



Hybrid phishing detection using joint visual and textual identity

Colin Choon Lin Tan^a, Kang Leng Chiew^{b,*}, Kelvin S.C. Yong^a, Yakub Sebastian^c,
Joel Chia Ming Than^a, Wei King Tiong^b

^a Faculty of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, Jalan Simpang Tiga, 93350 Kuching, Sarawak, Malaysia

^b Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

^c College of Engineering, IT and Environment, Charles Darwin University, Ellengowan Drive, Casuarina, NT 0810, Australia

ARTICLE INFO

Keywords:

Phishing detection
Web security
Website identity
Computer vision
Logo detection
Brand names

ABSTRACT

In recent years, phishing attacks have evolved considerably, causing existing adversarial features that were widely utilised for detecting phishing websites to become less discriminative. These developments have fuelled growing interests among security researchers towards an anti-phishing strategy known as the identity-based detection technique. Identity-based detection techniques have consistently achieved high true positive rates in a rapidly changing phishing landscape, owing to its capitalisation on fundamental brand identity relations that are inherent in most legitimate webpages. However, existing identity-based techniques often suffer higher false positive rates due to complexities and challenges in establishing the webpage's brand identity. To close the existing performance gap, this paper proposes a new hybrid identity-based phishing detection technique that leverages webpage visual and textual identity. Extending earlier anti-phishing work based on the website logo as visual identity, our method incorporates novel image features that mimic human vision to enhance the logo detection accuracy. The proposed hybrid technique integrates the visual identity with a textual identity, namely, brand-specific keywords derived from the webpage content using textual analysis methods. We empirically demonstrated on multiple benchmark datasets that this joint visual-textual identity detection approach significantly improves phishing detection performance with an overall accuracy of 98.6%. Benchmarking results against an existing technique showed comparable true positive rates and a reduction of up to 3.4% in false positive rates, thus affirming our objective of reducing the misclassification of legitimate webpages without sacrificing the phishing detection performance. The proposed hybrid identity-based technique is proven to be a significant and practical contribution that will enrich the anti-phishing community with improved defence strategies against rapidly evolving phishing schemes.

1. Introduction

Phishing is any criminal act of using genuine-looking websites to deceive users into disclosing confidential information such as account passwords, credit card numbers, etc. A typical phishing website is constructed by replicating visual cues and textual content of an existing legitimate website (Aleroud & Zhou, 2017; Chiew, Yong, & Tan, 2018). Recent statistics indicated that phishing attacks are still blooming, where an average of 54,924 unique phishing websites were reported monthly between January and March 2020 (Anti-Phishing Working Group, 2020).

Studies have shown that manually recognising phishing websites can be difficult even for users who previously underwent anti-phishing education training, therefore users should not be expected to perform this risky task on their own (Alsharnouby et al., 2015; Arachchilage

et al., 2016; Ubung et al., 2019). As such, blacklist-based automated detection mechanisms have been built into mainstream browsers to assist in warning users when known phishing websites are loaded, while unknown or zero-day phishing websites are still able to slip past the detection. Despite the limitations of blacklist-based detection, they remain commercially successful owing to its lightweight computational overhead and low rate of misclassifying legitimate websites (da Silva et al., 2020; Gupta et al., 2017; Yang et al., 2021). Therefore, in designing practical phishing detection solutions, achieving a low rate of misclassifying legitimate websites is an important requirement (Gupta et al., 2017). Anti-phishing researchers have been focusing more on enhancing the phishing detection rate, while adverse issues that impact the detection of legitimate webpages remain largely unattended. To address this performance gap, we attempt to study the possible

* Corresponding author.

E-mail addresses: ctan@swinburne.edu.my (C.C.L. Tan), klchiew@unimas.my (K.L. Chiew), kscyong@swinburne.edu.my (K.S.C. Yong), yakub.sebastian@cdu.edu.au (Y. Sebastian), jcmthan@swinburne.edu.my (J.C.M. Than), wktiong@unimas.my (W.K. Tiong).

<https://doi.org/10.1016/j.eswa.2023.119723>

Received 20 July 2022; Received in revised form 8 December 2022; Accepted 19 February 2023

Available online 23 February 2023

0957-4174/© 2023 Elsevier Ltd. All rights reserved.

improvement of one particular type of phishing detection technique called the *identity-based technique*.

Identity-based phishing detection techniques ascertain whether the portrayed identity (e.g., brand name, logo) presented in the webpage content corresponds to the actual page identity observed through the URL (Jain & Gupta, 2021). This is typically done by querying the search engine using the website logo or brand keywords to retrieve search results that can pinpoint the brand's original website. Since identity-based techniques do not depend on dynamically evolving adversarial patterns in the phishing webpage, it has consistently achieved superior performance in terms of phishing identification (i.e., high true positive rate) (Rao & Pais, 2019).

Nevertheless, identity-based techniques are more susceptible to false positives due to complications and challenges in extracting the website's portrayed identity and obtaining relevant search results for identity verification (Jain & Gupta, 2018; Rao & Pais, 2019; Van Dooremaal et al., 2021). Due to the diverse presentation styles and visual layout of webpages, it is challenging to derive efficient techniques that can reliably detect and extract previously unseen graphical brand elements such as the website logo (Gupta & Jain, 2020). Although state of the art machine learning techniques such as deep learning can be used to detect the website logo (Bozkir & Aydos, 2020), its detection capability is primarily limited to previously trained logos, thus making it less practical for phishing detection applications. Meanwhile, other techniques that strive to detect previously unseen logos in webpages are lacking the ability to model discriminative features utilised by human vision, thus limiting the logo detection accuracy (Chiew et al., 2015). On the other hand, existing identity-based techniques relying solely on textual identity may not yield accurate brand keywords as the portrayed identity when employed against webpages with limited textual content or contain mainly non-ASCII text (Tan et al., 2016).

In this paper, we advance the state-of-the-art of identity-based phishing detection techniques by enhancing the detection and extraction accuracy of the portrayed identities, as well as integrating visual and textual identities in a hybrid framework to improve the efficacy of website identity verification. Through a series of experiments, we empirically demonstrated that our method promises greater practicality and robustness than singular identity-based techniques. The main contributions from this research include the following:

- (a) Novel features that mimic human vision are proposed to improve the detection and extraction of the website logo, which is vital for phishing detection techniques utilising visual identities. For example, one of the proposed features is colourfulness, which has yet to be capitalised in existing techniques (Chiew et al., 2015). We argue that colourfulness is an essential consideration that guides humans in performing logo identification in webpages and, hence must be included in any computer vision-based system for logo detection (Van Dooremaal et al., 2021).
- (b) The proposed technique advances the emerging body of knowledge on the hybridisation of visual and textual identities for phishing detection. As evidenced in Van Dooremaal et al. (2021), the hybridisation of visual and textual identities is one of the emerging research directions aimed at overcoming the existing performance limitations. Therefore, our work establishes the foundation for other anti-phishing researchers to further develop improved methods based on hybrid identities.
- (c) The proposed hybrid identity-based technique achieves a lower rate of misclassifying legitimate webpages without sacrificing the high detection rate for phishing webpages. This performance improvement is attributed to the proposed technique's capability in adapting to a wider range of websites with varying design quality which can complicate the extraction of brand identity elements such as website logos and keywords. To achieve such level of robustness, the proposed technique integrates complementary visual and textual identity discovery components that work hand-in-hand to establish the webpage identity and achieve lower false positives.

The remainder of this paper is organised as follows: Section 2 reviews and discusses identity-based phishing detection techniques. Section 3 puts forward the proposed method. In Section 4, we describe the experimental setup, present the results, and discuss our research findings. Finally, Section 5 concludes the paper and suggests some future research directions.

2. Related work

Based on the scope of our research, the literature presented in this section focuses on identity-based phishing detection techniques. Identity-based techniques for phishing detection can be accomplished using textual or visual brand identities. We first discuss techniques that utilise textual brand identities, followed by visual brand identities and a combination of both.

In regular web browsing, users that intend to visit a particular website but are unsure of its actual URL will normally engage the help of a search engine. Tan et al. (2016) leveraged the very same concept to find the target domain by querying the search engine with identity keywords. A weighted URL tokens system was proposed, which utilises the structure of URLs in a webpage to extract identity keywords in the form of single keywords (uni-gram) or multiple coexisting keywords (multi-gram). These keywords are then searched using a search engine, where the target domain is derived from the search results. Based on experimental results, the proposed technique achieved competitive true positive and true negative rates of 99.68% and 92.52%, respectively. Although the proposed technique can also work on non-English webpages, it is still unable to accommodate webpage textual content written in non-ASCII languages.

Rao and Pais (2019) assume the domain name and page title as textual brand descriptors. They proposed an identity-based method by formulating dynamic search queries using the domain name and page title before feeding them to the search engine. If the potential target website is not found in the search results, the suspicious webpage will be declared as phishing. Otherwise, a similarity measure between the suspicious webpage and the potential target website will be used to determine the page's legitimacy. The proposed method achieved a true positive rate of 97.77% and a promising true negative rate of 99.36%. In an earlier technique by Jain and Gupta (2018), the identity verification process was also accomplished similarly. However, false negative detection may occur if the phishing webpage utilises mostly internal hyperlinks that point to the same phishing domain.

In the work of Peng et al. (2019), the behaviour and properties of phishing websites were studied in-depth. One of their analysis involved finding the target brand for 1500 samples of phishing webpages. Their analysis is carried out by extracting visible text from the webpage screenshot using optical character recognition (OCR) techniques. The less important terms were then filtered from the extracted text using a text-mining algorithm called RAKE (Rapid Automatic Keyword Extraction). Lastly, the remaining terms were searched on Google, and the top result is taken as the target brand. The authors observed that the target for phishing webpages targeting popular brands such as PayPal and Microsoft can be resolved with accuracy up to 99.8%, while the accuracy dropped to 88% for less popular brands. If actual legitimate webpage samples are tested using this technique, the less popular brands will suffer misclassifications as well. As such, more studies in the area of website identity verification is desirable to close the current performance gap.

A more recent identity-based anti-phishing approach came from Liu and Fu (2020). Given a query webpage, hyperlinks were obtained from the HTML source and used to retrieve the corresponding linked webpages. Based on this initial pool of webpages, the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm was employed to extract keywords, which were then submitted to Google Search to retrieve webpages listed in the top-10 search results. This expanded set of webpages and their corresponding page linking structure were processed

using a web graph model for phishing classification. Unfortunately, the effectiveness of the proposed technique seems to be doubtful, where only 83.98% of true positive rate was achieved.

Haruta et al. (2017) proposed an anti-phishing technique based on visual similarities of webpage content. A whitelist database of legitimate websites was constructed as reference profiles, which contains the screenshots, Cascading Style Sheets (CSS) rules, and the domain name of legitimate websites. To test whether a query webpage is phishing, a screenshot is taken and searched through the stored visual profiles to determine how similar they are. The CSS is also compared to the ones in the whitelist to determine whether they are identical. By utilising CSS matching, the proposed technique claimed to suppress the false negatives incurred for screenshots that failed visual similarity matching due to subtle differences in appearance. Experimental results suggest that the proposed method achieves only 78% of true positive rate. In another work, Mao et al. (2017) utilised CSS similarity alone and obtained better results with 97.9% and 100% of true positive and true negative rates, respectively. Mishra and Gupta (2018) proposed a more efficient method to perform similarity analysis. The researchers introduced a preliminary filter that performs full and partial string matching on the whitelisted domains before computing the CSS similarity level. The proposed method of Mishra and Gupta (2018) achieved true positive and true negative rates of 93.27% and 100%, respectively. Nevertheless, keeping the database up-to-date remains a challenge, underlining the common downside of any phishing detection technique that employs similarity analysis.

More recently, Dalgic et al. (2018) proposed a multi-level representation of visual features for webpage visual similarity computation. In addition to a standalone feature vector representing the full screenshot, the technique includes features computed from finer partitions of the screenshot. Results showed that the best performance was achieved when using Scalable Colour Descriptor (SVD) visual features and Support Vector Machine (SVM) classifier. The proposed technique reportedly outperformed an earlier method proposed by Bozkir and Sezer (2016) where the Histogram of Oriented Gradients (HOG) is used to derive visual features. Nevertheless, the proposed technique can only detect phishing webpages imitating website brands that are already included in the database.

As observed from preceding works, the efficacy of visual similarity-based techniques is often hampered by challenges in maintaining an up-to-date and exhaustive image database of legitimate websites. To address this weakness, Chiew et al. (2015) capitalise on the fact that legitimate websites' logos are indexed by Google and hence querying its reverse image search engine may return the target domain where the logo originated. Machine learning algorithms were used to select the logo from the pool of images involved in rendering the query webpage, where 99.8% and 87.0% of true positives and true negatives were achieved, respectively. Further analysis revealed that the false positives occur on 11.2% of legitimate webpage samples with incorrect logo detections, which were apparently caused by less discriminative image features and suboptimal pre-processing filtering actions.

In subsequent work, Chiew, Choo, et al. (2018) exploited the website favicon as a replacement for the website logo and employed additional features to suppress the false positive rates. The main advantage of utilising website favicon for identity-based technique is the increased consistency in extracting the favicon image. As highlighted in Section 1, it is rather complex to detect and extract a logo image from the webpage content since web developers can freely utilise different Document Object Model (DOM) elements to display the logo. However, the authors observed that 91 out of 3500 legitimate websites were without favicons, hence additional features were employed to suppress the false positive rates. Thus, utilising logo images from the webpage content is still considered a better approach for visual identity-based phishing detection techniques.

In the work of Van Dooremaal et al. (2021), the authors proposed to detect the brand logo in the webpage without prior training by

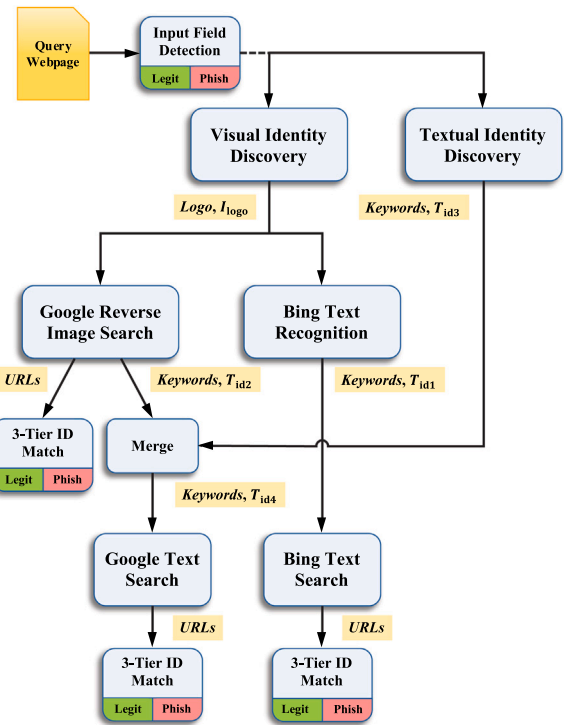


Fig. 1. Overview of the proposed hybrid identity-based phishing detection technique.

utilising connected component analysis and heuristics filter. The proposed technique employed Otsu's thresholding, morphological closing and contour detection to obtain potential logo regions in the webpage screenshot image. Since the potential regions may include arbitrary regions of non-interest, a heuristic filter is applied based on height, width, area, colour count, dominant colour percentage, and spatial location information. Nevertheless, the authors have found that utilising visual identity (e.g., website logo) alone may not consistently resolve to the correct target identity. To complement this limitation, the authors included the webpage title as an additional identity representation in their proposed technique. Therefore, it may be desirable for anti-phishing researchers to begin looking into hybrid identities when designing identity-based phishing detection models.

In short, logo detection is an important mechanism for visual identity-based phishing detection techniques. However, the problem of detecting unknown logos in the webpage content remains an ongoing challenge in the field of phishing detection. In addition, the performance limitations of utilising a single type of website identity have motivated researchers to begin exploring how hybrid identities may improve the performance of classifying both phishing and legitimate websites.

3. Method

To enhance phishing detection accuracy, we propose a hybrid identity-based technique through the combination of visual and textual identity discovery components. Fig. 1 shows the architecture of our proposed hybrid identity-based phishing detection method.

The method incorporates an input field detector as the first line of defence. This is how it works. When a webpage contains zero input fields, it will not be able to receive any personal information from the user. Such a webpage is considered harmless. To standardise the consistency of sample labelling, this type of webpage will be instantly declared legitimate. If a webpage is found to contain at least one input field, the processing continues. In subsequent stages, we apply visual identity discovery and textual identity discovery. For visual

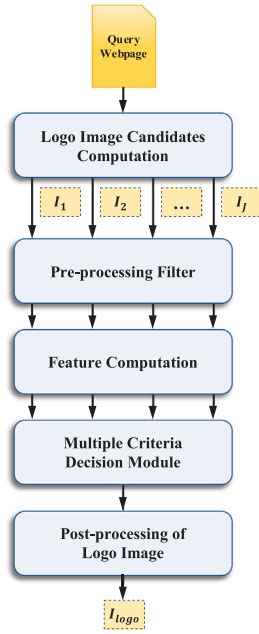


Fig. 2. Internal structure of the visual identity discovery component.

identity discovery, a novel logo detection technique is proposed based on the features and expert knowledge inspired by the human vision in recognising the website logo. For textual identity discovery, we adapted the previous algorithm proposed by Tan et al. (2016) to increase the robustness of the proposed technique against webpages with weak visual identity. Together, the two algorithms produce a tightly integrated phishing detection framework that leverages a multi-faceted approach to verify the identity of the webpage under scrutiny. We further discuss on each component of the method in the sections that follow.

3.1. Visual identity discovery

Visual identity discovery aims to identify and extract the website logo from a webpage, which will later be queried on a reverse image search engine to find the targeted identity. The internal structure of the proposed visual identity discovery component is shown in Fig. 2.

3.1.1. Logo image candidates computation

Logo image candidates is defined as a collection of visual sub-regions in the webpage content, where one of the candidates is the website logo. The process of computing logo image candidates starts by rendering the webpage in a browser using a browser automation framework. In a rendered webpage, DOM elements are representations of HTML tags, which serve as placeholders for textual contents and images to be positioned at various locations. Fig. 3 shows a sample webpage screenshot, where bounding boxes were drawn at selected DOM elements to indicate the logo image candidates.

For efficiency purposes, only DOM elements that potentially contain an image are considered in our logo detection analysis. We notice that most website logos can be seen immediately (i.e., without scrolling) once the page has fully loaded. Hence, it is reasonable to exclude DOM elements that are not visible in the browser window's viewing area or otherwise known as the *viewport*.

In subsequent stages of the visual identity discovery component, these logo image candidates will be analysed using a series of image processing algorithms.

3.1.2. Pre-processing filter

Some logo image candidates may exist in unusual dimensions. For instance, menu icons may be exceptionally small, while advertisements are usually large. These candidates are unlikely to be the website logo and thus can be filtered out, while the remaining logo image candidates will undergo the feature computation process as shown in Fig. 2. Chiew et al. (2015) proposed two simple filtering rules, namely, by excluding images with width or height of less than 10 pixels, as well as monochrome images. Through empirical testing, we adapted and extended these filtering rules to exclude logo image candidates with the following undesirable properties:

$$w \leq w_{\min} , \quad (1)$$

$$h \leq h_{\min} , \quad (2)$$

$$w > \frac{1}{3} W , \quad (3)$$

$$h > \frac{1}{3} H , \quad (4)$$

$$\frac{w}{h} > r_{\max}^{w \rightarrow h} , \quad (5)$$

$$\frac{h}{w} > r_{\max}^{h \rightarrow w} , \quad (6)$$

where w = width of image

h = height of image

w_{\min} = minimum width of image, set to 15 pixels as a start

h_{\min} = minimum height of image, set to 15 pixels as a start

W = width of viewport

H = height of viewport

$r_{\max}^{w \rightarrow h}$ = maximum ratio of $w : h$, set to 15.0 as a start

$r_{\max}^{h \rightarrow w}$ = maximum ratio of $h : w$, set to 2.0

3.1.3. Proposed visual features for logo detection

Given a pool of image candidates from the webpage content, we propose an algorithm that computes a set of features that serve as criteria in deciding which image is most likely to be the website logo. Our approach is akin to solving a multiple criteria decision-making (MCDM) problem. According to MCDM theory, a set of possibly conflicting criteria are considered simultaneously to select the best option from among a set of possible options (Mardani et al. 2015). For ease of presentation and consistency with machine learning notions, the criteria to select the website logo image will continue to be referred to as features. To characterise the inherent properties of a website logo image, we propose a total of eight features F_1, F_2, \dots, F_8 as follow:

(i) F_1 : Vertical position

Some researchers have suggested that website logos are mostly found near the top region of the webpage, which is likely an intentional design strategy to maximise visitors' attention (Chang et al., 2013; Whinton, 2016). Therefore, the vertical position of a logo image candidate is proposed as the first feature and is computed as the normalised y-coordinate using Eq. (7).

$$F_1 = \frac{y}{H} , \quad (7)$$

where y = y-coordinate of logo image candidate within the viewport

H = height of viewport

(ii) F_2 : Horizontal position

It is also fairly common for website logos to be aligned towards the left side of the webpage (Fessenden, 2017). This characteristic can be modelled by the horizontal position of a logo image candidate and is computed as the normalised x-coordinate using Eq. (8).

$$F_2 = \frac{x}{W} , \quad (8)$$



Fig. 3. Example of a webpage screenshot indicating the logo image candidates.

where x = x -coordinate of logo image candidate within the viewport
 W = width of viewport

(iii) F_3 : Vertical region

To supplant the measurement of vertical position, we also included the vertical region where a logo image candidate resides. Specifically, the viewport is divided into upper and lower half regions, such that the feature is computed according to Eq. (9).

$$F_3 = \begin{cases} 0, & \text{if } 0 \leq y < \frac{1}{2}H; \\ 1, & \text{if } \frac{1}{2}H \leq y < H. \end{cases} \quad (9)$$

(iv) F_4 : Standard deviation

Logo images tend to have greater contrast and vividness as opposed to natural images. This characteristic can be measured by using the standard deviation of pixel intensity values. Larger values of standard deviation translate to greater perceived contrast in the image. Given an original colour (i.e., RGB) logo image candidate, it is first converted to an 8-bit grayscale image before the standard deviation is computed by using Eq. (10). The colour to grayscale image conversion is performed using the standard conversion method, i.e., the weighted sum

method (OpenCV team, 2019).

$$F_4 = \sqrt{\frac{1}{N} \sum_{i=1}^N (h_i - \mu)^2}, \quad (10)$$

where N = total number of pixels
 h_i = pixel intensity value of the i -th pixel
 μ = mean of the pixel intensity value

(v) F_5 : Padding space

Based on empirical observations, our method assumes that website logos are tightly fitted or in other words, have minimal padding space along the edges of the image. In contrast, regular user interface (UI) elements such as menu icons and labels may have larger padding space, as illustrated in Fig. 4.

To characterise this heuristic, the padding space in the logo image candidate is computed as follows. Given a logo image candidate I_{RGB} , it is first converted to a grayscale image I_{gray} . If the width of I_{gray} is larger than its height (i.e., horizontally long), the padding space will be computed using strips of column pixels. To aid our presentation, Fig. 5 is also referred to. The computation begins from the left side of I_{gray} . In each iteration, a single strip of column pixels is sampled and its standard deviation σ is computed using Eq. (10). Note that σ in this case is the F_4 value. The sampling process is repeated until



Fig. 4. Visual comparison showing the actual logo of Capital One¹ website (left) having minimal padding spaces as compared to most non-logo image candidates (right).

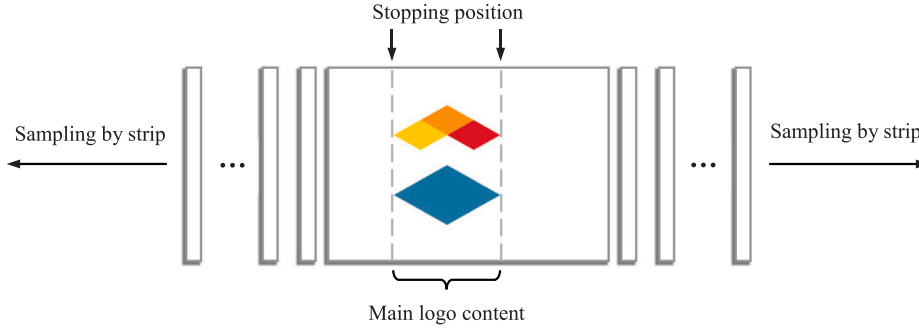


Fig. 5. Sampling of column strips for computation of padding space.

$\sigma \neq 0$, which indicates the presence of visually distinct content in the image. At this point, the current count of column strips with $\sigma = 0$ are recorded, and similar computation is performed by sampling strips of column pixels starting from the right side of the image. The final count of column strips with $\sigma = 0$ is divided with the width of the logo image candidate to obtain the percentage of padding space as the value for feature F_5 . Note that if the image has a larger height than its width (i.e., vertically long), computation is performed by sampling strips of row pixels instead.

(vi) F_6 : Distance to dimension ratio of unity

Based on general observations on our dataset samples, logo images seem to be mostly rectangular, while non-logo images such as UI elements tend to be square-shaped. To further investigate this potential occurrence, we examined a public dataset compiled by Su et al. (2019), namely, the QMUL-OpenLogo dataset² consisting of 352 logo classes. Analysis revealed that only 20.5% of the logo images have a width-to-height ratios that fall within the range of 0.9 and 1.1, which is consistent with our observation. However, this finding contradicts Chiew et al. (2015) who observed that the width-to-height ratio for most logo images are near to unity or 1.0 (i.e., square ratio). Since it is plausible that a smaller proportion of website logo images may exist in square-shape forms, it is necessary to consider all the proposed features collectively to facilitate the correct identification of the website logo. For this feature, we assume that logo image candidates with rectangular dimensions are more likely to be the website logo. To enumerate this characteristic, feature F_6 is computed by using Eq. (11).

$$F_6 = |1.0 - r^{h \rightarrow w}|, \quad (11)$$

where $r^{h \rightarrow w}$ = height to width ratio of the logo image candidate

(vii) F_7 : Colourfulness

Conventional features for website logo detection are typically derived from pixel intensity values of the grayscale image (Chiew et al., 2015). Such practice limits the effectiveness of logo detection since the colour information is completely ignored. We postulate that colourfulness is a key aspect that helps humans to differentiate between logo and non-logo images in a webpage (Van Dooremaal et al., 2021). Thus, we exploit the colourfulness feature for logo detection.

Given an RGB image, Hasler and Suesstrunk (2003) proposed to compute a colourfulness metric based on the mean and standard deviation of opponent colour components (i.e., yellow-blue and red-green). This is one of the earliest metrics for image colourfulness. However, this colourfulness metric is occasionally found to be poorly correlated with users' perceptual judgements (Reinecke et al., 2013; Zerman et al., 2019).






As such, we define a novel colourfulness metric as shown in Algorithm 1. First, let us present the underlying intuition. In RGB theory, a pure grey colour has equal values in the red, green and blue colour components as shown in Table 1, where higher values correspond to lighter shades of grey and smaller values indicate darker shades of grey. Deviations in any of the colour components will offset the perceived colour away from pure grey and make it appear to be more colourful. Therefore, it can be hypothesised that larger deviations among the three values in an RGB tuple will result in a more colourful shade. In short, our proposed colourfulness metric measures the extent that a given logo image candidate deviates from the grayscale colour scheme. Preliminary tests showed that our proposed colourfulness metric yields more reasonable values that correspond nearer to perceived colourfulness as compared to the earlier metric proposed by Hasler and Suesstrunk (2003).

Based on line 2 in Algorithm 1, a logo image candidate I_{RGB} will first undergo a dominant colour extraction process. For this purpose,

¹ <https://www.capitalone.com/>

² <https://qmul-openlogo.github.io/>

Table 1
Different colour shades and their corresponding RGB values.

Colour shade	Colour description	RGB values
	Dark gray	(60, 60, 60)
	Medium gray	(120, 120, 120)
	Light gray	(180, 180, 180)
	Light pink	(240, 180, 180)
	Light blue	(120, 180, 220)

Algorithm 1 Computation of feature F_7 .

```

Input:  $I_{RGB}$   $\triangleright I_{RGB}$  = logo image candidate
Output:  $F_7$ 
1: Begin
2:  $dominant\_colour\_array = KMEANSCLUSTERING(I_{RGB})$ 
3: for all  $RGB\_value \in dominant\_colour\_array$  do
4:    $\hat{\sigma} = COMPUTENORMALISEDSTANDARDDEVIATION(RGB\_value)$ 
5:   if  $\hat{\sigma} > 0.1$  then
6:      $SD\_array \leftarrow \hat{\sigma}$ 
7:   end if
8: end for
9: if  $EMPTY(SD\_array)$  then
10:    $F_7 = 0.0$ 
11: else
12:    $F_7 = COMPUTEMEAN(SD\_array)$ 
13: end if
14: End

```

the k-means clustering algorithm is invoked, with the number of clusters set to 5 in order to extract up to a maximum of 5 dominant colours. Here, we do not discuss the theoretical concepts of k-means clustering, which is a well-established machine learning algorithm that can be readily referred to in VanderPlas (2016). For each set of dominant colours, the standard deviation of its RGB values is computed, then normalised by division with a constant of 120. This constant is derived by considering an RGB tuple with the widest value separation among the colour components, namely, the minimum and maximum values of 0 and 255, respectively. For example, an RGB tuple of (0, 0, 255) will lead to the largest possible standard deviation of 120.21. The mean of the normalised standard deviation $\hat{\sigma}$ of each dominant colour will be assigned as the value for feature F_7 .

(viii) F_8 : Distance to the common minimum threshold area

The proposed features for logo detection would be incomplete without considering the size of the logo image candidates. Although the size of a typical website logo image is rather flexible, it should surpass a common minimum threshold to be sufficiently visible. Hence, we utilise the minimum height w_{min} and minimum width h_{min} as discussed in Section 3.1.2 to compute the common minimum threshold area using Eq. (12). When given a logo image candidate, we compute feature F_8 using Eq. (13).

$$A_{min} = w_{min} \times h_{min} \quad (12)$$

$$F_8 = \left| \frac{A}{A_{min}} - 1.0 \right| \quad (13)$$

Table 2
Weight setting of each feature.

Feature	Description	Weight
F_1	Vertical position	0.5
F_2	Horizontal position	0.4
F_3	Vertical region	0.2
F_4	Standard deviation	0.2
F_5	Padding space	0.3
F_6	Distance to dimension ratio of unity	0.3
F_7	Colourfulness	0.3
F_8	Distance to the common minimum area	0.2

Table 3
New parameter values in the relaxed filtering rules.

Parameter	Initial value (as defined in Section 3.1.2)	New value
w_{min}	15.0	10.0
h_{min}	15.0	10.0
$p^{w \rightarrow h}_{max}$	15.0	20.0

where A_{min} = common minimum threshold area
 A = area of logo image candidate

3.1.4. Multiple criteria decision module

For each webpage sample, the proposed features were used to rank the logo image candidates, where the highest-ranked candidate was picked as the website logo. To perform the ranking, we utilised compromise programming which is one of the MCDM methods. *Compromise programming* provides a way to systematically combine values of different features into a single distance metric value L_p as shown in Eq. (14). After the L_p value for every logo image candidate have been computed, the optimum decision (i.e., logo image) is made by selecting the candidate having the smallest L_p value, which corresponds to the highest ranking.

$$L_p(j) = \left[\sum_{k=1}^8 W_k^p \left| \frac{F'_k - F_k(j)}{M_k - m_k} \right|^p \right]^{\frac{1}{p}} \quad (14)$$

where $L_p(j)$ = distance metric for j -th logo image candidate
 F'_k = ideal value of the k -th feature
 $F_k(j)$ = value of the k -th feature for j th logo image candidate
 M_k = maximum value of k -th feature among all logo image candidates
 m_k = minimum value of k -th feature among all logo image candidates
 W_k^p = weight of k -th feature
 p = parameter controlling the influence of deviation magnitude on distance computation

When the p parameter is increased, the distances contributed by larger deviations are amplified while distances of small deviations are diminished. In our experiment, the value of p is set to 1 so that all deviations, regardless of magnitude, will have an equal influence on L_p . In determining the weight of the features, we regard certain features to be more important in distinguishing the website logo and hence, should be awarded higher weights. By empirical testing, the optimal weights of the feature were assigned as specified in Table 2.

In Section 3.1.2, filtering rules were applied to exclude logo image candidates with irregular properties. However, in some rare conditions, the website logo itself may have irregular properties, causing it to be excluded from the list of candidates. To account for these exceptional cases, we employed a secondary analysis with relaxed filtering rules as specified in Table 3.

With the inclusion of additional logo image candidates that pass the secondary pre-processing stage, the analysis continues with feature computation and selection of logo image via compromise programming,



Fig. 6. Examples of website logos rendered using actual text.

as described in Section 3.1.3 and Section 3.1.4. The final logo image will be selected from the output of either the primary or secondary analysis based on the conditions defined in Algorithm 2.

Algorithm 2 Selection of final logo with relaxed filtering rules.

Input: $I'_{\text{logo}}, I''_{\text{logo}}$ $\triangleright I'_{\text{logo}}$ = logo image from primary analysis
Output: I_{logo} $\triangleright I''_{\text{logo}}$ = logo image from secondary analysis

```

1: Begin  $\triangleright I_{\text{logo}}$  = final logo image
2: if  $\text{EMPTY}(I'_{\text{logo}})$  then
3:    $I_{\text{logo}} = I''_{\text{logo}}$ 
4: else
5:    $y'_{\text{logo}} = \text{GETVERTICALPOSITION}(I'_{\text{logo}})$ 
6:    $y''_{\text{logo}} = \text{GETVERTICALPOSITION}(I''_{\text{logo}})$ 
7:   if  $y'_{\text{logo}} \leq \frac{1}{3}H$  and  $y''_{\text{logo}} \leq \frac{1}{3}H$  then
8:      $A'_{\text{logo}} = \text{GETAREA}(I'_{\text{logo}})$ 
9:      $A''_{\text{logo}} = \text{GETAREA}(I''_{\text{logo}})$ 
10:    if  $A''_{\text{logo}} > A'_{\text{logo}}$  then
11:       $I_{\text{logo}} = I''_{\text{logo}}$ 
12:    else
13:       $I_{\text{logo}} = I'_{\text{logo}}$ 
14:    end if
15:  else
16:    if  $y'_{\text{logo}} > \frac{1}{3}H$  and  $y''_{\text{logo}} \leq \frac{1}{3}H$  then
17:       $I_{\text{logo}} = I''_{\text{logo}}$ 
18:    else
19:       $I_{\text{logo}} = I'_{\text{logo}}$ 
20:    end if
21:  end if
22: end if
23: End

```

Note that after undergoing relaxed filtering rules, it is still possible that no logo is detected (i.e., I_{logo} is empty). There are two possibilities that give rise to this scenario:

- The website logo is rendered using actual text in the webpage instead of loaded from an image file.. Based on examples in Fig. 6, textual logos may appear similar to image-based logos at a glance. This scenario occurs because our proposed logo detection analysis only considers DOM elements that potentially contain an image, as described in Section 3.1.1.
- The webpage indeed does not have any logo. This usually happens for webpages with weak visual identity such as personal blogs, forums, archive sites, etc., which tend to employ less appealing visual styles and reduced visual artefacts.

In this condition, the proposed visual identity discovery component is unable to detect any website logo. This limitation can be addressed by employing a complementary component, namely, the textual identity discovery, which will be presented in Section 3.2 later.



Fig. 7. Segmented logo with noisy background.

3.1.5. Post-processing of logo image

Post-processing is employed on the selected logo I_{logo} to optimise it for search engine queries. Two optimisation processes are performed here, namely, background recolouring and elimination of padding space.

Background recolouring is performed on the selected logo image to reduce noise that appears in the transparent regions. Recall that the DOM elements in a webpage are layered on top of one another. Hence, if the selected logo has transparent regions, the backgrounds that are underneath the logo will be exposed. Fig. 7 illustrates this problem, where the Tumblr logo is surrounded by a background image. Segmenting the screenshot directly at the DOM element of the selected logo will produce a Tumblr logo with noisy background, which may lead to inaccurate results when processed by the reverse image search engine.

To overcome this limitation, we employ an active DOM manipulation approach to recolour the background of the selected logo with a solid colour. Specifically, the dominant colour of the background regions underneath the logo is identified using the same algorithm utilised earlier in the computation of the colourfulness feature, namely, the k-means clustering (VanderPlas, 2016). By manipulating the CSS *background-color* attribute of the logo's DOM element, the dominant colour can then be overlaid underneath the logo as a solid colour background, replacing the original noisy background, as illustrated in Fig. 8.

The second optimisation process, i.e., elimination of padding spaces, is performed on the selected logo image to discard regions that do not contain any distinctive characteristics of the logo. The presence of padding spaces is undesirable as it is found to degrade the recognition performance of reverse image search engines, where further discussion on this finding is available in Section 4.8.

To eliminate the padding spaces in a given logo image, we performed iterative sampling on strips of row and column pixels. This technique is similar to the computation of padding space feature in Section 3.1.3, where strips of row and column pixels in the selected logo image were sampled iteratively. If a pixel strip has a standard deviation value of zero, it will be eliminated. Elimination will stop once a non-zero standard deviation value is encountered, indicating the presence of visually distinct content. Fig. 9 summarises the process of eliminating padding spaces in the selected logo image.

After the elimination of padding spaces, w_{\min} and h_{\min} filtering rules are reapplied to decide whether or not to discard the apparent logo image. This is to ensure that the selected logo, after undergoing padding spaces elimination, can still fulfil the typical minimum dimension as defined in Section 3.1.2. If the selected logo fails in the final filtering rules, we will end up without a logo to be utilised for the subsequent stage. In this case, the webpage is deemed as having a weak visual identity, where we will rely fully on the textual identity discovery component for further analysis.

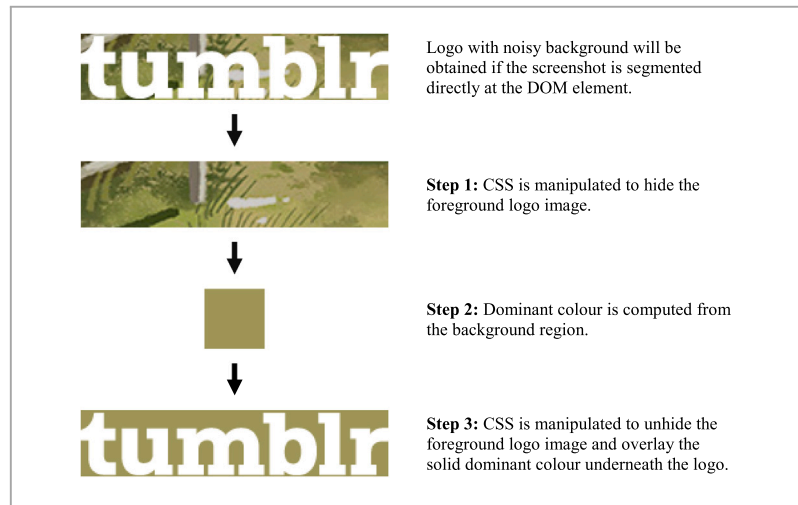


Fig. 8. Process of obtaining the selected logo with a clean background.

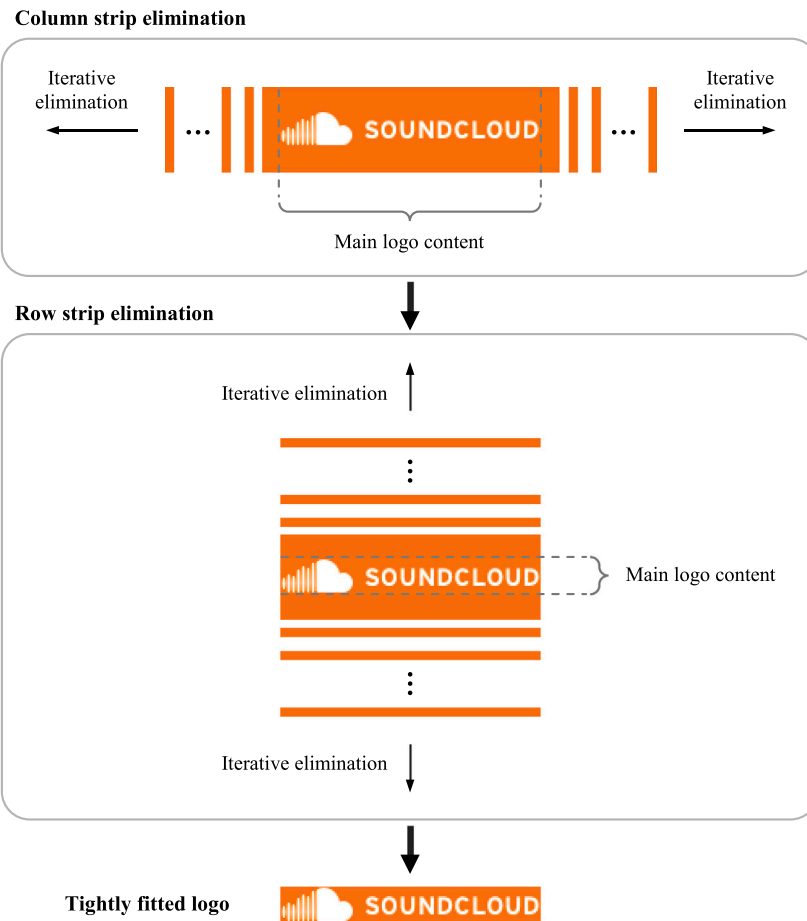


Fig. 9. Elimination of padding spaces in the selected logo image.

3.2. Textual identity discovery

Textual identity discovery aims to extract identity keywords from a webpage. Using a search engine, the identity keywords serve as a textual query to find the targeted identity. To derive the textual identity discovery component, we adopted a computational approach from our prior work (Tan et al., 2016). The approach leverages the structure of URLs in a webpage to extract identity keywords in the form of single keywords (uni-gram) or multiple coexisting keywords (multi-gram).

An overview of the proposed textual identity discovery component is shown in Fig. 10.

The process begins by feeding the given webpage into an N-gram generator to extract plain texts from the webpage content, which are then used to generate uni-grams and multi-grams. The N-grams are pre-processed by removing whitespaces and hyphens. Acronyms of the N-grams are also computed as aliases for subsequent string matching operations. If an N-gram does not contain any uppercase letters, it will be discarded.

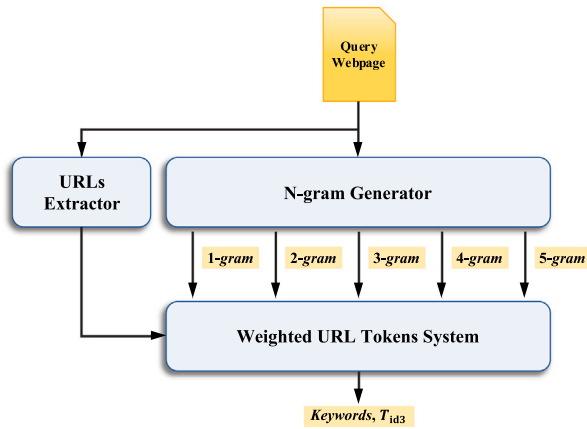


Fig. 10. Overview of the textual identity discovery component.

The weighted URL tokens system is basically a weight generator that leverages the URLs structure to assign weights to the N-grams. These URLs are extracted from the given webpage content, specifically from the *href* and *src* attribute in the HTML source codes. Forward slashes are used as delimiters to split each URL into tokens, where each N-gram will be compared against the tokens. When a string match is encountered, the N-gram is assigned a weight according to a predefined equation. Detailed computation of the weights can be found in Tan et al. (2016). For uni-gram, five words with the highest weight are selected and subjected to a set of elimination rules. As for multi-grams, only the highest weighted multi-gram is selected. Finally, the uni-gram and multi-gram keywords are merged to produce the final identity keywords T_{id3} .

3.3. Establishing the webpage identity

Using the combination of visual and textual identity discovery components, multi-faceted identity-related queries are formulated and submitted to the search engine to establish the identity of the webpage. If any of the entries in the search results matches the query webpage's domain name, then a legitimate label is obtained or vice versa. In other words, we check whether the query webpage (represented by the identity-related query) is capable of scoring a hit to domain names listed in the search results.

3.3.1. Querying reverse image search engine with the logo image

Once the logo image is selected, our method queries Google Reverse Image Search³ to find matching image data from the search engine database. Note that alternative reverse image search engines are also available such as those from Baidu⁴ and Yandex.⁵ For this study, we followed the recommendations by Chiew et al. (2015) and Van Dooremaal et al. (2021) to utilise the well-established reverse image search service by Google. An example of results returned by Google Reverse Image Search is shown in Fig. 11.

From the returned results, URLs were extracted from results subset-A and subset-C (as labelled in Fig. 11), where the domain names were compared against the query webpage domain name using an indirect identity comparison technique from our prior work in Tan et al. (2016), namely, the 3-tier identity matching component. The 3-tier identity matching component will be presented more thoroughly in Section 3.3.4. At this stage, it is sufficient to know that the output of the

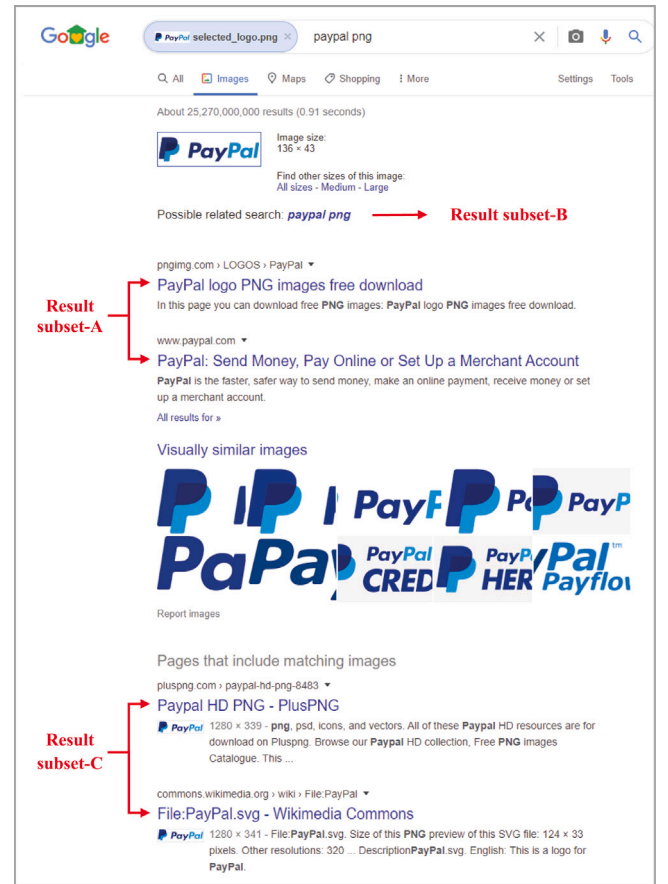


Fig. 11. An example of search results returned by Google Reverse Image Search.

3-tier identity matching component is a binary decision, e.g., phishing or legitimate.

Result subset-B, on the other hand, is a set of keywords possibly related to the image content. When a logo image is submitted and correctly recognised, result subset-B will reflect the identity keywords or brand name of the logo image. In the next subsection, we will explain how result subset-B is utilised.

3.3.2. Querying search engine with textual keywords

Note that result subset-B from the Google Reverse Image Search provides the identity keywords T_{id2} as denoted in Fig. 1. Here, T_{id2} was combined with T_{id3} (output from the textual identity discovery component) to form an optimised set of identity keywords T_{id4} . Using T_{id4} , a regular textual search was invoked on Google Search Engine, and the results were analysed. Similar to Fig. 11, we collected the entries (i.e., URLs) in the search results and leveraged the 3-tier identity matching module to obtain a binary result indicating the legitimacy of the query webpage.

3.3.3. Querying search engine with visual keywords extracted from the logo image

Most website logo consists of shape-based and name-based identifiers. Fig. 12 shows examples of both identifiers in some website logos. The name-based identifier is often the website domain name, brand name or identity keywords. Therefore, in addition to T_{id4} , we propose to extract the name-based identifier in the logo as additional information to facilitate in finding the webpage identity. For this purpose, we employed Microsoft Text Recognition — an Optical Character Recognition

³ <https://www.google.com/imghp?sb=1>

⁴ <https://image.baidu.com/>

⁵ <https://yandex.com/images/>

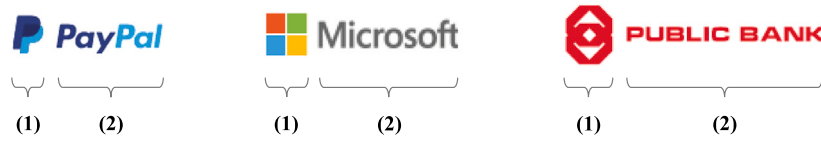


Fig. 12. Example of shape-based identifier (1) and name-based identifier (2) in some website logos.

(OCR) service hosted on the Bing Visual Search⁶ page. Note that the utilisation of Microsoft Text Recognition in this study is intended for proof-of-concept purposes, and it can be replaced with open source technologies in future work.

The textual output from the OCR process forms the keyword set T_{idl} , which was subsequently fed to Bing Text Search. From the search result entries, URLs were extracted and processed by the 3-tier identity matching module to verify the query webpage's legitimacy.

3.3.4. 3-Tier identity matching module

This subsection describes the proposed method to validate the query webpage identity against potential target identities. Normally, this process involves matching the domain name of the query webpage against domain names in the search result entries. However, country-specific websites can be falsely detected as phishing if conventional domain name matching is employed. For example, given a query webpage from the global eBay site (ebay.com), it is logical for us to classify it as legitimate if the search result contains its country-specific websites such as ebay.com.my, ebay.co.uk, etc. Note that within the domain names, "ebay" is known as the Second-level Domain (SLD) while ".com.my" and ".co.uk" are country-code Top-level Domain (ccTLD). Country-specific websites are commonly deployed by large organisations to achieve better engagement with visitors in a local region. Some companies may also deploy affiliated websites that are completely different in terms of the domain name but are both having the same IP address (i.e., hosted on the same server), though this practice is less common.

As such, we adopted the 3-tier identity matching module from our prior work in Tan et al. (2016) to resolve the aforementioned identity discrepancies. In short, the 3-tier identity matching module exploits domain name matching, ccTLD matching, and IP address matching to achieve a more logical ownership identification, which contributes to improved detection accuracy for legitimate websites. Note that the 3-tier identity matching module is invoked at various stages in the proposed phishing detection technique, particularly when there is a list of search result entries that needed to be matched with the query webpage domain name. The final output of the 3-tier identity matching module is a binary decision of either phishing or legitimate. If a query webpage is classified as legitimate at any stage in the proposed phishing detection technique, the remaining processes will be skipped. In other words, a legitimate result provides an early exit route to reduce unnecessary computational processing.

4. Results and discussions

4.1. Dataset preparation

To construct the dataset, we collected complete webpage samples which include the HTML documents and their linked resources (e.g., images, CSS files, JavaScript files, etc.) required for proper rendering at our local machine. Using the GNU Wget⁷ tool, the webpage samples were downloaded from two disjoint periods, namely, from September to December 2018 and from January to March 2020. Legitimate webpages were collected based on the Alexa top 1 million list,⁸ while phishing webpages were sourced from URLs in PhishTank⁹

and OpenPhish¹⁰ archive. The downloaded webpage samples were then filtered by first removing samples with duplicate URLs. A more intensive and stringent check was also employed, whereby each local webpage sample was loaded in the browser to verify the fidelity of its content and visual layout. Moreover, samples with duplicate visual content and presentation styles were kept to a minimum. In short, we constructed the first dataset (DS-1) consisting of 500 phishing samples, and another 500 legitimate samples sourced from the top 1000 rank of the Alexa top websites list. The second dataset (DS-2) consists of 250 legitimate webpages ranked in between the top 5000 to 6000 of the Alexa top websites list.

4.2. Experimental setup

We developed a software prototype in Python. Real-time rendering of webpages in the browser and the manipulation of webpage content were made possible using the Selenium Webdriver¹¹ framework. Textual or image-based queries to search engines and scrapping of search results were also performed programmatically using Selenium Webdriver.

For image processing operations, the OpenCV¹² computer vision library was used. Experiments were carried out on a desktop computer equipped with an Intel Core 2 Quad 2.66 GHz CPU, 8 GB RAM and Windows 7 Professional 64-bit operating system. We evaluated the performance of the proposed technique using true positive rate (TPR), true negative rate (TNR), and accuracy, which are computed as follows:

$$TPR = \frac{|\text{Samples correctly classified as phishing}|}{|\text{Actual phishing samples}|} \times 100, \quad (15)$$

$$TNR = \frac{|\text{Samples correctly classified as legitimate}|}{|\text{Actual legitimate samples}|} \times 100, \quad (16)$$

$$\text{Accuracy} = \frac{|\text{Samples correctly classified}|}{|\text{Actual phishing and legitimate samples}|} \times 100. \quad (17)$$

4.3. Evaluation I – performance of the proposed technique on high and low-ranked websites

In this experiment, the performance of the proposed technique was evaluated using both datasets DS-1 and DS-2. Besides assessing the phishing detection performance, this experiment also allows us to see how the proposed technique performs on different sets of legitimate websites which may vary in terms of visual and content presentation quality. Specifically, we will investigate whether low-ranked legitimate websites can negatively impact the true negative rate, as claimed by Chiew et al. (2015). Table 4 shows the performance of the proposed phishing detection technique on the two datasets.

Results showed that the proposed technique, when evaluated on DS-1, achieved a fairly high true positive rate of 99.4%, indicating its efficacy in detecting phishing webpages. Based on the operating mechanism of identity-based detection techniques, obtaining an outstanding true positive rate is effortless, since these techniques are almost guaranteed not to find any fundamental brand identity relations in phishing webpages. On the other hand, results showed that true

⁶ <https://www.bing.com/visualsearch/Microsoft/TextRecognition>

⁷ <http://www.gnu.org/software/wget/>

⁸ <http://s3.amazonaws.com/alexa-static/top-1m.csv.zip>

⁹ <https://www.phishtank.com/>

¹⁰ <https://www.openphish.com/>

¹¹ <https://www.selenium.dev/>

¹² <https://opencv.org/>

Table 4

Performance of the proposed phishing detection technique on datasets DS-1 and DS-2.

Dataset	TPR (%)	TNR (%)	Accuracy (%)
DS-1	99.40	97.80	98.60
DS-2	–	94.00	–

Note. TPR and accuracy metrics were not computed for DS-2 because it consists of only legitimate webpage samples.

Table 5

Performance comparison of visual, textual, and hybrid identity-discovery components in the proposed technique.

Identity discovery component	TPR (%)		TNR (%)	
	DS-1	DS-2	DS-1	DS-2
Visual	100.00	–	49.60	38.80
Textual	99.40	–	96.20	88.80
Hybrid	99.40	–	97.80	94.00

Note. TPR was not computed for DS-2 which consists of only legitimate webpage samples.

negatives are slightly lower at a rate of 97.8% while the overall accuracy is at 98.6%. Comparing the true negative rates between high-ranked legitimate websites (DS-1) and the low-ranked ones (DS-2), a difference of 3.8% can be observed. This discrepancy seems to support the findings in earlier studies claiming that the authenticities of low-ranked legitimate websites are more difficult to verify. Low-ranked or unpopular legitimate websites often lack strong, easily identifiable, and discriminative identity indicators such as the brand logo. For example, non-commercial organisations may not have the necessity to put up a strong brand image or invest in designing a distinct logo that has typical logo characteristics. Moreover, these organisations may not employ Search Engine Optimisation (SEO) techniques on their websites, causing them to be poorly indexed and granted inferior ranking by search engines. Looking at the high-ranked and low-ranked legitimate websites as a whole, it is apparent that their webpage content presentation and visual layout can vary widely and this often becomes a performance bottleneck for conventional identity-based techniques. Therefore, in our proposed technique, textual identity is a necessary component to reinforce the visual identity so that a robust brand identity can still be obtained despite the erratic nature of webpage content that users will normally encounter.

4.4. Evaluation II – performance comparison of visual, textual, and hybrid identity discovery components

In the second experiment, we evaluated the performance of the two main components in the proposed technique, namely, visual identity discovery and textual identity discovery, as well as the combination of both components. The results are presented in Table 5.

Results indicate that both visual and textual identity discovery components achieved comparable performance in terms of phishing detection rate, attaining 100.0% and 99.4% of true positive rates, respectively. When looking at the true negative rates, the visual identity discovery component underperformed significantly, scoring only 49.6% on high-ranked legitimate websites and 38.8% on low-ranked legitimate websites.

The results do not appear to correlate with the findings of Chiew et al. (2015), where 87.0% of true negative rate was achieved, although it should be noted that their datasets and system architecture were different from ours. To gain more insights on this result, further investigations were conducted on samples misclassified by the visual identity discovery component, which will be reported in Section 4.6.

For the textual identity discovery component, reasonable performance was observed on DS-1 with a true negative rate of 96.2%. When DS-2 was tested, the true negative rate suffered a 7.4% drop as

Table 6

Performance comparison between the proposed technique and Tan et al. (2016).

Technique	TPR (%)	TNR (%)	Accuracy (%)
The proposed technique	99.40	97.80	98.60
Tan et al. (2016)	99.40	94.40	96.90

compared to DS-1. By aggregating both identity discovery components to form a hybrid component, an improved true negative rate of 97.8% was achieved on DS-1, while the performance drop was suppressed to 3.8% when evaluated on DS-2.

Based on experimental results, the joint visual and textual identity discovery components are proven to be advantageous, which enabled the proposed technique to achieve greater performance and stability on diverse classes of legitimate websites. Establishing the correct identity on legitimate websites has been a prevalent challenge for identity-based phishing detection techniques (Jain & Gupta, 2018; Rao & Pais, 2019; Van Dooremaal et al., 2021). Our proposed technique contributes to reducing the misclassification rate for legitimate websites while maintaining a good phishing detection rate. Thus, the aggregation of visual and textual identity discovery components can be seen as a reinforcement to the overall performance of the proposed anti-phishing technique, enabling it to achieve better accuracies in classifying both high and low-ranked legitimate websites.

4.5. Evaluation III – performance benchmarking with existing identity-based phishing detection technique

In this evaluation, the proposed technique was benchmarked against an existing identity-based phishing detection technique (Tan et al., 2016). Tan et al. (2016) is the most relevant work for comparison because it is based on the similar objective of finding the target identity and it was previously benchmarked against other works and proven to be one of the most superior identity-based methods. To ease the performance comparison, dataset DS-1 was utilised to evaluate both techniques.

Based on the results in Table 6, both techniques attained the same true positive rate. This finding is not surprising, as our proposed technique adopts a similar identity keywords extraction algorithm as Tan et al. (2016). The performance difference is more distinct when looking at the true negative rates, where our proposed technique outperformed (Tan et al., 2016) by 3.4%. This indicates that our proposed technique, which utilises the combination of visual and textual identity discovery components is able to correctly resolve the identity for a larger portion of legitimate webpages. The results of this evaluation substantiate the findings in Section 4.4, suggesting that the integration of visual and textual identity discovery is vital for improving the overall accuracy.

4.6. Analysis of misclassified legitimate samples

In line with the objective of our study, the misclassified legitimate samples were analysed further to gain better insights into the factors that contributed to the incorrect results. As such, the legitimate samples in dataset DS-1 were chosen and examined, where the results are presented according to three possible outcomes of the logo detection: (a) *correct logo detected*; (b) *wrong logo detected*, and; (c) *no logo detected*. For each outcome, we counted the number of legitimate and phishing labels assigned by the proposed technique, where a legitimate label is considered a correct decision (i.e., true negative) while a phishing label represents a wrong decision (i.e., false positive). Since there exist no ground truth data for comparison, the results of (a) and (b) were compiled by manually judging the correctness of the detected logo. By manually comparing the selected logo image to the webpage URL and overall webpage content, we were able to make informed decisions on whether the logo image was correctly detected or not. Fig. 13 shows the

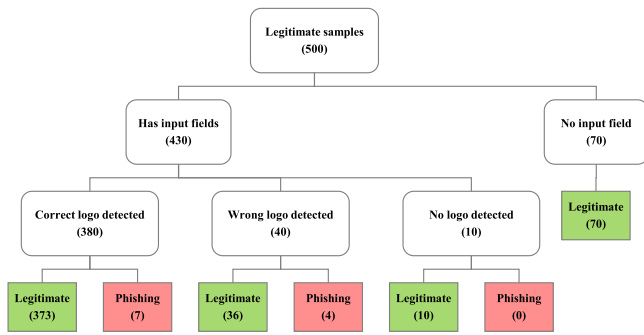


Fig. 13. Breakdown analysis of phishing and legitimate labels by logo detection outcome of legitimate samples in dataset DS-1.

breakdown analysis of phishing and legitimate labels according to the logo detection outcome. Note that the proposed technique first checks for the presence of input fields in a given webpage, and will straight-away classify it as legitimate if no input field is found. Therefore, our analysis here is focused on a subset of legitimate samples (430 in total) in dataset DS-1 where input fields exist and thus will get processed by the subsequent components in the proposed technique.

Results indicate that 380 out of 430 legitimate samples in dataset DS1 have their logos correctly detected, which is equivalent to a logo detection accuracy of 88.4%. Within this pool of webpage samples, 98.2% of them were correctly detected to be legitimate webpages.

On another 40 samples in dataset DS-1, the detected logos were found to be incorrect. Out of these 40 samples, the proposed technique managed to attain near to 90% of correct decisions, even when wrong logos were detected. This shows that the proposed technique is rather robust in coping with the irregular nature of webpage content, which is not always favourable for the detection of website logos (Gupta & Jain, 2020). For example, a small number of webpages in this sample pool were found to be written in languages that flow from right to left such as the Arabic language. Such webpage usually mirrors the traditional horizontal layout of conventional webpages utilising left-to-right language (e.g., English, Mandarin, etc.). As a result, important identity cues such as brand names and the website logo will also be placed towards the right edge of the page (Bozkir & Aydos, 2020; Whinton, 2016). Fig. 14 shows a sample webpage from our dataset that utilises right-to-left language flow or mirrored horizontal layout, resulting in the website logo being placed on the top right region. Nevertheless, studies have shown that violating the left-aligned logo placement convention will weaken users' ability to recall the brand name (Whinton, 2016). Therefore, logo detection errors contributed by this category of webpages are expected to be less often since such design convention is rather uncommon.

Based on our analysis, incorrect logos were also found to originate from webpages utilising logos with black-and-white colour scheme. This is not surprising, as our proposed technique regards a colourful image to be more representative of a website logo. In addition, some logos that exist as square images have failed to be detected as well. Examples of some square-shaped logos are shown in Fig. 15. As part of our future work, it may be possible to mitigate these limitations by improving the feature design and optimising the weight of the related features.

On another 10 samples, the proposed technique was unable to extract any logo. Note that this does not mean that all the samples within this set are truly without a logo. Analysis was performed on this set of 10 samples to determine the reasons for failure in detecting the correct logo, where the findings are tabulated in Table 7. Interestingly, based on Fig. 13, notice that the correct decision labels (i.e., Legitimate under *No logo detected* category) were still successfully assigned to this pool of samples. This is owing to the integration of visual and textual

Table 7

Analysis of results on samples where no logo is detected.

Number of samples (out of a total of 10)	Website logo actually exists	Reason for failing to detect the correct logo
3	Yes	Logo is rendered using actual text (non-image).
2	Yes	Logo exists in square form, causing it to be perceived as more likely to be a non-logo (feature F_6).
2	Yes	Logo uses a black and white (i.e., less colourful) colour scheme.
1	Yes	Logo failed to load when the downloaded webpage sample is tested on the local machine.
1	Unsure	No visual element that stands out as the logo. Considered as a webpage with low quality design.
1	Yes	After undergoing padding spaces elimination in post-processing, the logo failed to fulfil the typical minimum dimension as defined in Section 3.1.2.

identity discovery components, where the absence of a logo does not completely paralyse the efficacy of the proposed technique. The textual identity discovery component serves as a fall-back method in cases where the visual identity discovery component extracted the wrong logo or failed to extract any logo, thus the proposed phishing detection technique is still able to maintain a robust performance.

From the pool of 380 correctly detected logos, 7 samples were given wrong decision labels (i.e., phishing). Upon investigation, we found that those samples were mostly caused by the failure of Google Reverse Image Search to correctly identify the brand name represented by a given logo. When this happens, the reverse image search result will return unrelated keywords as a possible guess for the image content, as well as unrelated search results corresponding to that possible guess. We suspect that unfavourable properties (e.g., low resolution) and insufficient shape details in the logo image might be preventing Google Reverse Image Search from correctly recognising it. Alternatively, it could also be due to the logo that was yet to be indexed into the reverse image search database. Chiew et al. (2015) reported that Google Reverse Image Search can sometimes return incorrect results if a logo image lacks distinctive visual patterns or closely resembles other logos in terms of colour and shape.

It is important to note that, when Google Reverse Image Search fails to recognise the correct identity from the logo (i.e., results subset-A and subset-B shown in Fig. 11 are incorrect), the final outcome for the legitimate webpage is not necessarily incorrect (labelled as phishing). Under this circumstance, Google Reverse Image Search may still be able to return webpages through result subset-C that are found to contain the exact match of the submitted logo. If the target domain is found among the entries of result subset-C, then a correct decision (i.e., legitimate) can still be obtained.

Additionally, there are two supplementary components that attempt to establish the website identity by using a textual approach. The textual identity component extracts a set of textual keywords from the webpage content and merges it with another set of textual keywords from result subset-B produced by Google Reverse Image Search. The resultant keywords, when searched on Google using regular text search, may yield a correct decision if the target domain is found in the search results. In addition, it is also possible for the webpage to be correctly classified as legitimate if the Microsoft Text Recognition is able to extract correct identity keywords from the logo, resulting in a subsequent Bing Text Search that returns results containing the target domain. In short, only 1.8% of legitimate samples (with correct logos extracted) were misclassified, and it occurs as a result of all aforementioned supporting components failing to determine the correct identity. This underlines the advantage of the proposed technique which consists of



Fig. 14. Screenshot of a sample webpage utilising right-to-left language flow.



Fig. 15. Examples of logos that exist as square images.



Fig. 16. Examples of non-tightly fitted image (left) and tightly fitted image (right) of a sample logo.

tightly integrated identity discovery components. These identity discovery components reinforce one another to cope with diverse classes of webpage and irregularities in the content design, allowing the proposed technique to achieve improved robustness and reduced false positives.

4.7. Analysis of misclassified phishing samples

Adversaries normally craft a phishing webpage to resemble the original targeted legitimate webpage (i.e., near duplicate) or intentionally alter its content to increase stealthiness and avoid detection. As a result, phishing webpages are more likely to contain deformed visual styles and incomplete content that do not conform to typical design conventions. Therefore, it is generally harder for a phishing webpage to establish its identity, giving better chances for identity-based anti-phishing techniques to correctly detect it as phishing. This advantage is evidenced by the high true positive rate shown by our proposed hybrid identity-based technique.

To provide better insights into the factors that resulted in false negative detections, the phishing samples were analysed. Findings suggest that one of the factors is due to the presence of a small banner in the webpage that contained the brand name or logo of the free hosting company that the phishing webpage was hosted. This brand name or logo could be mistakenly extracted by our proposed technique and resulted in a false negative detection. Next, we also found a phishing webpage that alleged to represent an arbitrary brand and has successfully gotten itself indexed by Google. Our proposed technique extracted the brand name which was then searched on Google and matched to its own domain listed in the search results. In other words, the phishing webpage seemed to be imitating a non-existent brand.

Nevertheless, we believe that the aforementioned deceptive scenarios are counter-productive to the phishers' intent and thus, are hardly employed by them. By hosting phishing webpages that contain

banners from the free hosting server, the phishers face a greater risk of triggering users' suspicion. In addition, imitating a non-existent brand name is not profitable since the perpetrators cannot utilise the stolen credentials at a real website.

4.8. Undesirable properties of logo image that interfere with search results

Based on our experiments, we found instances where the logo image may contain certain undesirable properties that can jeopardise the results returned by the search engine. These properties and the proposed corrective steps are discussed in the subsections that follow.

4.8.1. The logo image is not tightly fitted

A tightly fitted logo image is defined as a logo image with minimal padding space at the edges of the image. Padding spaces are plain regions that do not contain any distinctive characteristics of the logo image. Examples of a non-tightly fitted logo image and a tightly fitted logo image are shown in Fig. 16.

In Chiew et al. (2015), some logo images were reportedly extracted in non-tightly fitted formats, which degraded the results of Google Reverse Image Search. Hence, our proposed technique employed post-processing on the selected logo image as described in Section 3.1.5 to facilitate effective segmentation of tightly fitted logo images.

4.8.2. The logo image exists in an image sprite

An image sprite is a single image file containing a collection of sub-images. Using CSS properties, web developers can choose which sub-images to be displayed in the webpage. In modern websites, image sprites are commonly used to reduce concurrent network requests. An example of image sprite is shown in Fig. 17. When an image sprite containing the brand logo is submitted to Google Reverse Image Search, interference from non-related sub-images could hinder the

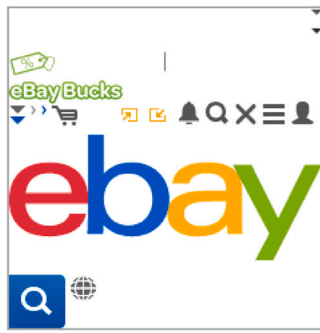


Fig. 17. Example of an image sprite in the eBay¹³ website.

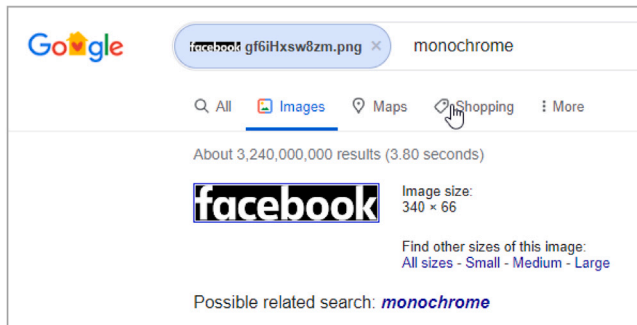


Fig. 18. Outcome of Google Reverse Image Search after a query was performed using a raw image file of the Facebook logo with transparent background.

actual logo from getting recognised correctly (Chiew et al., 2015). Therefore, the non-related sub-images in an image sprite must be discarded. As described in Section 3.1.1, our proposed technique alleviates this problem by manipulating the DOM of a rendered webpage to compute logo image candidates from the visible elements displayed hence the non-related sub-images can be excluded effectively.

4.8.3. The logo image with transparent background

When analysing the raw image files used to load the webpage logos, we discovered some logo images that utilise transparent backgrounds. When submitted to Google Reverse Image Search, the transparent regions will get converted to a solid colour, which may degrade the accuracy of the returned results. Fig. 18 shows a Facebook¹⁴ logo with transparent backgrounds that got converted to black backgrounds after we submitted it to Google Reverse Image Search, resulting in an appearance that is inconsistent with the standard Facebook logo with a blue background, thus affecting the accuracy of the search results. This issue is also addressed in our proposed method using the DOM manipulation approach described in Section 3.1.1, thus avoiding the direct use of inappropriate raw image files that may give rise to aforementioned logo background transparency issues.

5. Conclusion and future works

Conventional phishing detection techniques are struggling to keep up with the continual change of phishing tactics by adversaries, which motivated security researchers to explore more on a particular category of phishing detection technique called the identity-based technique. Although identity-based detection techniques are highly effective in catching phishing webpages, the complexities and challenges in establishing the webpage's brand identity have made these techniques to be

inferior in classifying legitimate webpages. To address this performance gap, we propose a hybrid detection technique that integrates visual and textual identity discovery components to enhance the establishment of the webpage identity. In the visual identity discovery component, novel features that mimic human vision were introduced to facilitate the detection and extraction of the website logo from the query webpage. On the other hand, the textual identity discovery component adopted a prominent identity keywords extraction algorithm from our prior work to assist in reducing the false positive rate. Through the experiments, the proposed technique demonstrated promising performance with an overall detection accuracy of 98.6%. When benchmarked against an existing technique, the false positive rate of our proposed technique has shown up to 3.4% of improvement, which is consistent with our objective of reducing the misclassification of legitimate webpages without sacrificing the phishing detection performance. Such improvement carries a significant practical value as users routinely encounter legitimate webpages with varying designs and visual layouts, hence a low false positive rate shall facilitate a better user experience. Since the hybridisation of visual and textual identities is becoming one of the new anti-phishing research directions, our work advances the emerging body of knowledge and facilitates other anti-phishing researchers to further develop improved methods based on hybrid identities.

As part of our future work, the accuracy and robustness of the proposed logo detection technique can be further enhanced. It is worth exploring additional aspects involved in human vision's identification of website logos, such as the perceived visual complexity and layout of the overall webpage content. To expand the practical merits of the proposed identity-based phishing detection technique, further work may study how distinctive identity properties in phishing websites (e.g., detected logo, target brand name) can be exposed to users in a way such that they increase the users' ability to identify a phishing webpage.

CRedit authorship contribution statement

Colin Choon Lin Tan: Conceptualization, Methodology, Investigation, Software, Data curation, Writing – original draft. **Kang Leng Chiew:** Supervision, Validation, Funding acquisition. **Kelvin S.C. Yong:** Resources. **Yakub Sebastian:** Writing – review & editing. **Joel Chia Ming Than:** Visualization. **Wei King Tiong:** Project administration.

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.

Data availability

Data will be made available on request.

Acknowledgements

The funding for this project is made possible through the research grant obtained from UNIMAS, Malaysia under the Dana Pelajar PhD [Grant No: F08/DPP/1649/2018]. This work is also supported by the Sarawak Foundation Tun Taib Scholarship Scheme, Malaysia.

References

- Aleroud, A., & Zhou, L. (2017). Phishing environments, techniques, and countermeasures: A survey. *Computers & Security*, 68, 160–196.
- Alsharnouby, M., Alaca, F., & Chiasson, S. (2015). Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, 82, 69–82.
- Anti-Phishing Working Group (2020). Phishing activity trends report, 1st quarter 2020. URL: https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf.

¹³ <https://www.ebay.com/>

¹⁴ <https://www.facebook.com/>

- Arachchilage, N. A. G., Love, S., & Beznosov, K. (2016). Phishing threat avoidance behaviour: An empirical investigation. *Computers in Human Behavior*, 60, 185–197.
- Bozkir, A. S., & Aydos, M. (2020). LogoSENSE: A companion HOG based logo detection scheme for phishing web page and E-mail brand recognition. *Computers & Security*, 95, Article 101855.
- Bozkir, A. S., & Sezer, E. A. (2016). Use of HOG descriptors in phishing detection. In *Proceedings of the 4th International Symposium on Digital Forensics and Security* (pp. 148–153).
- Chang, E. H., Chiew, K. L., Sze, S. N., & Tiong, W. K. (2013). Phishing detection via identification of website identity. In *Proceedings of the International Conference on IT Convergence and Security* (pp. 1–4).
- Chiew, K. L., Chang, E. H., Sze, S. N., & Tiong, W. K. (2015). Utilisation of website logo for phishing detection. *Computers & Security*, 54, 16–26.
- Chiew, K. L., Choo, J. S. F., Sze, S. N., & Yong, K. S. (2018). Leverage website favicon to detect phishing websites. *Security and Communication Networks*, 2018, Article 7251750.
- Chiew, K. L., Yong, K. S., & Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Systems with Applications*, 106, 1–20.
- da Silva, C. M. R., Feitosa, E. L., & Garcia, V. C. (2020). Heuristic-based strategy for phishing prediction: A survey of URL-based approach. *Computers & Security*, 88, Article 101613.
- Dalgic, F. C., Bozkir, A. S., & Aydos, M. (2018). Phish-IRIS: A new approach for vision based brand prediction of phishing web pages via compact visual descriptors. In *Proceedings of the 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies* (pp. 1–8).
- Fessenden, T. (2017). Horizontal attention leans left. Nielsen Norman Group, URL: <https://www.nngroup.com/articles/horizontal-attention-leans-left/>.
- Gupta, B. B., & Jain, A. K. (2020). Phishing attack detection using a search engine and heuristics-based technique. *Journal of Information Technology Research*, 13(2), 94–109.
- 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.
- Haruta, S., Asahina, H., & Sasase, I. (2017). Visual similarity-based phishing detection scheme using image and CSS with target website finder. In *Proceedings of the IEEE Global Communications Conference* (pp. 1–6).
- Hasler, D., & Suesstrunk, S. E. (2003). Measuring colorfulness in natural images. In B. E. Rogowitz, & T. N. Pappas (Eds.), *Human vision and electronic imaging VIII*, vol. 5007 (pp. 87–95).
- Jain, A. K., & Gupta, B. B. (2018). Two-level authentication approach to protect from phishing attacks in real time. *Journal of Ambient Intelligence and Humanized Computing*, 9, 1783–1796.
- Jain, A. K., & Gupta, B. B. (2021). A survey of phishing attack techniques, defence mechanisms and open research challenges. *Enterprise Information Systems*, 16(4), 527–565.
- Liu, X., & Fu, J. (2020). SPWalk: Similar property oriented feature learning for phishing detection. *IEEE Access*, 8, 87031–87045.
- Mao, J., Tian, W., Li, P., Wei, T., & Liang, Z. (2017). Phishing-Alarm: Robust and efficient phishing detection via page component similarity. *IEEE Access*, 5, 17020–17030.
- Mishra, A., & Gupta, B. B. (2018). Intelligent phishing detection system using similarity matching algorithms. *International Journal of Information and Communication Technology*, 12(1–2), 51–73.
- OpenCV team (2019). Color conversions. URL: https://docs.opencv.org/4.1.1/de/d25/imgproc_color_conversions.html.
- Peng, P., Xu, C., Quinn, L., Hu, H., Viswanath, B., & Wang, G. (2019). What happens after you leak your password: Understanding credential sharing on phishing sites. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security* (pp. 181–192).
- Rao, R. S., & Pais, A. R. (2019). Jail-Phish: An improved search engine based phishing detection system. *Computers & Security*, 83, 246–267.
- Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., & Gajos, K. Z. (2013). Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2049–2058).
- Su, H., Zhu, X., & Gong, S. (2019). Open logo detection challenge. In *Proceedings of the 2018 British Machine Vision Conference* (pp. 1–13).
- Tan, C. L., Chiew, K. L., Wong, K., & Sze, S. N. (2016). PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder. *Decision Support Systems*, 88, 18–27.
- Ubing, A. A., Binti Jasmi, S. K., Abdullah, A., Jhanjhi, N., & Supramaniam, M. (2019). Phishing website detection: An improved accuracy through feature selection and ensemble learning. *International Journal of Advanced Computer Science and Applications*, 10(1), 252–257.
- Van Dooremaal, B., Burda, P., Allodi, L., & Zannone, N. (2021). Combining text and visual features to improve the identification of cloned webpages for early phishing detection. In *Proceedings of the 16th International Conference on Availability, Reliability and Security* (pp. 1–10).
- VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. O'Reilly Media, Inc..
- Whitenton, K. (2016). Website logo placement for maximum brand recall. Nielsen Norman Group, URL: <https://www.nngroup.com/articles/logo-placement-brand-recall/>.
- Yang, L., Zhang, J., Wang, X., Li, Z., Li, Z., & He, Y. (2021). An improved ELM-based and data preprocessing integrated approach for phishing detection considering comprehensive features. *Expert Systems with Applications*, 165, Article 113863.
- Zerman, E., Rana, A., & Smolic, A. (2019). Colormet - Estimating colorfulness in natural images. In *Proceedings of the 26th IEEE International Conference on Image Processing* (pp. 3791–3795).