, but in Ortigosa-Hernández et al. (2017), when discussing class imbalance, the sample ratio they use is 20:1. In our constructed dataset, the ratio of legitimate and phishing samples is greater than 20:1, reaching 20.37:1, and the sample imbalance ratio is much higher than the other two published datasets mentioned in Section 2.3. In dataset(Mohammad et al., 2015), the ratio is 1:1; and in dataset (Vrbančič et al., 2020), it is 1.89:1. In reality, although the number of phishing websites is large,

Table 2 – Dataset Construction.

Class	Source	Number	Description
Phishing Samples	APWG	144	Including samples of 10 brands, such as Apple and Paypal
	APAC	51	Including samples of four brands –Facebook, Alibaba, Apple and PayPal
Legitimate Samples	DMOZ	1447	Legitimate websites randomly crawled from DMOZ (https://dmoz-odp.org/) through a web crawler
	DNS Recursive Resolution Log	1954	From DNS recursive resolution service (1.2.4.8) log, including random samples and confusing samples containing brand strings
	Google Search Engine	571	Samples obtained by retrieving brand names with the Google search engine.

the number is still extremely low compared to the total number of Internet websites. Setting up a 20.37:1 ratio of positive and negative samples mimics the real-world situation, where there are more legitimate websites. Moreover, in the constructed dataset, phishing samples cover 10 different phishing brands, including Apple, Alibaba, DBS, DocuSign, Microsoft, Facebook, Bank of America, Yahoo, PayPal and Banco Inter. The languages of sample pages in the dataset include English, German, French, Irish, Spanish, Lithuanian, Norwegian, Slovenian, Croatian, Latvian, Japanese, and Chinese, which increases the data complexity and detection difficulty.

Notably, from the DNS recursive resolution service (1.2.4.8) log, there are samples containing brand strings, such as www.paypalwarning.com and us.paypal-here.com. These samples containing brand strings are very confusing and can easily be misjudged based on the phishing detection model of URL analysis. Therefore, these samples underwent further manual verification and were determined to be legitimate websites.

Legitimate samples containing brand names were obtained with the Google search engine by randomly selecting 20 results from the first 100 returned results for each brand and the duplicate URLs that were extracted through the DNS log and DMOZ were removed. These URLs include brand websites and their associated websites, so most of their URLs contain brand strings or have similarities, and their web contents often contain brand keywords and images such as logos. These websites, which include https://www.dbs.com.gs/, https://www.dbs.com.cn/, https://students.dbs.ie/, and https://secure.id.dbsdigibank.com, are obfuscated and easily misjudged as phishing.

We made the constructed dataset public, and it can be accessed at http://phishing.tvlib.net/casedata.html. This dataset not only contains the 107-dimensional CASE features described in this paper but also provides URLs, webpage source codes and webpage screenshots of all the samples. Compared with the dataset currently available on the Internet, this dataset generally has the following advantages:

- In this dataset, the ratio of legitimate and phishing samples is 20.37:1, which greatly increases the difficulty of classification and helps to mimic the real-world situation where there are many more legitimate websites.
- Legitimate samples collected through the DNS recursive resolution log and Google search engine contain many brand-related websites. Their URLs have brand similarity, and the web content contains brand keywords and pictures; hence, the websites are very obfuscated and easy to misjudge as phishing. In addition, the dataset contains webpages in more than ten different languages, which further increases the detection difficulty.
- The dataset supports users in verifying most of the 107-dimensional features based on the raw data. In fact, 104 of them can be verified, except for the Alexa ranking, IP address and domain name registration information due to the expiration of phishing websites.
- The dataset also supports users in further extracting more statistical text features, such as URL statistical features, webpage bag of word features, and visual statistical features in more dimensions and in extracting favicons and logos through image segmentation.
- The webpage screenshots and webpage HTML data provided in this dataset can also be directly used as input for an experimental verification of phishing detection models based on deep learning algorithms, such as CNNs and LSTM.

5.2. *Evaluation of the CASE framework*

In this paper, we first evaluate the detection ability of different feature spaces under the CASE framework on the above-mentioned dataset. Random forest has better comprehensive performance than other commonly used algorithms, but to make the comparison result more intuitive, the results of Adaboost and SMO are also listed. The experimental results are presented in Table 3.

Table 3 presents the experimental results of the above-mentioned machine learning algorithms in the three feature spaces of C, C+S, C+S+A+E. The reason for choosing these three feature spaces is based on the following considerations: In the CASE feature framework, counterfeiting (C) is the key feature of phishing and stealing (S) is a necessary means for phishing websites to make profits. Both of them are essential features of phishing websites, but the evaluation (E) and affiliation (A) features are mainly used to reduce false detections. Therefore, evaluation and affiliation features, as subspaces of the CASE feature framework, have no strong interpretability, except for their influence on misdetection. A more detailed discussion is presented in Section 5.4.1.

Table 3 shows the performance of the three strong classification algorithms—SMO, Adaboost and random forest—on the three feature spaces of C, C+S and C+S+A+E under the CASE framework, including the comprehensive F1-measure

Table 3 – Experimental Results on Different Feature Spaces.

Algorithm	Feature Space	Recall /TPR	FPR	Precision	F_1-Measure
AdaBoost	C	0.7385	0.0060	0.8571	0.7934
(Stump)	C+S	0.7538	0.0063	0.8547	0.8011
	C+S+A+E	0.8872	0.0013	0.9719	0.9276
SMO	C	0.7333	0.0045	0.8882	0.8034
	C+S	0.7590	0.0048	0.8862	0.8177
	C+S+A+E	0.8051	0.0030	0.9290	0.8626
Random	C	0.8872	0.0050	0.8964	0.8918
Forest	C+S	0.8872	0.0023	0.9505	0.9178
	C+S+A+E	**0.8923**	**0.0005**	**0.9886**	**0.9380**

C: Counterfeiting features, A: Affiliation features, S: Stealing features, E: Evaluation features Model: Random Forest

and the single precision values, which all become sequentially better. This is in line with our expectations that phishing websites gain user trust through "counterfeiting (C)" and then "steal (S)" users' private and sensitive information, and A+E can effectively reduce the false detection rate. The random forest learning algorithm on CASE achieved the best results, and the comprehensive F1-measure was 0.938. This proves the high quality of the multiscale CASE features for phishing detection.

5.3. *Evaluation of the proposed multistage model*

The proposed multistage phishing detection model is implemented, and comparative experiments with one-stage detection models are conducted on the dataset.

5.3.1. *Whitelist filtering stage*

The whitelist we used in this experiment contains the second-level domain names of 10 brand websites. It is worth mentioning that the whitelist is extensible, and localized domain names such as 'amazon.co.uk' can also be included in real applications. The experiment on this dataset uses a small whitelist to demonstrate the role of the module. After this filtering stage, 23 normal samples in the training set are correctly filtered. The reason 23 instead of 10 URLs are filtered is that a brand URL uses a third-level domain name. There are 4,144 samples input in the second stage.

5.3.2. *Fast counterfeit filtering stage*

In this stage, the number of features extracted in this paper is 82, including brand string features, sensitive word features, favicon features, copyright features and resource request features. These features are extracted from the web page itself and do not depend on any third-party data, ensuring fast feature extraction. To ensure the recall rate, the decision rule at this stage is "if the detected web page contains any of the 82-dimensional features, it is input into the next stage as a suspected phishing website; otherwise, it is considered legitimate."

After the counterfeit filtering stage, 2,844 of the 4,144 samples were correctly classified as nonphishing, and 1,300 samples were identified as suspected phishing and were input into the next stage. This result also proves the brand counterfeiting property of phishing websites.

5.3.3. *Accurate recognition stage*

In this stage, the 107-dimensional multiscale CASE features are comprehensively utilized. In addition to the 82-dimensional features extracted in the previous stage, the logo feature and the sensitive word features extracted from the body are further included; that is, a total of 99-dimensional counterfeiting features are extracted. The 4-dimensional stealing features, including the form feature, the submit feature, the password feature, and the https feature, are extracted. In addition, the website IP address attribution and domain name registration time feature are extracted as affiliation features. The 2-dimensional evaluation features are extracted and include redirection features and Alexa ranking features.

In this paper, the random forest, SMO and AdaBoost classification algorithms are used. The weak classification of AdaBoost uses a decision tree stump, and SMO adopts a polynomial kernel. In particular, instead of training the classifier on the original training set, the classification model is trained on the dataset after the fast counterfeit filtering stage; that is, the model is trained on 1,105 normal samples and 195 phishing samples for model learning. Among the three different classification models, the random forest algorithm achieved the best results, and 184 phishing samples and 1,103 normal samples were correctly classified (see Table 4).

From Table 4, the proposed multistage model on the CASE feature framework with random forest achieved the best results, and its F_1-measure was 0.9659. This table also shows the experimental results of single-stage detection directly on the original dataset of 4,167 samples without filtering stages and further compares the results with a variety of existing phishing detection methods, including phishing detection methods based on classic machine learning methods (Geng et al., 2015; Xiang et al., 2011) and deep learning methods (Chen et al., 2018; Hiransha et al., 2018; Liang et al., 2020; Wei et al., 2019). Among them, CANTINA+ is a feature-rich machine-learning framework for detecting phishing websites and includes 15 different features(Xiang et al., 2011). Geng et al. (2015) proposed a brand identity and authorization feature-based phishing detection method, and bagging was used as the classifier.

In this table, it is noticed that classic machine learning algorithms with CASE features perform better than some deep learning-based methods. The possible reasons are as follows: first, the datasets constructed in this paper are very complex

Table 4 – Comparative Results between the Single- and Multi-Stage Models.

Model	Algorithm	Recall /TPR	FPR	Precision	F_1-Measure
One Stage Model	Adaboost(CASE)	0.8872	0.0013	0.9719	0.9276
	SMO(CASE)	0.8051	0.0030	0.9290	0.8626
	Random Forest (CASE)	0.8923	0.0005	0.9886	0.9380
	Bagging (C4.5) (Geng et al., 2015)	0.8462	0.0053	0.8871	0.8661
	Bayesian Network (Xiang et al., 2011)	0.7641	0.004	0.9030	0.8278
	CNN (Wei et al., 2019)	0.8308	0.0103	0.7983	0.8308
	CNN (Hiransha et al., 2018)	0.7641	0.002	0.9524	0.8436
	LSTM (Chen et al., 2018; Liang et al., 2020)	0.7333	0.0081	0.8251	0.7698
The Proposed Multistage Model	Adaboost(CASE)	0.9333	0.0018	0.9630	0.9479
	SMO(CASE)	0.8615	0.0041	0.9130	0.8865
	Random Forest (CASE)	**0.9436**	**0.0005**	**0.9892**	**0.9659**

and unbalanced; second, simply with the URL or HTML input, CNNs or LSTM cannot mine enough identifying information to support effective phishing site detection.

Moreover, the multistage phishing detection model with CASE features proposed in this paper is superior to the existing single-stage detection method. The core reason is that, based on the property of phishing attacks, the brand counterfeiting features are prioritized; that is, the samples with no signs of counterfeiting are preferentially eliminated. Then, the accurate recognition stage is implemented on a more targeted and balanced sample set because the remaining positive and negative samples are considered "suspected phishing". Hence, the learnt model is more robust.

5.4. *Further discussion on the proposed multistage model*

5.4.1. *Feature selection*

In the comparative experiments on different feature spaces, we find an interesting phenomenon. Although the comprehensive index F1-Measure achieves the optimal value (0.938) in the C+A+S+E feature space, the TPR value of the random forest (RF) model in the C+A feature space is 0.923, which is higher than the 0.8923 achieved in the C+S+A+E feature space. After the analysis, the RF (C+A) model correctly identified 7 phishing websites that were missed by the RF (C+S+A+E) model. These websites are characterized by "counterfeiting features" and are "newly registered domain names", which is easy to understand because most phishing sites have a very short life cycle and look similar to legitimate ones. However, the price of RF (C+A)'s high TPR is that 8 legitimate websites that are correctly identified in the C+S+A+E space are incorrectly judged as phishing in C+A space, resulting in a higher FPR (0.015), which is 3 times the FPR in C+S+A+E space. In other words, websites with "counterfeiting features" and "newly registered domain names" are easily classified as phishing sites by the RF (C+A) model, including both the real phishing sites and some recently registered legitimate brand-related sites, and its high TPR comes at the cost of high FPR, which is disastrous in actual detection because a misjudgment can bring great losses to legitimate websites. Although C+A has no special significance, this phenomenon inspires us to perform feature selection on the 107-dimensional joint features in the CASE feature framework and find a feature subspace with better comprehensive performance.

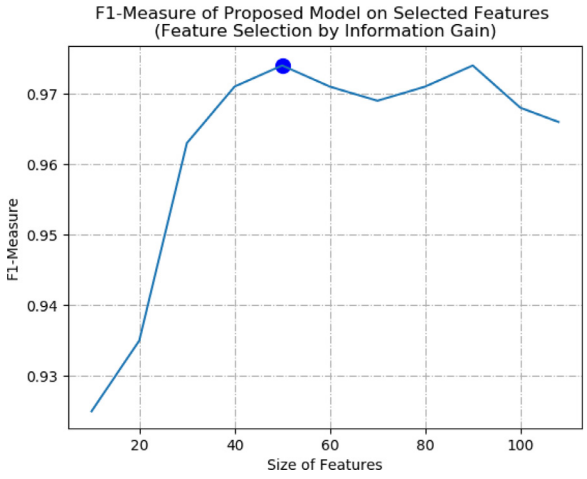

Fig. 4 – Model performance based on different dimensional features.

The feature selection experiment is carried out with the information gain algorithm Information gain and mutual information for machine learning. Based on the selected features of different dimensions, the performance evaluation is performed by random forest. The experimental results are shown in Fig. 4.

The horizontal axis of Figure 4 is the feature dimension, and the vertical axis is the F1-Measure value. It can be seen from the figure that there is some redundancy among the extracted 108-dimensional features. The best result is 0.974, which is obtained with the 50-dimensional features. The 50-dimensional features include 44-dimensional counterfeit features, 2-dimensional stealing features, 2-dimensional affiliation features, and 2-dimensional evaluation features. This also proves that the extracted multiscale features under the framework of CASE complement each other.

5.4.2. *Discussion on the accurate recognition stage*

To determine how effective an additional fusion of deep semantic features may be in the accurate recognition stage, we referred to Zhang et al.'s work (Zhang et al., 2017) and extracted the word2vec features of all words on a webpage. The dimension of the word vector was set as 128, and then the word vector features were further mapped to the document

vector with the arithmetic mean method in Zhang et al. (2017). The obtained deep semantic document vector was also 128-dimensional. In the accurate recognition stage, if the model is based only on the 128-dimensional semantic features, the performance of random forest is only 0.8966, which is not satisfactory. Then, we also tried to linearly fuse the 128-dimensional deep semantic feature and the 107-dimensional CASE feature extracted in this paper, and the classification algorithms were trained on the new feature space for phishing detection. Among them, Adaboost performed the best with 700 iterations. Its F_1 value reached 0.9716, which exceeds the optimal value when deep semantic features are not considered. However, the performance of random forest was not as good when only CASE features were used; that is, performance degradation occurs, which implies that deep semantic features and CASE features interfere with each other in these two classifiers. In short, the word2vec feature has a certain discrimination ability, but determining how to effectively integrate it with the CASE features proposed in this article requires an in-depth follow-up study.

5.4.3. *Model efficiency analysis*

There were 2,867 of the 4,167 samples filtered in the first two stages, and less than 32% of the samples are needed to extract a series of heavyweight features in the accurate recognition stage, which includes downloading images and extracting logo comparative features, third-party service-dependent Alexa ranking features and domain name registration features. Compared to the extraction of the two types of heavyweight features, web page acquisition and page-based feature extraction in the fast counterfeit filtering stage is very quick and efficient. This shows that the overall detection efficiency increased by approximately 60% on this dataset. In the real web environment, the efficiency is improved by more than 60%. The reasons are as follows: a. The dataset contains multiple obfuscated samples obtained from search engines (i.e., these samples are input into the accurate recognition stage); however, the proportion of confusing samples in the real web environment is smaller. b. A very short whitelist containing 10 domain names is used, but in practice, the size of the whitelists used can reach tens of thousands or even more. The proactive phishing discovery experiment in the real web environment in the following section demonstrates this.

6. Phishing discovery in the real web environment

To test the efficiency of the multistage model proposed in this paper in a real web environment, a monthly report experiment that lasted for one year was carried out. A local domain name recursive resolution service log was selected as the data source. The service serves approximately 100 thousand Chinese users. The whitelist is collected from DMOZ (http://www.dmoz-odp.org/) and Baidu (http://site.baidu.com/ and contains 13,276 domain names. Considering that the uptime of phishing websites is very short, to save bandwidth and computing resources, a temporary whitelist in addition to the fixed whitelist was also used. The temporary whitelist contained three days' worth of host names that were recognized as nonphishing websites. The period of three days was chosen because statistics show that the life cycle of most domestic phishing websites in China is less than three days (CNNIC, 2016).

From Jan. 1st, 2018 to Dec. 31st, 2018, 3,415 phishing sites covering 67 brands were newly discovered. The number of phishing sites detected per month is shown in Fig. 5. The largest number of phishing websites occurred in November, and the potential cause was the Double Eleven Shopping Spree.

Of the 67 detected phishing brands, detailed information on more than 5 websites and their brands is shown in Table 5. All the discovered sites were reported to APAC and were disposed of in a timely manner.

Table 5 – Information about the Newly Discovered Phishing Sites.

Brand Name	Number of Phishes	Type
China Mobile	883	Mobile communication
Agricultural Bank of China	696	Electronic banking
MI.COM	132	E-Commerce
Bank of China	86	Electronic banking
Bank of Communication	84	Electronic banking
CITIC Securities	68	Securities investment
ICBC	23	Electronic banking
Facebook	19	Social media
China Construction Bank	16	Electronic banking
China Merchants Bank	15	Electronic banking
Taobao	15	E-Commerce
Apple	13	Mobile communication
China Everbright Bank	10	Electronic banking
People's Bank of China	9	Electronic banking
Merchants Securities	9	Electronic banking
Ping An Bank	9	Electronic banking
WeChat	9	Mobile communication
Credit Information Center	6	Credit investigation service
Halifax	6	Electronic banking

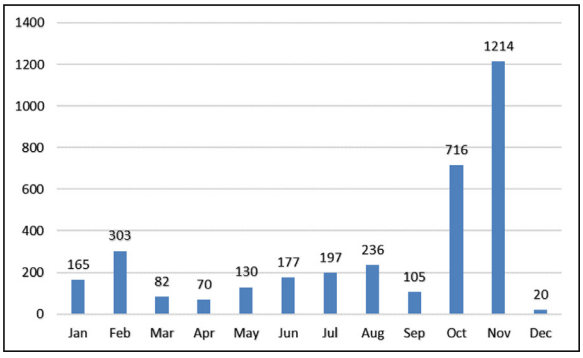

Fig. 5 – Monthly statistics on the number of phishing sites detected.

The domain name recursive resolution service used in this paper has a small user scale, the daily resolution is only approximately 1 billion, and the host after deduplication is less than 3 million. According to the statistics, after the temporary whitelist is fully implemented, approximately 98.7% of the queries are excluded in the first stage. After the second stage, only approximately 0.05% of the suspicious hosts enter the third stage. This means that an ordinary server can be used to complete detection, and the method is practical and efficient.

According to the statistics released by Google, Google 8.8.8.8 and 8.8.4.4 provide DNS resolution to 10% of the world's Internet users. As one of the largest public DNS resolution services, hundreds of millions of users generate more than 1 trillion queries every day. It is conceivable that if the method proposed in this paper is based on large DNS resolution data similar to Google DNS, the discovery capability of phishing websites will be greatly improved.

7. Conclusion and future work

This paper analyzes the pattern of phishing attacks and proposes the antiphishing feature framework—CASE—which depicts the features of phishing from four different perspectives. Then, a robust multistage phishing detection model is proposed. On a constructed complex dataset, comparative experiments between single-scale and CASE features and between single- and multistage models were carried out. Further discussions on feature selection, accurate recognition stage and model efficiency are also made. The experimental results prove the effectiveness of both the CASE framework and the multistage model with CASE in improving phishing detection performance. Finally, this method was implemented on the real web environment to prove its efficiency in reality.

Future work may include the following: 1) In terms of feature extraction, the feature dimensions under the CASE framework can be extended, for example, extending affiliation features by extracting domain name registrants and domain name resolution server features and extending visual counterfeiting features by extracting web page screenshots (Goldberg and Levy, 2014); 2) In terms of the detection model, further exploration of the model layer fusion strategy, such as visual similarity detection based on deep learning (LeCun et al., 2015), to facilitate different submodels for different scale features is necessary; 3) In terms of the experimental verification, building a larger dataset containing more brands and more languages is necessary and can effectively support the validation of novel detection models.

Declaration of Competing Interest

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

CRediT authorship contribution statement

Dong-Jie Liu: Data curation, Writing – original draft. **Guang-Gang Geng:** Methodology, Writing – review & editing. **Xiao-Bo Jin:** Investigation. **Wei Wang:** Supervision.

Acknowledgement

This research was supported by the National Key R&D Plan of China under grant No. 2018YFB1003701 and the Natural Science Foundation of Guangdong Province under grant No. 2021A1515011314.

REFERENCES

Aggarwal S, Kumar V, Sudarsan S. Identification and detection of phishing emails using natural language processing techniques. In: Proceedings of the 7th International Conference on Security of Information and Networks; 2014. p. 217–22.

Akinyelu AA, Adewumi AO. Classification of phishing email using random forest machine learning technique. J. Appl. Math. 2014;2014.

Ali S, Shahbaz M, Jamil K. Entropy-based feature selection classification approach for detecting phishing websites. In: 2019 13th International Conference on Open Source Systems and Technologies (ICOSST); 2019. p. 1–6.

Apwg phishing attack trends reports list, http://www.apwg.org/resources/apwg-reports/.

Arachchilage NAG, Love S, Beznosov K. Phishing threat avoidance behaviour: an empirical investigation. Comput. Hum. Behav. 2016;60:185–97.

Bahnsen AC, Bohorquez EC, Villegas S, Vargas J, González FA. Classifying phishing urls using recurrent neural networks. In: 2017 APWG symposium on electronic crime research (eCrime). IEEE; 2017. p. 1–8.

Basit A, Zafar M, Liu X, Javed AR, Jalil Z, Kifayat K. A comprehensive survey of ai-enabled phishing attacks detection techniques. Telecommun. Syst. 2020:1–16.

Basnet RB, Sung AH, Liu Q. Feature selection for improved phishing detection. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer; 2012. p. 252–61.

Bilge L, Kirda E, Kruegel C, Balduzzi M. Exposure: Finding malicious domains using passive DNS analysis.. In: Ndss; 2011. p. 1–17.

Castao, F., Fidalgo, E., Alegre, E., Chaves, D., Sanchez-Paniagua, M., 2021. State of the art: Content-based and hybrid phishing detection.
Chen W, Zhang W, Su Y. Phishing detection research based on LSTM recurrent neural network. International Conference of Pioneering Computer Scientists, Engineers and Educators(ICPCSEE 2018), 2018.
Chiew KL, Chang EH, Tiong WK, et al. Utilisation of website logo for phishing detection. Comput. Secur. 2015;54:16–26.
CNNIC, 2016. Global chinese phishing sites report.
Consumer data under attack: The growing threat of cyber crime : https://www2.deloitte.com/tr/en/pages/risk/articles/consumer-data-under-attack.html.
Dhamija R, Tygar JD. The battle against phishing: dynamic security skins. In: Proceedings of the 2005 symposium on Usable privacy and security; 2005. p. 77–88.
Feroz MN, Mengel S. Phishing url detection using url ranking. In: 2015 ieee international congress on big data. IEEE; 2015. p. 635–8.
Garera S, Provos N, Chew M, Rubin AD. A framework for detection and measurement of phishing attacks. In: Proceedings of the 2007 ACM workshop on Recurring malcode; 2007. p. 1–8.
Geng G-G, Lee X-D, Wang W, Tseng S-S. Favicon-a clue to phishing sites detection. In: 2013 APWG eCrime Researchers Summit. IEEE; 2013. p. 1–10.
Geng G-G, Lee X-D, Zhang Y-M. Combating phishing attacks via brand identity and authorization features. Secur. Commun. Netw. 2015;8(6):888–98.
Girod, B.,. Digital image processing. https://web.stanford.edu/class/ee368/Handouts/Lectures/2014_Spring/Combined_Slides/4-Histograms-Combined.pdf. Stanford University, 2013.
Global phishing survey:trends and domain name usein 2h2014 http://docs.apwg.org/reports/APWG_Global_Phishing_Report_2H_2014.pdf.
Goldberg Y, Levy O. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. Comput. Sci. 2014.
Gulli A, Pal S. Deep learning with Keras. Packt Publishing Ltd; 2017.
Gupta R. Comparison of classification algorithms to detect phishing web pages using feature selection and extraction. Int. J. Res. - GRANTHAALAYAH 2016;4(8):118–35.
Hiransha M, Unnithan NA, Vinayakumar R, Soman K, Verma A. Deep learning based phishing e-mail detection. Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal.(IWSPA), 2018.
Information gain and mutual information for machine learning, https://machinelearningmastery.com/information-gain-and-mutual-information/.
Jain AK, Gupta B. Towards detection of phishing websites on client-side using machine learning based approach. Telecommun. Syst. 2018.
Jain AK, Gupta BB. Phishing detection: analysis of visual similarity based approaches. Secur. Commun. Netw. 2017;2017.
Kang LC, Chang EH, Sze SN, Wei KT. Utilisation of website logo for phishing detection. Comput. Secur. 2015;54(OCT.):16–26.
Keerthi SS, Shevade SK, Bhattacharyya C, Murthy K. Improvements to platt's smo algorithm for svm classifier design. Neural Comput. 2014;13(3):637–49.
Le A, Markopoulou A, Faloutsos M. Phishdef: Url names say it all. In: 2011 Proceedings IEEE INFOCOM. IEEE; 2011. p. 191–5.
LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
Liang Y, Deng J, Cui B. Bidirectional LSTM: An Innovative Approach for Phishing URL Identification; 2020. p. 326–37.
Liaw A, Wiener M, et al. Classification and regression by randomforest. R News 2002;2(3):18–22.
Likarish P, Jung E, Dunbar D, Hansen TE, Hourcade JP. B-apt: bayesian anti-phishing toolbar. In: 2008 IEEE International Conference on Communications. IEEE; 2008. p. 1745–9.
Ma J, Saul LK, Savage S, Voelker GM. Beyond blacklists: learning to detect malicious web sites from suspicious urls. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining; 2009. p. 1245–54.
Maurer M-E, Herzner D. Using visual website similarity for phishing detection and reporting. In: CHI'12 extended abstracts on human factors in computing systems; 2012. p. 1625–30.
Moghimi M, Varjani AY. New rule-based phishing detection method. Expert Syst. Appl. 2016;53:231–42.
Mohammad, R., Thabtah, F. A., McCluskey, T., 2015. Phishing websites dataset.
Moradpoor N, Clavie B, Buchanan B. Employing machine learning techniques for detection and classification of phishing emails. In: 2017 Computing Conference. IEEE; 2017. p. 149–56.
Naga Venkata Sunil A, Sardana A. A pagerank based detection technique for phishing web sites. In: 2012 IEEE Symposium on Computers Informatics (ISCI); 2012. p. 58–63.
Oliveira D, Rocha H, Yang H, Ellis D, Dommaraju S, Muradoglu M, Weir D, Soliman A, Lin T, Ebner N. Dissecting spear phishing emails for older vs young adults: on the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems; 2017. p. 6412–24.
Ortigosa-Hernández J, Inza I, Lozano JA. Measuring the class-imbalance extent of multi-class problems. Pattern Recognit. Lett. 2017;98:32–8.
Pandey PK, Singh SK. Phishing diagnosis: a multi-feature decision tree-based method. Int. J. Eng. Adv. Technol. 2019;9:4353–9.
Phishing attacks increase 350 percent amid covid-19 quarantine https://www.pcmag.com/news/phishing-attacks-increase-350-percent-amid-covid-19-quarantine. Published March 30, 2020.
Phishing statistics: What every business needs to know, https://blog.dashlane.com/phishing-statistics/.
Popular phishing techniques used by hackers, https://www.hackingloops.com/popular-phishing-techniques-used-by-hackers/.
Rao R, Pais A. Detection of phishing websites using an efficient feature-based machine learning framework. Neural Comput. Appl. 2018;31:3851–73.
Rätsch G, Onoda T, Müller K-R. Soft margins for adaboost. Mach. Learn. 2001;42(3):287–320.
Russell I, Markov Z. An introduction to the weka data mining system. Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, 2017. 742–742
Sahingoz OK, Buber E, Demir O, Diri B. Machine learning based phishing detection from URLs. Expert Syst. Appl. 2019;117:345–57.
Subasi A, Molah E, Almkallawi F, Chaudhery TJ. Intelligent phishing website detection using random forest classifier. In: 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA). IEEE; 2017. p. 1–5.
Tajaddodianfar F, Stokes J, Gururajan A. Texception: a character/word-level deep learning model for phishing URL detection. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing; 2020. p. 2857–61.
Vrbančič G, Fister Jr I, Podgorelec V. Datasets for phishing websites detection. Data Brief 2020;33:106438.
Wang G, Liu H, Becerra S, Wang K, Belongie SJ, Shacham H, Savage S. Verilogo: proactive phishing detection via logo recognition. Department of Computer Science and Engineering, University of California; 2011.
Wei B, Hamad R, Yang L, He X, Wang H, Gao B, Woo WL. A

deep-learning-driven light-weight phishing detection sensor. Sensors 2019;19:4258.
Xiang G, Hong J, Rose CP, Cranor L. Cantina+ a feature-rich machine learning framework for detecting phishing web sites. ACM Trans. Inf. Syst. Secur. (TISSEC) 2011;14(2):1–28.
Xiang G, Hong JI. A hybrid phish detection approach by identity discovery and keywords retrieval. In: Proceedings of the 18th international conference on World wide web; 2009. p. 571–80.
Zhang Q, Bu Y, Chen B, Zhang S, Lu X. Research on phishing webpage detection technology based on CNN-BiLSTM algorithm. J. Phys. 2021;1738(1):012131.
Zhang X, Zeng Y, Jin X-B, Yan Z-W, Geng G-G. Boosting the phishing detection performance by semantic analysis. In: 2017 ieee international conference on big data (big data). IEEE; 2017. p. 1063–70.

Dong-Jie Liu is currently working toward the Ph.D. degree with Computer Network Information Center, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China. Her research interest includes machine learning, network security and blockchain.

Guang-Gang Geng received his Ph.D. degree from the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a professor with the College of Cyber Security, Jinan University, Guangzhou. His current research interest include machine learning, web abuse detection and web search.

Xiao-Bo Jin received the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009. He is currently an associate professor with Department of Intelligent Science, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China. His current research interests include web mining and machine learning, pattern recognition, and neurocomputing.

Wei Wang received the Ph.D. degree in Nankai University, Tianjin, China. He once worked in CNNIC and Google and is currently a professor in Computer Network Information Center, Chinese Academy of Sciences, Beijing, China. His research interest includes domain name and blockchain.