Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 436-443, Ekim 2019

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 436-443, October 2019

Research Article

Brand Recognition of Phishing Web Pages via Global Image Descriptors

Esra Eroglu[1*], Ahmet Selman Bozkir[2] and Murat Aydos[3]

[1] Başkent University, Faculty of Commercial Sciences, Department of Management and Information Systems, Ankara, Turkey (ORCID: 0000-0002-6140-6894)
[2] Hacettepe University, Faculty of Engineering, Department of Computer Engineering, Ankara, Turkey (ORCID: 0000-0003-4305-7800)
[3] Hacettepe University, Faculty of Engineering, Department of Computer Engineering, Ankara, Turkey (ORCID: 0000-0002-7570-9204)

(This publication has been presented orally at HORA congress.)

(First received 1 August 2019 and in final form 25 October 2019)

(**DOI:** 10.31590/ejosat.638397)

ATIF/REFERENCE: Eroğlu, E., Bozkır, A. S. & Aydos, M. (2019). Brand Recognition of Phishing Web Pages via Global Image Descriptors. *European Journal of Science and Technology*, (Special Issue), 436-443.

Abstract

Phishing attacks, which have exponentially increased in recent years, are a form of cyber attack aiming to steal sensitive credentials of innocent users. In general, the attackers attempt to deceive users by creating and submitting a fake but visually similar version of a legitimate web page, which has already been in usage. In this study, we suggest an approach for recognition of phishing web pages by utilizing two global image descriptors namely GIST and local binary patterns (LBP) which have never been employed in phishing web page recognition literature. Moreover, in order to obtain a discriminative representation, we have experimented two kinds of visual feature extraction scheme such as (1) "holistic" and (2) "multi-level patches". While we have only used whole web page screenshot in "holistic" scheme, screenshots were divided into equally sized smaller crops at growing number of levels during the implementation of "multi-level" patches scheme. In order to evaluate the proposed approach, we have employed a publicly available phishing web page dataset in literature including screenshots of both 14 different highly phished brands and legitimate web pages posing an open-set problem for researchers. Besides, the aforementioned dataset covers 1313 training and 1539 testing cases in total. The visual signatures extracted by use of GIST and LBP descriptors were then fed to various machine learning models such as SVM, Random Forest and XGBoost (regularized gradient tree boosting). According to the results of comprehensively conducted experiments, XGBoost has been found as the best learner. In line with this finding, we obtained 87.7% (GIST) and 83.1% (LBP) validation accuracy along with the representation of "multi-level patches". Consequently, it has been shown that preferred global image descriptors can be successfully employed for detecting and recognizing phishing web pages. In addition, average required time for processing one screenshot (around 1.12 sec.) with GIST descriptors indicates that the proposed scheme and GIST can be effectively used as a browser based plug-in for recognizing brands of phishing web pages.

Keywords: Phishing, Computer Vision, Machine Learning, GIST, LBP.

Genel Görsel Betimleyicilerden Faydalanarak Oltalayıcı Web Sayfalarında Marka Tanıma

Öz

İnternetin gelişmesiyle son yıllarda katlanarak artan kimlik avı saldırıları, masum kullanıcıların özel kimlik bilgilerini çalmayı amaçlayan bir siber saldırı şeklidir. Genel olarak saldırganlar, kullanımda olan meşru bir web sayfasının sahte ancak görsel olarak benzer

[*] Corresponding Author: Baskent University, Faculty of Commercial Sciences, Dept. of Management of Information Systems, Ankara, Turkey, ORCID: 0000-0002-6140-6894, eeroglu@baskent.edu.tr

bir sürümünü oluşturup kullanıcılara göndererek aldatmaya çalışırlar. Bu çalışmada oltalayıcı web sayfalarının hedef aldığı markaların tanınmasında alanyazınında denenmemiş olan iki genel amaçlı görsel betimleyicinin (GIST ve Local Binary Patterns) kullanıldığı bir yaklaşım önerilmektedir. Buna ilaveten ayırt ediciliği yüksek temsillerin elde edilebilmesi amacıyla "bütünsel" ve "çok seviyeli parçalama" gibi iki özellik çıkarım yaklaşımı denenmiştir. "Bütünsel" yaklaşımda tüm sayfa şipşakı girdi olarak kullanılırken "çok seviyeli parçalama" yaklaşımında tüm görsel, eşit büyüklükteki parçalar içeren çok katmanlı yapıda ele alınmıştır. Önerilen yaklaşımın performans ölçümünde, oltalama saldırılarına sıklıkla maruz kalan toplamda 14 farklı marka ile birlikte özgün web sayfalarına ait sayfa şipşaklarını içeren ve araştırmacılar açısından "açık küme" problemi teşkil eden bir veri kümesi kullanılmştır. Öte yandan, yukarıda belirtilen veri kümesi toplamda 1313 eğitim ve 1539 test örneğini kapsamaktadır. GIST ve LBP betimleyicileri kullanılarak çıkarılan görsel imzalar daha sonra SVM, Random Forest ve XGBoost gibi çeşitli makine öğrenme modellerine girdi olarak sunulmuştur. Kapsamlı deneylerin sonuçlarına göre, XGBoost en iyi sınıflandırıcı olarak tespit edilmiştir. Öte yandan geçerleme verisi üzerinde "çok seviyeli parçalama" temsili kullanılarak doğruluk kriterinde sırasıyla %87.7 (GIST) ve %83.1 (LBP) değerleri elde edilmiştir. Sonuç olarak seçilen genel görsel betimleyicilerinin oltalayıcı web sayfalarını tespit etme ve marka tanımada başarıyla kullanılabileceği gösterilmiştir. Ek olarak, bir sayfa şipşakının ortalama GIST betimleyicisimnden yararlanarak 1.12 saniyede işlenerek sınıflandırılabilmesi önerilen yaklaşımın oltalayıcı web sayfalarının tanınmasında bir tarayıcı eklentisi olarak da etkin ve verimli şekilde kullanabileceğini göstermetedir.

Anahtar Kelimeler: Oltalama saldırıları, Bilgisayarlı Görü, Makine Öğrenmesi, GIST, LBP.

1. Introduction

Phishing is a cyber attack aimed at deceiving users in order to share personal information of innocent users such as passwords, user names and ID numbers. In this kind of attack, web pages visually mimicking to their counrparts are delivered to the users in order to capture their sensitive information. Besides, targeted users are discovered through social engineering tecniques. The general purpose of phishing attacks is financial fraud through imitation. However, there are many different types of this attack and they are usually classified according to who the target and the attacker are. In phishing cloning, an attacker uses a legitimate e-mail that has already been sent and copies its content to a similar e-mail with a link to a malicious site. *Spear phishing* usually targets a specific person or organization. In *Pharming*, an attacker would poison a DNS record and, in practice, redirect visitors to a legitimate website. *Whaling* is a kind of fishing that targets important and wealthy individuals such as CEOs or civil servants [1].

Throughout the world, phishing attacks have become one of the most popular methods used for targeted attacks. According to the reports prepared by Anti Phishing Working Group, in the first quarter of 2019, the total number of detected phishing sites was determined as 180,768. This number was estimated as 151.014 in the fourth quarter of 2018. Brazil was the country with the highest share of attackers by 21.66%, followed by Australia. In addition, the banking sector is ranked first in the number of attacks, the share of attacks on credit institutions increased by 5.23%. It increased to 25.78% compared to the fourth quarter of last year [2].From a worldwide perspective, phishing attacks have been an increasing attack for almost last two decades. Likewise, there exists an on-going race between attackers and anti-phishers. Therefore, it can be deduced that the phishing is still not a solved problem.

Anti phishing solutions can be categorized in many different ways. However, according to Rao and Pais [12], they can be grouped under four categories (Fig.1). These are list-based techniques, heuristic-based techniques, vision-based techniques, and machine-based techniques. List-based techniques, as Google Safe Browsing API [13] employs, divide web pages into black and white lists based on URL information. In this way, the web site that can perform the attack protection is provided. However, due to the fact that a new phishing web page stays operational in a very short time, the blacklist needs to be updated regularly and rapidly leading these kind of solutions vulnerable to "zero-hour" attacks. Broadly speaking, in heuristic-based approaches, information sourced from text, image and URL of web pages are collected and utilized by feature extraction in order to create a decision function built by various machine learning techniques. It should be noted that, rule based methods do also exist in heuriscti based approaches. According to the list-based approach, it is provided to make less mistakes in detecting phishing pages. In the machine-based approach; It focuses on the application of machine learning algorithms such as Random Forest (RF), logistic regression (LR), multilayer perceptron (MLP), Bayesian network (BN), support vector machine (SVM) [17] on features extracted from web pages. These methods can work more efficiently in large data sets, depending on the hand crafted features selected as the feature set [12].

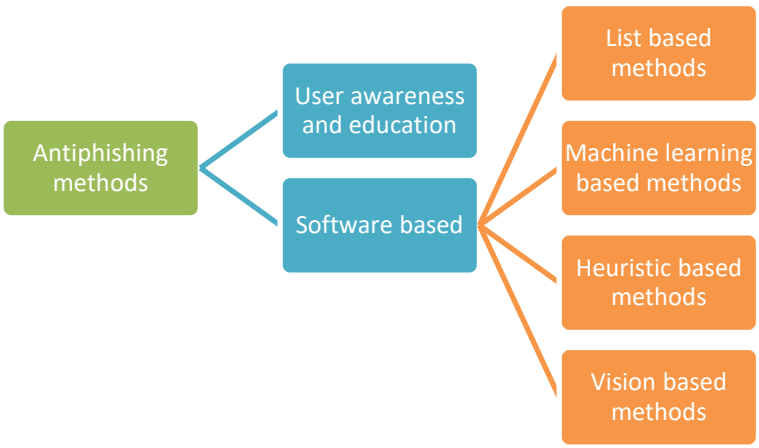

Figure 1. Taxonomy of anti phishing solutions [12]

Recently, since phishing web pages are visually similar to their countparts, vision-based approaches have emerged in order to create effective and efficient classifiers. In general, vision based approaches attempt to extract a visual sigtnature (i.e. feature vector) from the source web pages by utilizing local or global image descriptors. These sigtnatures are either compared or used to create a multi-class classifiers.

Vision based anti-phishing literature covers numerous works employing different underlying approaches. From these studies, [3] have attempted phishing detection bu leveraging machine learning methods along with performing corner analysis using content-based features in a manner of heuristic scheme. Further, in [4], authors have gathered wrapper based features and applied feature selection strategies for phishing detection. So, they have employed the best features in the dataset. In [5], Zhang et al., have suggested an approach which considers spatial layout of web pages. They have constructed an r-tree based indexing technique for determining the visual similarity among the web pages under suspicion. Rao and Ali [6], have proposed a scheme which based on matching SURF features extracted from legitimate and phishing web pages. According to their idea, screenshots of phishing web pages can be identified through SURF based pairwise matching. In another study [8], the images and URL information are utilized in order to detect phishing web pages. For the vision based part, authors have used the "ImgSeek" tool to detect visual similarities between online hosted images and the ones located under investigation. Though, their work is accurate, the proposed approach requires a third party service and the effectiveness is highly dependent on query and retrieval quality of the dependent service. In [9], visual similarity between suspicious and legitimate web page pairs have been studied through earth mover's distance metric (EMD), a measure of the distance between two objects. Although, their results are satisfying, their proposal is not scalable due to underlying feature extraction and optimization strategy. In another vision based study [10], a scale and rotation invariant descriptor namely CCH (i.e. Color Context Histogram) has been used to find visual similarities between legitimate and suspicious web pages. Apart from vision based works, there also exist studies utilizing different source of information or tecniques such NLP [12] and blacklisting [7]. Sahingoz et al., have curated more than 20 handcrafted features which will be extracted from only URL of web pages. Most of these features have been first extracted through NLP methods and were fed to Random Forest classifiers. Their detection accuracy has been reported over 97%. However, these features are prone to be easily discovered by attackers leading to a vulnerable detection mechanism.

Compared to conventional methods such as blacklisting or heuristic methods, computer vision based approaches in phishing detection have some advantages. First of all, in order to gain credit, phishing assets must mimic to their legitimate counterparts. Otherwise, users can easily understand that they are surfing on a fake version of the targeted web page. Second, vision based methods are generally robust to content manipulations carried out by phishers. In other words, vision based analysis is invariant to the underlying HTML source code and tricky web element substitions such as replacing text parts with image/flash based contents. Third, vision based methods consider only the rendered web page screenshot which yields an invariance to HTML versions. Fourth, vision based studies constitute a robust scheme against zero-hour attacks which is a big shortcoming of blacklist based methods. As a disadvantage, vision based methods are resource consuming methods which make them hard to be employed in a high throughput backend.

In this study, we suggest a phishing detection and brand recognition mechanism by employing two global image descriptors (i.e. GIST and LBP) which have been widely used in computer vision. According to our best knowledge, this study is the first for employing these descriptors in an anti phishing scheme. Moreover, we have applied two different feature extraction scheme: (1) holistic and (2) multi-level patches in order to gain more discriminative information from the rendered web page screenshots. Experiments and evaluations carried out on a publicly available dataset (i.e. Phish-Iris dataset) along with use of three diffent machine learning methods (SVM [17], Random Forest and XGBoost [14]) have revealed that, GIST [15] based features have outperformed the LBP ones in terms of accuracy, true positive and false positive rate. Moreover, the run-time speed of GIST+XGBoost based inference has been found suitable for various environments such as plug-in in web browsers or e-mail servers.

The rest of this paper is organized as follows. In section 2, the utilized image descriptors and the way we represent them have been demonstrated. Section 3 briefly introduces the dataset we have used during the experiments. In section 4, details of methodology and application are presented. Next, section 5 reports the results of the experiments. Section 6 serves a comparative study carried out with Histogram of Oriented Gradients [11]. Finally section 7 concludes the paper.

2. Generating and Representing Visual Signatures of Web Pages

In this study, we have employed two different global image descriptors. The global image descriptors in computer vision, deal with generating a discriminative descriptive feature vector extracted from whole input image. The produced image descriptors can than be used for various purposes such as pair-wise similarity or dissimilarity comparison and data driven machine learning applications. Our approach is actually based on inducting machine learning models with feature vectors obtained via descriptors of GIST and local binary patterns. The rest of this subsection briefly introduces the details of these descriptors and representation of "the multi-level patch" inspired from the work of [20]

2.1. GIST Descriptor

The GIST descriptor is used to solve the problem of object identification by focusing on the outline of an object in the field of computer vision. In the study conducted by Oliva and Torralba in 2001, spatial envelope of a scene or image according to various characteristics were identified [15]. The concept of spatial envelope mentioned in the study is a low-dimensional representation of a

scene showing the correlation between the framework of the surface and the properties of the objects in it. The spatial envelope in the model is similar to the characteristics of the known space in everyday life. In other words, it must contain the objects of certain shapes and sizes within certain dimensions. The spatial envelope is in fact the problem of scene classification, and taking into account a number of characteristics of the scenes, combining similar scenes will create a solution to the problem. Therefore, 5 basic features, which are naturalness, openness, roughness, expansion and ruggedness representing the structure of the space, were needed [15].

$$G_{\theta_i}^{s} = Cexp\left(\frac{-\left(x_{\theta_i}^2+y_{\theta_i}^2\right)}{2\sigma^{2(s-1)}}\right) exp\left(2\pi j(u_0 x_{\theta_i} + v_0 y_{\theta_i}\right) \quad (1)$$

$$x_{\theta_i} = xcos\theta_i + ysin\theta_i \quad y_{\theta_i} = -xcos\theta_i + ycos\theta_i \quad (2)$$

As given in (1) and (2), in order to perform feature extraction using the GIST descriptor, first, the image is separated into $n \times n$ blocks to prevent loss of information and to extract the correct properties. Each block is processed by Gabor filters in different scales and directions. Then, a vector is obtained by collecting values from blocks [16, 17]. An image consists of 3 color channels (R, G, B) and is defined as a 4×4 dimensional spatial cell and consists of 1 GIST descriptor 2 finer 8 orientations and 1 coarser 4 orientations. Accordingly, the vector obtained from the GIST descriptor of the one image is 3x (4x4) x (8 + 8 +4) = 960 [2, 3]. The GIST descriptor has been used to identify traffic scenes in another study [18]. Considering the characteristics of *openness* and *naturalness*, it is ensured that the motorways are separated from other roads and closed spaces [18].

2.2. Local Binary Patterns (LBP)

In 1996, Ojala et al. [19] first developed the Local Binary Patterns (LBP) algorithm for pattern classification. LBP has been applied to many computer vision tasks, including face recognition, pedestrian detection, and scene categorization. The LBP algorithm identifies each pixel with two codes and analyzes the textures of a local patch by comparing the center pixel to neighboring pixels.

LBPs construct local representations of textures via comparing each pixel with its surrounding neighborhood of pixels. In order to create a LBP descriptor we first convert the input image to a single channel grayscale format. Followingly, for each pixel in input image, we choose neighborhood of size r surrounding the center pixel. Next, LBP value is computed for this center pixel and saved as two dimensional array having the same width and height. As illustrated in Fig. 2, for a fixed 3 × 3 neighborhood of pixels on a grid, we take the center pixel (highlighted in red) and assume it as threshold value against its neighborhood of 8 pixels. If the intensity value of the center pixel is greater-than-or-equal to its neighbor, then its value is set to 1. In contrast, if it is less than the center pixel, it is set to 0. In this way, by counting in clock-wise or counter clock-wise fashion we obtain a 8 bit binary feature vector that yields an integer value ranging between 0 and 255. We can then generate a single feature vector summarizing whole input image having 256 bins.

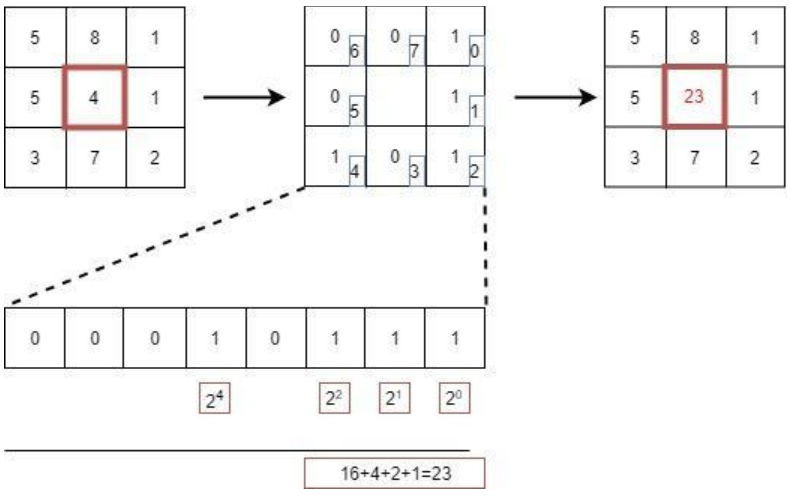

Figure 2. Example of the LBP algorithm

2.3. Multi-level Patch Representation

Though, the purpose of global imge descriptors is to eventually generate a feature vector by considering the whole image given as input, we additionaly propose to use a finer grained and spatial information preserving multi-level patch representation. Indeed, this concept is initially suggested by the seminal work of Lazebnik et al. [20] which enables generating spatial information preserving *spatial pyraming matching* scheme. Nonetheless, Lazebnik et al. [20] has suggested this idea to be employed with local image descriptors such as Scale Invariance Robust Features (SIFT). In this sense, they enabled to accumulate bag of visual words (i.e. histogram generation) by also preserving their spatial relation by dividing the spatial feature space into equal sized rectangular regions within a growing number of levels. Thus, for each succeeding level, we could have more "cells" for better feature localization. With this improvement, captured visual cues were enabled to be more accurately matched.

Similarly, we have adopted this idea to capture both holistic and finer details on web page screenshots by dividing the 2D image screenshots into 2×2 and 3×3 segments and build a concatenated feature vectors for each screenshot sample. This procedure has been visualized and illustrated in Fig. 3 given below.

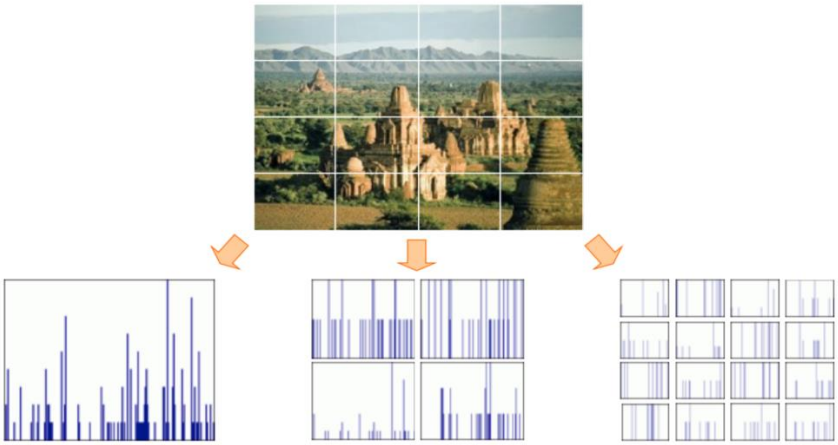

Figure 3. Example of spatial multi-level patch pyramid scheme [20]

3. The Dataset

As stated before, there exists numerous works in anti-phising literature. However, the number of vision based suggestions is relatively low [21]. Moreover, due to the fact that this study focuses on web page screenshots, the necessity of a suitable and labeled dataset is crucial. For this reason, we have searched in literature and have found a suitable dataset called as "Phish-Iris" provided by [21] which involves 2852 screenshot samples in total covering 14 distinct highly phished brands and legitimate instances. Note that, "Phish-Iris" dataset is a publicly and freely available dataset for academic purposes and it can be downloaded from the URL of "https://web.cs.hacettepe.edu.tr/~selman/phish-iris-dataset/" According to the definitions of the dataset creators, "Phish-Iris" dataset has been collected between the March-May 2018. The distribution of the brand samples in both training and testing groups were given in Table 1 below.

Table 1. Phish-IRIS Dataset

Brand Name	Training Samples	Testing Samples
Adobe	43	27
Alibaba	50	26
Amazon	18	11
Apple	49	15
Bank of America	81	35
Chase Bank	74	37
Dhl	67	42
Dropbox	75	40
Facebook	87	57
Linkedin	24	14
Microsoft	65	53
Paypal	121	93
Wellsfarno	89	45
Yahoo	70	44
Other	400	1000
Total	1313	1539

4. The Application and Experiments

In this study, firstly, by use of GIST descriptor, visual feature extraction was performed. To this end, we first obtained 960 dimensional feature vectors by extracting GIST descriptors in a holistic manner. Second, we have applied multi level patching in order to produce finer detailed image descriptors that will eventually build single concatanated and larger "multi-level" feature representation. At this stage, each image has been recursively divided into 2×2=4 and 3×3=9 equal parts. Hence, for the multi-level representation, we have processed either 1+4 = 5 or 1+4+9 = 13 patches in total. As a result, we have eventually generated descriptions of screenshots by employing either single holistic fashion or multi-level pyramidal like scheme. For GIST features, we have utilized "pyleargist" library developed for Python programming language. The same procedure have also been followed during the phase of LBP description generation. These procedures have been depicted in Fig 4.

Following to feature vector generation, we have built three classification models including Random Forest, Support Vector Machine and Xgboost methods for predicting the class of screenshot samples in test group.The training process was performed on a

computer equipped with an Intel® Core ™ i7 4700HQ processor and 16 GB of memory. The machine learning modeling has been carried out on Ubuntu platform by employing several Python libraries such as Scipy, Numpy, Matplotlib, Pandas and Sklearn. Detailed experiments conducted with different parameters were performed and the best results were examined in the next section.

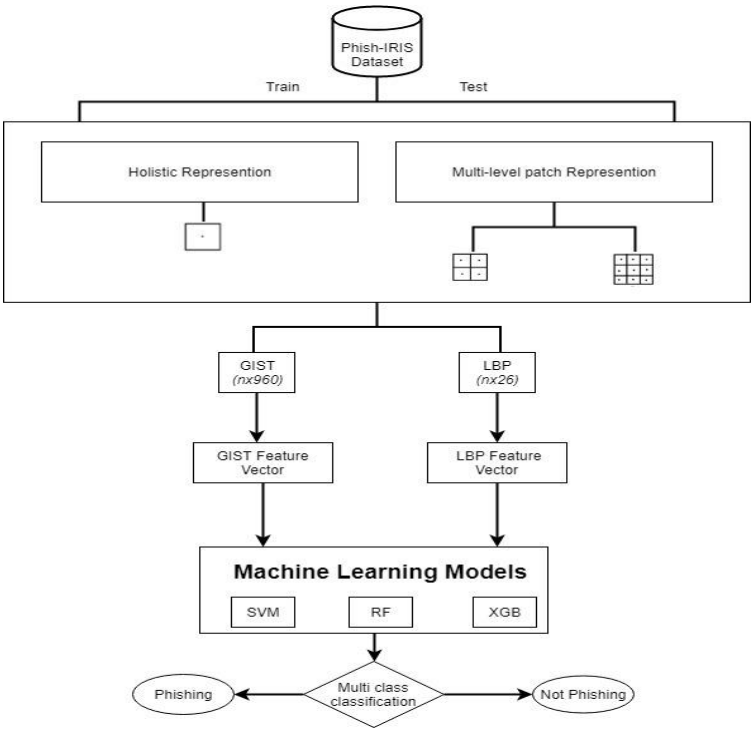

Figure 4. The workflow of the proposed study

5. Results and Discussion

This section gives experimental results of the classification algorithms regarding the underlying descriptor and representation scheme. We have both tested the training and testing dataset with created models. Moreover, the evaluation has been carried out by considering the metrics of accuracy, true positive rate (TPR), false positive rate (FPR) and F-1 measure. According to the results obtained with the GIST descriptor + XGBoost learning model, we have maximally achieved 87.771% accuracy along with 0.0084 FPR on test cases (See Table 2). Besides, as can be seen in Table 3, the LBP descriptor based modelling achieved the highest accuracy rate as 83.1% on XGBoost learner. Both of these best results have been obtained by using 3 levels in multilevel patch representation.

Table 2. Experimental results regarding the GIST Descriptor

Descriptor - Level	Learner	# of Features	Train Acc.	Test Acc.	TPR	FPR	Fl
GIST – Holistic	SVM	960	0.533	0.746	0.7465	0.018	0.75
GIST – Holistic	XGB	960	0.732	0.8583	0.8583	0.010	0.86
GIST – Holistic	RF	960	0.740	0.860	0.860	0.009	0.86
GIST – 2 levels	SVM	4800	0.549	0.757	0.757	0.017	0.76
GIST – 2 levels	XGB	4800	0.757	0.8719	0.8719	0.0091	0.87
GIST – 2 levels	RF	4800	0.755	0.858	0.858	0.01	0.86
GIST – 3 levels	SVM	13440	0.568	0.7868	0.786	0.01	0.79
GIST – 3 levels	XGB	13440	0.779	**0.8771**	0.8771	**0.0084**	**0.88**
GIST – 3 levels	RF	13440	0.768	0.8739	0.8739	0.009	0.87

Table 3. Experimental results regarding the LBP Descriptor

Descriptor - Level	Learner	# of Features	Train Acc.	Test Acc.	TPR	FPR	Fl
LBP – Holistic	SVM	26	0.272	0.629	0.629	0.0266	0.63
LBP – Holistic	XGB	26	0.602	0.751	0.751	0.0177	0.75
LBP – Holistic	RF	26	0.631	0.784	0.784	0.015	0.78
LBP – 2 levels	SVM	130	0.338	0.638	0.638	0.025	0.64
LBP – 2 levels	XGB	130	0.687	0.798	0.798	0.0143	0.8
LBP – 2 levels	RF	130	0.711	0.827	0.827	0.012	0.83

LBP – 3 levels	SVM	364	0.372	0.6621	0.662	0.024	0.66
LBP – 3 levels	XGB	364	0.732	**0.831**	0.831	**0.12**	0.83
LBP – 3 levels	RF	364	0.733	0.825	0.825	0.00124	0.83

A comparison covering GIST and LBP based results reveal that multi-level representation has a greater impact on LBP features. This implies that, extracting more detailed information has a positive impact when dealing with LBP based analysis. However, this regime does not significantly hold for GIST based study. As can be inferred from Table 2, working with more levels does not contribute much for GIST based learning. On the other hand, number of features in GIST based modeling indicates that training duration for GIST is higher than LBP since the representation for GIST requires much larger feature vectors due to concatenation. One another key finding is that, GIST based inference took 1.2 seconds for single image in average.

6. Comparative Study

In order to better reveal the effectiveness of the proposed scheme, we have conducted a comparative study by employing HOG (Histogram of Oriented Gradients) [11] descriptors. By definition, the HOG features produce a gradient based visual cues for revealing the corner-edge characteristic of the input image. In particular, HOG descriptor divides an image detection window into small connected regions called cells and calculates the histogram of the gradient directions or edge directions of the pixels within the cell for each cell followed by a normalization stage. Throughout the comparative study, we utilized the same data set and the same machine learning methods. We have either resized or cropped the input screenshot in order to have canonical input resolution which is a requirement for HOG based feature extraction. Cropping process yields an information loss at edges of screenshots whereas resizing distorts the edge structures. Furthermore, we have also preferred two different cell sizes (i.e. 32 and 64 pixels) which directly affect the performance of the obtained feature vectors. Therefore, we applied these two tecniques and obtained detailed results as can be seen in Table 4.

Table 4. Prediction results with HOG descriptors

Descriptor – Cell Size – Mode	**Learner**	**Train Acc.**	**Test Acc.**	**TPR**	**FPR**	**F1**
HOG – 32px cells – Cropped	XGB	0,714	0,8349	0.834	0.011	0.82
HOG – 32px cells – Cropped	RF	0,693	0,8258	0.825	0.012	0.81
HOG – 32px cells – Cropped	SVM	0,596	0,7543	0.754	0.017	0.73
HOG – 32px cells – Resized	XGB	0.719	**0,8408**	0.840	0.011	0.83
HOG – 32px cells – Resized	RF	0,71	0,8395	0.830	0.011	0.82
HOG – 32px cells – Resized	SVM	0,626	0,7673	0.767	0.016	0.75
HOG – 64px cells – Cropped	XGB	0,729	0,8245	0.824	0.012	0.81
HOG – 64px cells – Cropped	RF	0,705	0,8317	0.831	0.012	0.82
HOG – 64px cells – Cropped	SVM	0,579	0,74	0.74	0.018	0.72
HOG – 64px cells – Resized	XGB	0,747	0,8304	0.830	0.012	0.82
HOG – 64px cells – Resized	RF	0,722	0,8369	0.836	0.011	0.82
HOG – 64px cells – Resized	SVM	0,597	0,7563	0.756	0.017	0.74

According to the results, HOG features achieve 84.08% accuracy at best configuration. Experimental study reveals that Random Forest and XGBoost produce slightly similar results. Nevertheless, SVM (RBF kernel) has been clearly outperformed by RF and XGBoost learners. Compared to the best model created with HOG features, GIST based analysis is superior to HOG and LBP.

7. Conclusion

In this study, a new vision based multi-class phishing web page recognition scheme has been proposed and developed. For this purpose, we have utilized two different global image descriptors namely GIST and Local Binary Patterns. To our best knowledge, we have employed these two descriptors for the first time in phishing field. Furthermore, we have applied two distinct respresentation schema for visual signature generation. Detailed experimentation shows that GIST descriptors surpass the LBP based modeling in terms of accuracy, TPR and FPR. One another finding we have explored is that, along with having higher accuracy rate, XGBoost has several advantages such as GPU based training. The short duration of GIST based inference makes it a suitable, lightweight and practical scheme for being used as the first stage classifier in phishing detection mechanisms. As a future work, we plan to use convolutional neural networks for generating single yet deep feature vectors for better generalization and improved accuracy.

8. Acknowledge

Competing interests: The authors declare that they have no competing interests.
Authors' contributions: All authors read and approved the final manuscript.

References

[1] Jain, A. K., & Gupta, B. B. (2017). Phishing detection: analysis of visual similarity based approaches. Security and Communication Networks, 2017.
[2] Phishing Activity Trends Report 1st Quarter 2019, www.apwg.org • info@apwg.org
[3] Basnet, R. B., & Sung, A. H. (2014). Learning to Detect Phishing Webpages. J. Internet Serv. Inf. Secur., 4(3), 21-39.
[4] Ali, W. (2017). Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection. International Journal of Advanced Computer Science and Applications, 8(9), 72-78.
[5] Zhang, W., Lu, H., Xu, B., & Yang, H. (2013). Web phishing detection based on page spatial layout similarity. Informatica, 37(3).
[6] Rao, R. S., & Ali, S. T. (2015, April). A computer vision technique to detect phishing attacks. In 2015 Fifth International Conference on Communication Systems and Network Technologies (pp. 596-601). IEEE.
[7] Prakash, P., Kumar, M., Kompella, R. R., & Gupta, M. (2010, March). Phishnet: predictive blacklisting to detect phishing attacks. In 2010 Proceedings IEEE INFOCOM (pp. 1-5). IEEE.
[8] Hara, M., Yamada, A., & Miyake, Y. (2009, March). Visual similarity-based phishing detection without victim site information. In 2009 IEEE Symposium on Computational Intelligence in Cyber Security (pp. 30-36). IEEE.
[9] Fu, A. Y., Wenyin, L., & Deng, X. (2006). Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD). IEEE transactions on dependable and secure computing, 3(4), 301-311.
[10] Chen, K. T., Chen, J. Y., Huang, C. R., & Chen, C. S. (2009). Fighting phishing with discriminative keypoint features. IEEE Internet Computing, 13(3), 56-63.
[11] Dalal, N., Triggs, B. (2005). Histograms of oriented gradients for human detection. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA
[12] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345-357.
[13] Google Safe Browsing API, https://developers.google.com/safe-browsing/ (Online accessed: 13.7.2019)
[14] XGBoost Documentation, https://xgboost.readthedocs.io/en/latest/ (Online accessed: 13.7.2019)
[15] Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision, 42(3), 145-175.
[16] Wang, Y., Li, Y., & Ji, X. (2013, December). Recognizing human actions based on gist descriptor and word phrase. In Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC) (pp. 1104-1107). IEEE.
[17] Corinna Cortes, Vladimir Vapnik, "Support-vector networks", Machine learning, vol. 20, no. 3, pp. 273-297, 1995.
[18] Sikirić, I., Brkić, K., & Šegvić, S. (2013). Classifying traffic scenes using the GIST image descriptor. arXiv preprint arXiv:1310.0316.
[19] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions”, Pattern Recognition, vol. 29, pp. 51-59, 1996.
[20] Lazebnik, S., Schmid, C., & Ponce, J. (2006, June). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 2169-2178). IEEE.
[21] Dalgic, F. C., Bozkir, A. S., & Aydos, M. (2018, October). Phish-IRIS: A New Approach for Vision Based Brand Prediction of Phishing Web Pages via Compact Visual Descriptors. In 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-8). IEEE.