Journal of Theoretical and Applied Information Technology
20th November 2015. Vol.81. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SELECTION OF ROBUST FEATURE SUBSETS FOR PHISH WEBPAGE PREDICTION USING MAXIMUM RELEVANCE AND MINIMUM REDUNDANCY CRITERION

[1, 2]HIBA ZUHAIR, [3]ALI SELAMAT, [3]MAZLEENA SALLEH

[1]PhD Candidate, Faculty of Computing, Universiti Teknologi Malaysia, Malaysia.

[2]Senior Lecturer, Al-Nahrain University, Baghdad, Iraq

[3] Prof., Faculty of Computing, Universiti Teknologi Malaysia, Malaysia.

[4] Assoc. Prof, Faculty of Computing, Universiti Teknologi Malaysia, Malaysia.

E-mail: [1, 2]hiba.zuhair.pcs2013@gmail.com, [3]aselamat@utm.com, [4]mazleena@fsksm.utm.my.

ABSTRACT

Phishers usually evolve their web exploits to defeat current anti-phishing community. Accordingly, that becomes a serious web threat and puts both users and enterprises at the risks of identity theft and monetary losses day by day. In the literature, most computational efforts were dedicated to justify well-performed phishing detection against evolving phish exploits. However, facets like exploration of new and predictive features, selecting minimal and robust features compactness still raise as key challenges to optimize the detection scenarios over vast and strongly interrelated web. In this study, we proposed a set of new hybrid features, and refine it as few, maximum relevant, minimum redundant, and robust features as possible. In the presence of a machine learning classifier and some assessment criteria that recommended for this purpose, the reported results experimentally demonstrated that our remedial scenario could be used to optimize a phish detection model for any anti-phishing scheme in the future.

Keywords: *Hybrid Features, Maximum Relevance, Minimum Redundancy, Goodness, Stability, Similarity.*

1. INTRODUCTION

In the last decade, cyberspace has shown a rapid expansion of phishing. Phishers try to target users and enterprises to access their sensitive information. They imitate legitimate websites with some deceptive features to build their phishes [1] - [3]. Moreover, they continually evolve phishes by exploiting more sophisticated features in different feature spaces, such as webpage URLs and content. Thus, they can circumvent the existing phish detect approaches, causing more potential risks and monetary losses [4], [5]. Most of the literatures focus on methods of surviving phish attacks, hosted in webpages and the ways to improve the existing phishing detective approaches, such as the list-based, heuristics, hybrid and information flow-based methods [1], [2], [6]-[15]. The hybrid detective approaches somewhat outperform other approaches due to the use of classifiers and multiple types of features, i.e. hybrid features [16]-[40]. Thus, exploitation of new and most predictive features besides classification models is usually emerged as a problematic issue specifically against evolving phishing deceptions. Because new and predictive features may enrich the detective approach to effectively handle evolving deceptions over the rapidly distributed phish webpages over the Web [12]-[15]. A common way to deal with such issue is to assist the detective approach with a feature selection technique that chooses the most contributing features or feature subsets as well as exploring new features. For this reason, some prior researchers developed their proposals with the aid of feature selection techniques [10], [23]-[28].

Despite of their impressive achievements, such developments still sub-optimal perform against vast and evolving data like the web. We observed that it becomes an intricate issue to deal with hundreds billions of evolving web pages that strongly interrelated with a large number of features [41] -

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[43]. Mainly, high dimensional set of features may be inn-productive because it contains either irrelevant features or redundant ones; or both with respect to a specific class [44]. More importantly, the number of possible obtained feature subsets increases as the feature set's dimensionality increases. That, in turn, leads to ineffective and costly classification which is an intractable problem in phishing detection [45] -[47].

In this course, this study attempts to find the most advantageous feature subset (i.e. A near optimal feature subset) in terms of maximal relevance and minimal redundancy at once. To do so, it sought for new features that could be crafted by phishers, and experimentally investigated the most predictive ones by eliminating both the least relevant and most redundant features simultaneously. Through investigation and experimentation, new 58 hybrid features were extracted and then they were refined to subsets of highly productive features by using a distinct feature selection criterion. In addition, the obtained feature subsets were assessed in terms of their robustness and effectiveness in the presence of certain evaluation measures. It is hoped that this recommended scenario for feature selection, and robustness and effectiveness evaluation could help to optimize detection models for any existing hybrid based phishing detective scheme in the future. At present, our study focuses on investigating new hybrid features, selecting as few and effective features as possible, and emphasizing their significance assessment to promote phishing classification task. Other facets like real-world application is kept constant now, but it will be investigated in our future work.

The rest of this study is organized as follows. Section 2 briefly surveys the previous works in phishing detection domain; whilst, Section 3 critically appraises them. Then, Section 4 presents the promoting features as well as the supporting criteria for features selection, phishing induction and evaluation. Section 5, addresses the strategy and execution of dedicated experiments. To give a global insight on their outcomes, the experimental results are reported and discussed in Section 6. Finally, conclusions and future perspectives are presented in Section 7.

2. RELATED WORK

In the literature, various phishing detective approaches have been proposed by researchers to mitigate the increased phishing susceptibility. In general, researchers have categorized anti-phishing techniques into several groups due to the exploited features, detection scenarios, and information sources. For example, Han et al. [8] Have decomposed them into blacklist, whitelist, heuristic and hybrid-based approaches with respect to detection techniques and features they encompassed. On the other hand, Shahriar and Zulkernine [9] categorized them into: whitelists, blacklists, hybrid, standalone and random-based techniques due to the utilization of information sources and features. Contrarily, Islam and Abwajy [10] roughly isolated them into non-classification and classification based techniques due to the use of machine learning classifiers for phishing detection. Generally speaking, anti-phishing techniques can be categorized into non-classification based techniques such as: white lists of famous trustworthy URLs, blacklists of valid phish URLs, rule-based techniques and information flow; and classification-based techniques; namely hybrid techniques that assisted by machine learning and data mining techniques along with the usage of hybrid features. Given that machine learning classifiers outperform other techniques in many application domains. Most researchers have adopted them to develop intuitive phishing detection and prevention [3], [11]. Furthermore, some researchers have mostly relied on various machine learning classifiers and constructed them in single and ensemble design [1], [2]. In turn, the constructed classifiers could automatically examine a set of extracted hybrid features such as those of URL, web content, hosting information and online features [12]-[14]. In the light of classification based anti-phishing techniques, **Table 1** enlisted examples of them with their relative merits and demerits.

As depicted in **Table 1,** a Bayesian filter was developed by Likarish et al. [16] to identify phish websites based on retrieving tokens from the HTML document and constructing DOM (Document Object Model) with the aid of DOM parser. Then, researchers at Google Inc., Whittaker, Ryner & Nazif [17]; worked on the up-gradation of Google's phishing blacklist integrated with a classifier. Alongside, another anti-phishing technique was developed by Bergholz et al. [18] for phish email filtering by analysing several extracted features related to body, external and model-based on examining emails. The developed techniques involved two training phases one for model-based features and the other for the rest of the features.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Later, CANTINA+ was proposed by Xiang et al. [19] with the use of three classifiers and ten features derived from URLs and contents of the webpage as well as some online features for highly accurate results of detection on phishes. Meanwhile, Zhang et al. [20] introduced a linear classifier *Naïve Bayes* (*NB*) in order to detect eight textual and visual features on suspected websites for phishness prediction. The used classifier returned a normalized number reflecting the likeliness of the suspect website to be phished or non-phished. Likewise, a *Supervised Machine Learning* (*SVM*) classifier was developed by He et al. [6] to predict phishness on examined webpage by exploiting webpage identity and some textual features. Textual features are extracted using a well-known information retrieval method to be deployed in the classification process. Contrarily, a phish webpage detector was proposed by Li et al. [7] Based on visual features and DOM objects on the webpage content that learned and tested over datasets by using Semi-Supervised Machine Learning (*TSVM*) classifier. Further, Kordestani & Shajari [21] applied three classifiers including Naïve Bayes (*NB*), Supervised Machine Learning (*SVM*) and Random Forest (*RF*) on a randomly selected dataset to predict phishes in suspected websites. They were deployed for phishness prediction with the presence of URL and online features. Then, Gowtham & Krishnamurthi [22] extracted fifteen features that were trained by using Supervised Machine Learning (*SVM*) classifier and a whitelist through two modules. The first module involved identifying features of the examined website against a pre-defined whitelist of legitimate ones. The second module predicts phishness of the examined webpage according to its login form features via *SVM* classifier. However, the application of the aforesaid proposals encountered some trade-offs related to the processing of large and realistic data sets, the extraction of hybrid features, the analysis of their heterogeneity, increasing storage requirements and processing time as well as some costly misclassifications. These trade-offs are degraded within the performance of such proposals and made phishing detection more prohibitive.

On the other hand, the researchers made their final decisions based on the potentiality of deployed features to predict phishness with minute amounts of valid phish misclassifications and loss of valid non-phish instances. They maintain some feature selection methods to cope with a high dimensional feature space. For instance, Pan and Ding [23] proposed phishing detector based on applying Supervised Machine Learning (*SVM*) classifier and extracting both textual and Document Object Model (*DOM*) features from the examined webpages. They employed two major components of their detector including an information retrieval strategy to extract textual features and Chi-squared (χ^2) criterion to select the most effective features. Then, Ma et al. [24] experimentally analysed seven webpage and page rank features with the aid of a feature weighting method for phish website classification and deployed two classifiers that varied in their classification accuracy due to the selected features. Later, Toolan & Carthy [25] evaluated 40 features that are mostly used in the literature for both phish and spam e-mails filtering. They ranked the most informative ones among three datasets by using Information Gain (*IG*) analysis. The prediction accuracy was varied among all the three datasets due to the selected set of features in the presence machine learning classifier. Khonji, Jones & Iraqi [26] enhanced classification performance by selecting the most effective subset of most commonly used 47 features. All Information Gain (*IG*), Wrapper Feature Based Selection (*WFS*) and Correlation Based Feature Selection (*CFS*) were deployed with classifiers to predict phish emails. The classification results differed due to the used feature selection method and the number of selected features., Alongside, Basnet, Sung & Liu [27] analysed high dimensional feature space, including 177 features extracted from both the content and the URL of websites to select the best feature subset. Several subsets are considered using Wrapper Feature Based Selection (*WFS*) and Correlation Based Feature Selection (*CFS*) feature selection methods. They trained over dataset with the aid of Logistic Regression (*RF*) classifiers. But the selection of contributing features varied among different feature selection methods and classifiers causing different detection results. Then, Zhang et al. [28] developed an automatic detection approach for Chinese e-business websites by incorporating unique features extracted from the URL and contents of the website. The extracted features were further trained and tested via four classifiers including Logistic Regression (*RF*), Naïve Bayes (*NB*), Random Forest (*RF*) and Sequential Minimum Optimization (*SMO*). Features were evaluated using Chi-squared (χ^2) statistic criterion based on the used classifiers. Even though the aforesaid studies considered some traditional feature selection methods to rank the most predictive features, they rarely addressed the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

problems of features' irrelevance and redundancy that encountered when features were being combined in a feature subset. Specifically, this became an intricate issue in dealing with a huge dataset of real-world phishing. Capturing, processing and classifying such retrieved datasets from the web are really exhaustive. Web page classification with a high certainty of phishing features causes irrelevance and redundancy problems and involves huge computational costs. Therefore, the aforesaid proposals continue to be limited in terms of detection accuracy, performance, the rate of sensitivity, and interpretability to the evolving phishes.

Continuous development of more effective anti-phishing techniques with the key factors of zero sensitivity and optimum phish detection became an urgent necessity. It is acknowledged that improvement in the detection capability of classification technique can be maintained by using multi-tier classifier (i.e. an ensemble classifier). For instance, Aburrous et al. [29] designed phishing detector, particularly for e-banking websites by using an ensemble classifier which was composed of both K-nearest neighbour (*K-NN*) and Supervised Machine Learner (*SVM*) to obtain better detection results. Further, Zhuang, Jiang & Xiong [30] developed a detection model comprising of several phases such as feature extractor, training, ensemble classifier, and cluster training. The proposed model relies on extracting hybrid features from webpages and training them by using ten classifiers built as an ensemble classifier to achieve better prediction. Later, Hamid & Abwajy [31] proposed a multi-tier detector for phish emails filtering with the aid of *AdaBoost* and Sequential Minimum Optimization (*SMO*) classifiers in an ensemble design. Moreover, they used clustering strategy to set profiles of the best predictive features and they tested them across three large scale datasets. Some critical limitations, including large size and imbalanced datasets, the limit of cluster size and error rates are encountered. Even though, those prior researchers have empirically proven that their proposals could outperform their competitors for phishing detection. Their proposals still encountered some trade-offs related to handling different categories of features along with their increased storage requirements, processing time and costly misclassifications. More precisely, their proposed works were limited in dealing with big and realistic data like that of the web. Further, they still have shortages on how to mitigate the rapid vastness and advancement of phishing deceptions and activities over the web. Such shortages might have degraded the performance of their proposal works and made phishing detection more prohibitive. Upon such problematic voids, a customary emphasis of the aforesaid shortages is strongly discussed in the next section.

3. KEY CHALLENGES

In the reviewed literature, the exploited features for phishing prediction were roughly characterized into webpage content, URL and online features on the basis of their nature and parts of the webpage where they were exploited [32]-[41]. However, these features and categories vary in their prediction susceptibilities against phishing deceptions. Each feature category may have negative impacts on the overall performance of anti-phishing and future implications regarding to its limited prediction susceptibility as described in **Table 2** [3], [9], [14], and [22]. On the other hand, evolving phishing deceptions particularly involve new features crafted by phishers to bypass the current anti-phishing community because some of them were rarely considered and identified in the literature. Today, phishers can impersonate their target websites by hiding some links for users' redirection to their own fake webpages, and obfuscating the client-side scripting components like JavaScript, PHP and ASP, etc. As such phishers are able to install suspicious, malicious and spy codes into the client's computer for further damages; and create multiple replicas of their targets for pharming purposes, i.e. redirecting as many visitors as possible to the same fake website. Moreover, they modify some applets, Flash objects and ActiveX controls in the source file of their targets to submit their cookies and fake advertisements through the web banners [3], [9], and [14]. They also target the URLs of webpages presented in any language rather than English, e.g. Chinese e-business web pages [9] and [30]. The commonly used features could not be potentially predictive against those new and sophisticated ones. **Table 3** emphasizes that the surveyed anti-phishing techniques fall short in their prediction susceptibilities against some kinds of phishing features such as: Cross Site Scripting (XSS), Embedded Objects and cross language exploits in novel variants of phishes [3], [9], [14], and [16]-[31]. In the presence of such causality between new phishing exploits and the prediction susceptibilities of current anti-phishing techniques, a key challenge on the optimal thwarting of new variants of phishes rises day by day. Thus, further exploration of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

features is emerging as a key research agenda in phishing detection domain. That, in turn, will lead to major prediction improvement with low latency of misclassifications which is the main goal of almost anti-phishing techniques.

In the course of tolerating with a big data like that on the web, anti-phishing techniques assisted by features still have some shortages like complex computations, time consuming and requirements of external resources. In the presence of high dimensional space of features which may contain many non-contributing ones, a costly amounts of misclassifications need for attention and being resolved straightforwardly [42]-[47]. To come up with the vastness of webpages and their enormous variety of features, a minimal and effective subset of the most contributing features must be selected for both dimensionality reduction of feature space and the best prediction susceptibility of phish and non-phish classes [45]-[46]. Therefore, some anti-phishing techniques (**Table 1**), are assisted by traditional methods for feature selection. Such methods typically filter out the original set of extracted features into minimal subsets of most predictive ones. However, they often deployed sub-optimal feature subsets for phishing prediction due to some constraints as described in **Table 4**. Constraints like the dimensionality of the feature space, the type of features, the heterogeneity of their values, and the existence of irrelevant and redundant features; limit the assisted features selection methods and then degrade the overall performance of anti-phishing techniques [23]-[28]. More precisely, the shortages of assisted feature selection methods become more challenging day by day in realistic application [23]-[28]. Therefore, an alternative and remedial methods are required to assess and select a set of the most relevant and least redundant features. In turn, they may demonstrate the discriminating power of anti-phishing techniques on phishes among a given stream of the web with least misclassification costs.

4. METHODOLOGY

4.1 New Features

Based on the aforesaid observations, 58 hybrid features that expectedly being crafted by phishers on their fake webpages are nominated for investigation in this work. Additionally, their prediction susceptibility against phishing is experimentally assessed. These examined features are specifically extracted from two different sources: webpage's URL and content. Thus, two feature categories are taken into consideration through the experiments, the first feature category is a group of 48 features mostly including cross site scripting and embedded objects features; whereas the second feature category is a group of 10 URL features extracted from webpage's URL. To extract such features, the j^{th} webpage (W_j) is represented as a feature vector by using the standard document representation that is usually used for text classification. Then all feature vectors extracted from m-dimensional training dataset are represented as feature matrix M such that $M = \{W_1 \quad W_2 \quad W_m\}$; where m indicates the number of feature vectors included in M. Each entry feature vector W_j in M consists of its feature indexes and their corresponding values alongside its corresponding class label as the first column as follows [6], [7], [15], [45], and [46]:

$W_j = \{C_j, (f_{j,1}, v_{j,1}), \quad (f_{j,2}, v_{j,2}), \ldots \quad (f_{j,n}, v_{j,n})\}$;

Where n is the number of features and C_j is the label of the class.

4.2 Feature Selection Criterion

Based on the review of previous works assisted by traditional feature selection techniques. Traditionally, assisted feature selection techniques relied on either features ranking or feature subset selection. Features ranking based methods select features according to their discriminating power on instances related to different classes. Feature subset selection based techniques, mainly seek for a minimal and effective feature subsets through specific search strategies [44]. However, such strategies become inefficient against high dimensional feature space because they rarely underscore redundancy problem along with the relevance problem over a large feature space (i.e. high dimensional datasets). Then, they yield a trade-off between results optimality and computational efficiency [44] and [47]-[50].

One of possible ways to identify a nearly optimal feature subset for effective phishing detection, is a specific criterion (*mRMR*) which is presented herewith. *mRMR* [44], [47]-[50] removes irrelevant and redundant features simultaneously over a high dimensional feature space (i.e. high dimensional dataset). And it assesses features' relevance and redundancy constraints independently of any classification algorithm. Thus, it provides minimal subsets of the most predictive with less computation and unexhausted searching strategy over all possible combinations of features [47]-

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[50]. Further, it could be incorporated with other feature selection criteria like filter-based and wrapper based criteria. For this purpose, it has been adopted by the literatures of related fields like pattern recognition, high dimensional data processing and genes expression [50] and [51]. However, it is scarcely adopted in the literature of phishing detection in spite of treating phishing detection as a pattern recognition problem by almost previous works.

The feature's relevance refers to the mutual information between the feature itself and the class label vector to maximize. According to [44] and [50], a feature can be categorized in terms of its relevance into: most and least relevant as well as irrelevant. Most relevant infers that a feature is necessary to choose an optimal feature subset such that it can affect the class distribution whenever it is removed. While, least relevant refers to the necessary feature for choosing an optimal subset under a specific condition. Otherwise, irrelevant feature is not necessary for optimal feature subset selection at all [44] and [50]. On the other hand, the feature's redundancy denotes the mutual information among the feature itself and other features in the same combination [47]-[51]. For instance, two features are said to be redundant if their values are completely correlated to each other's within a set of features [44], [50] and [51]. Given a feature set F, a feature in the feature set (F_i), and a feature subset (S_i) such that$S_i = F - \{F_i\}$, the aforesaid concepts are clearly defined as what follows [44, 50]:

Definition 1 (maximal relevant feature)
A feature F_i is maximally relevant if: $P(C|F_i, S_i) \neq P(C|S_i)$.

Definition 2 (minimal relevant feature)
A feature F_i is minimally relevant if: $P(C|F_i, S_i) \neq P(C|S_i)$
and $\exists S_i' \subset S_i, such\ that\ P(C|F_i, S_i') \neq P(C|S_i')$.

Definition 3 (irrelevant feature)
A feature F_i is said to be irrelevant if: $\forall\ S_i' \subseteq S_i, such\ that\ P(C|F_i, S_i') \neq P(C|S_i')$

Definition 4 (redundant feature)
Given M_i as Markov Blanket for F_i, a feature F_i is redundant if it is weakly relevant and has a Markov Blanket M_i on F such that:
$P(F - M_i - \{F_i\}, C|F_i, M_i) =$
$P(F - M_i - \{F_i\}, C|M_i)$

The criterion of *mRMR,* involves three simultaneous feature assessment on both its maximal relevance and its minimal redundancy as given in *Equations 1, 2* and *3* [51]-[53]. *Equation 1* excludes the features set S that highly depends on the target class c. Whereas, *Equation 2* eliminates features that are highly dependent on each other without compromising their discriminability. Finally, *Equation 3;* $max\Phi(D,R)$, combines both constraints in *Equations 1* and *2* as follows [51]-[53]:

$$max\, D(S,c), D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i, c) \quad (1)$$

$$min\, R(S), R = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i, x_j) \quad (2)$$

$$max\, \Phi(D,R), \Phi = D - R. \quad (3)$$

Equation 1 results the feature set S with m features x_i that have the highest dependency $D(S, c)$ on the target class c. Then, the mean value of all mutually informative features x_i with respect to class c is computed with $D(S, c)$. *Equation 2* calculates the highest dependency $R(S)$ among the resultant features x_i, and x_j by selecting mutually exclusive features. The criterion of *mRMR* is defined by combining D and R simultaneously in $\Phi(D,R)$ as described in *Equation 3* [51]-[53] Resultantly, this criterion filters out a minimal set of selective features that are the maximal relevant and minimal redundant features among the extracted hybrid features.

4.3 Phishness Induction Criterion

Generally speaking, most of classification-based anti-phishing techniques as those were previously discussed in Section 2, are trying to map an input data to an output data using a specific induction function. An established induction rule γ automatically assess the relevance of input feature vector to a specific class, e.g. phish and non-phish. Thus, the induction function maps the extracted feature vector W_j into an output vector Y_j with the aid of an induction rule γ such that $Y_j = f(W_j, \gamma)$ [6] and [11]. This induction function applies to all m feature vectors that obtained from m dimensional training dataset during the learning task to produce the *classification model*. During the testing task, the same features are extracted from un-labelled instance, which is represented as a feature vector W_{new}, and learned with the previously generated *classification model* to produce its corresponding classification label as either $W_{new}^{\sim}$ or $W_{new}^{\sim}$ [6] and [11]. Herewith a certain induction function denoting by *Support Vector Machine* (*SVM*) classifier, the tested feature vector W_{new} of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

input website is classified into either *phish* (W_{new}^{+}) or *non-phish* (W_{new}^{-}). The *SVM* classifier is the most commonly used classifier to obtain the optimal separating hyper plane between two classes [56] and [57]. It guarantees the lowest level of error rate because of its generalization ability and handling of high dimensional feature space by producing two output class labels: +1 and -1, respectively [56] and [57]. Basically the induction by the *SVM* classifier implemented as follows: *W* denotes all the web pages in the training dataset such that $W = \{W_1 \quad W_j \quad W_m\}$ and W_j is the feature vector of each web page as$W_j = \{C_j, w_{j,1} \quad w_{j,i} \quad w_{j,n}\}$, where *m* and n are the number of feature vectors and features in each feature vector, respectively. Then, $w_{j,i}$ denoting the value of i^{th} feature index for each j^{th} feature vector W_j, where$0 \leq w_{j,i} \leq 1$, $i = 1,2,3,\dots,n$ and $j = 1,2,3,\dots,m$. given that $W = \{W_j\}_{j=1}^{m}$ is a set of m training feature vectors or alternatively the *m*-dimensional feature matrix [6], [11], [56], and [57]. Each W_j is labelled by $C_j \in \{1,-1\}$ with $C_j = 1$ and $C_j = -1$ which indicates the membership of W_j in the class 1 and the class 2 through *Equation 4* [56] and [57]:

$$f(x) = \sum_j \alpha_j \gamma_j K(W_{new}, W_j) + b \qquad (4)$$

Where α_i and b are obtained by a quadratic algorithm, W_{new} is the unlabelled website and W_j is the feature vector of each training website. The function $K(W_{new}, W_j)$ maps the space of input webpage to higher dimensions where training webpages in the dataset are learned individually [56] and [57]. Furthermore, features were varied in their values between continuous and categorical values, i.e. binary or numeric values. Such heterogeneous features in their values were optimized as follows: the binary features were computed as the union of their corresponding features, while the numeric features were combined by taking the smallest value of the corresponding features. Given a selective feature subset that is a minimal subset of most relevant and least redundant features; would be used with the aid of the *SVM* classifier over the training dataset to generate the *phish detection model*. *Phish detection model* is a consolidated classification model that could be further deployed as an inducer of an unlabeled webpage's class during the testing task [3], [6], [7], and [10].

Consequently, to overcome the heterogeneity of features values, all feature values are discretized into the interval of values [0…1], and they represented as numeric number. Some feature values were computed as the union of their corresponding features to represent a binary value (i.e. either 1 or 0) that refers to the presence or absence of that feature in the examined webpage. Meanwhile, other features were combined by taking the smallest value of the corresponding features and they represented in values in the specified interval [47] and [48]. To implement the SVM, a machine learning tool from the Waikato Environment for Knowledge Analysis (WEKA) was used.

4.4. Evaluation Criterion

As consequence, the expected level of accuracy that a generated classification model could fit by using extracted features over a dataset, is evaluated too. For this purpose, some of the most commonly used measurements like *Precision*, *Recall* and *F-measure* that enlisted in **Table 5** are utilized. To do so, each instance in a dataset is binary classified into two classes either positive or negative classes (i.e. phish or non-phish) [56] and [57]. Herewith the binary classification, the performance of the generated *phish model* is assessed by the aforesaid set of theoretical measurements. Each of them rely on constraints of *TP, FP* and *FN* [1], [17], [29], [56] and [57]. The True Positive (*TP)* indicates the rate of correctly classified phish instances. The False Positive (*FP)* refers to the rate of wrongly classified legitimate instances as the phishing ones. The False Negative *(FN)* indicates the wrongly labeled phish instances as legitimate ones [1]. The outcomes with maximal *Precision* value state the maximal positive webpages that are classified. However, those of maximal *Recall* value denote the minimal prediction error. Then, the resultant *F-measure* outcome scores denote the initial induction of *phish model* with the aid of selective features [1], [17], [29], [56] and [57].

Regarding, to assess the prediction susceptibility of the extracted features, the *Area Under the Curve* (AUC) was set for the proposed features. This tool is commonly used to represent the efficiency of the classifier under the *Receiver Operating Characteristic (ROC)* curve [1] and [35]. *ROC* is a graph shows the relationship between sensitivity and specificity of a classifier [1], [17], and [35]. In this work, the *AUC* is directly calculated for all features without drawing the *ROC* curve by using calculations presented in **Table 5**.The scalar value

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of *AUC* denotes how much the individual feature could discriminate phish and non-phish classes. Once the *AUC* values of all features are set, the observation of the most contributing feature becomes easier. The higher feature in *AUC* would be the maximal in its contribution to the purpose of phishing detection [35]. Prediction susceptibility evaluation demonstrates whether the extracted features are expected to be crafted by phishers and they are able to predict their evolving deceptions.

On the problem at hand (*i.e. obtaining near optimal feature subset*), it is a noteworthy issue to highlight whether the selective subsets are nearly optimal subsets for phishing classification model. Thus, the outcomes of the simultaneous discarding criterion of redundant and irrelevant features *(mRMR)* are quantified on their goodness, stability and similarity over the collected datasets. Thus, specific measures that adopted by prior researchers in different fields are recommended in this work (**Table 5**) to evaluate the selection outcomes [51]-[53]. Such measures can be considered as comparison baselines for any further study on feature selection effects to phishing detection. The higher outcome features subset in *Goodness, Stability and Similarity* specifics, denotes the near optimal subset which in turn would increase the classification accuracy and speed up the classification task [51]-[53]. To the best of our knowledge, this type of robustness evaluation with the aid of the aforesaid measures is scarcely underscored in the literature of phishing detection despite of its significance for both feature selection and phishing classification model. Evaluation was implemented on collected datasets that set to extract 58 hybrid features. Then, a comparison was implemented on the robustness of the best chosen feature subset across the outputs of the aforesaid evaluation and classification models. More details on dataset collection, feature selection, and evaluations of robustness and effectiveness would be summarized in next section.

5. EXPERIMENTS

Our experimental strategy involves collecting the preliminary datasets (i.e. webpages aggregated from legitimate and phishing data archives). Further, it involves four steps conducted for implementation and assessment: features extraction, prediction susceptibility assessment, assessment of robustness, assessment of effectiveness. The first step focuses on extracting the original set of features. Then, the second step demonstrates which feature category is the most contributing to phishing classification. Whereas; the rest two steps underscore the robustness of feature selection outputs, and highlight their significance to the classification task. This experimental strategy was conducted to emphasize our study's objective and to promote its contribution.

5.1. Experimental Setup

A preliminary set of real world webpages, 500 living phishing webpages and 500 valid legitimate webpages were downloaded in 30 days from September to November 2014. Specifically, the phishing pages were downloaded from two publically available sources; the *Phish Tank* and the *Castle Cops* archives. The *Alexa's top sites* archive was used as the source of legitimate webpages. The collected webpages hosted by websites of the most targeting financial organizations by phishers. They involve: homepage, registration forms and login functionalities.

5.2. Features Extraction

This step was conducted to extract the original set of 58 hybrid features from their relative parts on webpages. Additionally, it states what features and what feature categories that have mostly been exploited by phishers as advanced deceptions on their evolving phish variants. Thus, two categories of features were extracted and grouped as two feature groups. Webpage content group consists of 48 features extracted from the source code and HTML tags of collected webpages. URL features group composed of 10 features extracted from URL indicators. It is noteworthy to mention that the group of hybrid features presented herewith, is constructed by combining all the features belonging to the former feature groups, as enlisted in **Table 6**.

5.3. Assessment of Prediction Susceptibility

In this step, the extracted features were examined on their prediction susceptibilities in order to identify (i) the most predictive features and (ii) the most contributing feature category that could maximize phishing classification accuracy and minimize classification sensitivity. To do so, the SVM classifier was run three times over the training dataset with the use of three feature groups (i.e. three different categories of features: webpage content features, URL features, and hybrid features) accordingly. As plotted in Figures 1, 2,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and 3; the reported results of learning SVM classifier with all features groups show the best performed feature group in filtering phish instances among its competitors. The concerns of specificity, sensitivity and misclassification indicate the effects of such features and their overlap reduction to distinguish phish variants. Thus, the proposed hybrid features group could possibly be assigned to predict multiple phish variants because of their hybridity, i.e. their variety of potentials. In turn, this will help in circumventing the phishers' attempts to bypass the existing anti-phishing techniques. Additionally, features encompassed in the group of hybrid features are evaluated and ranked with respect to their computed *AUC* scores. The higher feature in *AUC* is ranked as the higher on its prediction susceptibility among the others. Unlike further assessment steps that highlight features contributions as different compactness, this assessment examines the prediction susceptibility of features individually.

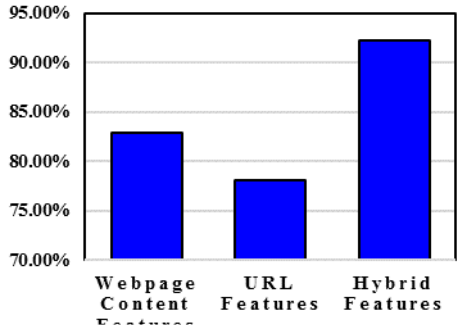


*Figure 1: Percentages of **TP** in terms of the category of features space.*

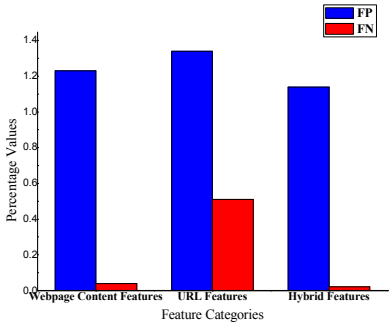


*Figure 2: Prediction in terms of the category of features, and percentages of both **FP** and **FN**.*

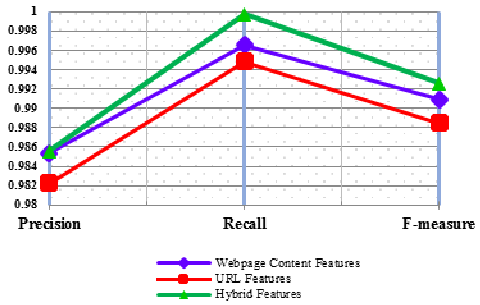


*Figure 3: Classifier's performance merits in terms of features categories, **Precision**, **Recall** and **F-measure**.*

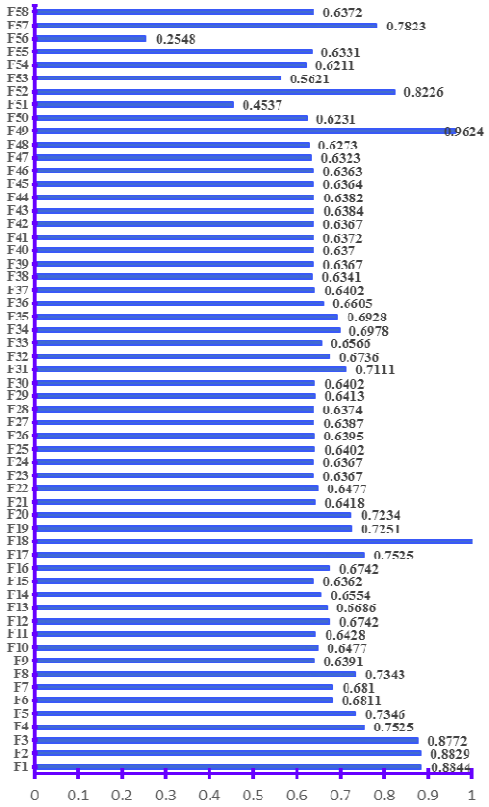


*Figure 4: **AUC** scores show the individual features' prediction susceptibilities.*

5.4. Assessment of Robustness

To further assess the predictive features as compact subsets, they were selected as different subsets with respect to their relevance and redundancy concerns by using *mRMR* criterion.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Resultantly, five subsets of most relevant and least redundant features are selected and considered as candidate subsets for classification task. The robustness of the five selective feature subsets was assessed with the aid of three promoting measures: Goodness, Stability and Similarity. Such assessment provides a general view on the robustness that each selective feature subset could state among its competitors. **Table 7** enlisted the selective subsets with their constituent features in terms of their Goodness, Stability and Similarity scores. It can be observed from **Table 7**, that the selective feature subsets *3, 4* and *5* reported the best scores than their competitors.

5.5. Assessment of Effectiveness

In this context, **Figure 5** provides a comparative view of the classifier's performance by using the aforesaid selective subsets of features.

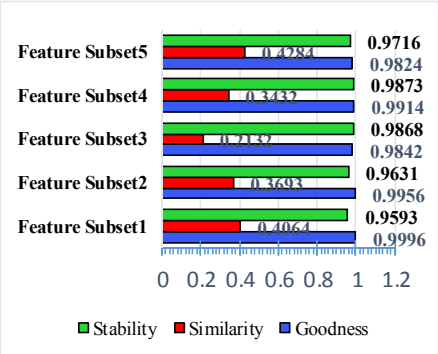
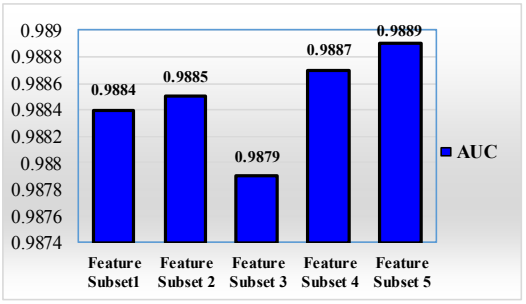


*Figure 5: Robustness of feature selection outcomes in terms of **Goodness**, **Stability** and **Similarity** scores.*

Afterwards, their contributions to improve classification performance was assessed by learning them with the aid of SVM classifier over the training dataset in five runs. Further, their performance outcomes were underscored in the form of their *AUC* scores as can be seen in **Figure 6**. This assessment emphasizes the outperformed *classification model.* That, in turn, states the significance of the suggested feature subsets to enhance phishing detection approach in terms of classification accuracy and sensitivity in the future.

*Figure 6: Overall classification performance using the selective subsets of features w.r.t their **AUC** scores.*



6. DISCUSSION

As can be observed from **Figures 1, 2 , 3, 4, 5,** and **6** as well as **Tables 6** and **7**; the overall reported results are encouraging towards deploying the presented features as hybrid and predictive features and as selective subsets for phish website detection. The only differences are their variation of prediction susceptibility, the robustness and effectiveness of their compactness. In summary, the findings infer the following:

- Reported results in **Figures 1, 2,** and **3** restates the significance of features hybridity to detect phish variants. Phishers usually exploit different types of features in their own phish webpages. Moreover, a phisher may exploit multiple variant phish webpages for the same legitimate webpage. Therefore, deployments of different features (i.e. hybrid) leads to a quite bit high scalable detection against phish webpages in classification based detection approaches and then less amounts of misclassification cost. Moreover, both features' hybridity and their values' heterogeneity can be considered as the baselines for well performed classification models. As such, hybrid features promote the overall performance of classification models with low latency.
- It can be observed from **Figure 4**, Therefore, In **Figure 4**, the overall resultant *AUC* scores are very encouraging towards deploying all the proposed hybrid features as predictive features on evolving phish webpages. However, the variation of their prediction susceptibilities infers that amongst 58 hybrid features, there are 20 hybrid features can be nominated as the most predictive ones for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

phishing detection. Upon their statistics plotted in **Figure 4**, it is revealed that features *F1, F2, F3, F18, F49* and *F52* outperform the others. They scored the highest values of *AUC:* 0.8844, 0.8829, 0.8772, 1.0000, 0.9624 and 0.8226, respectively. However, features like *F9, F11, F15, F51, F53* and *F56* less predictive than others due to their lowest *AUC* values: 0.6391, 0.6428, 0.6362, 0.4537, 0.5621 and 0.2548, respectively.

- **Figure 5** qualifies the robustness of the selected subsets in terms of goodness, stability and similarity scores over the collected dataset. It clearly shows that the 3rd, 4th and 5th selective subsets preserve the best cases of goodness (i.e. quality) among the others. This implies the significance of reducing feature set's dimensionality, removing both redundant and noisy features to define the best features subset, and the ability to obtain multiple robust feature subsets from high dimensional feature space. Indeed, such feature subset will improve the effectiveness of phishing classification model. More interestingly, such feature subsets are needed to effectively detect phishing websites in realistic applications. Unlikeliness, the 1st and 2nd feature subsets yield least goodness cases among their competitors.
- **Figure 5** outlines how the obtained 3rd, 4th, and 5th feature subsets are notably more stable over the tested datasets than their competitors. That, in turn, emphasizes the significance of their inter-dependencies between the features included in the same chosen feature subset. Features chosen on their inter-dependencies can compose a stable subset under different classification scenarios and datasets. In contrary, 1st and 2nd subsets vary in their stability even though they yield high goodness scores.
- In the context of overall outputs' similarity (**Figure 5**), it can be observed that the five outputs are relatively dissimilar over all the datasets. However, 3rd, 4th and 5th subsets are notably dissimilar (i.e. their reported similarity scores are less than 1.0) which point out that their constituent features partially overlapped. More interestingly, such dissimilarity implies that these feature subsets are complementary to each other's. They are diversely predictive and contributing a promising avenue to improve the classification performance. In particular, this dissimilarity reveals that the applied feature selection criterion (mRMR) can be effectively exploited and integrated as an assisted feature selection technique to any phishing detection approach. Despite this, both likelihood and variations between selective subsets of features are crucial issue in a machine learning based detection approaches.
- For the problem domain at hands (i.e. phish webpage detection); all the outcomes reported in **Table 7** and portrayed in **Figure 6** pay attention to the crucial importance of discarding feature's irrelevance and redundancy at once. The *mRMR* criterion enables us to improve the detection performance in the context of using as few, most relevant and least redundant features as possible. Interestingly, feature subsets chosen by *mRMR* have the best cases of robustness. Further, they reported best case of effectiveness whenever they applied on SVM classifiers over the collected training and testing datasets.
- Based on the overall findings, we can infer that (*mRMR*) could promise near optimum subsets of features (i.e. a minimal, robust and effective feature subsets) through more intensive experimentations and over more challenging datasets. Moreover, both hybrid and most predictive features that presented in this study are promoting to detect phishing specifically in real world application. However, the exact answer for optimum subsets of features cannot be provided unless further assessments conducted in terms of classification accuracy, specificity and sensitivity across different classification models. This case of study will be addressed in our future work.

7. CONCLUSION AND FUTURE WORK

With regard to the problem of effectively classifying phish exploits over more challenging data like the web; identifying and deploying optimal feature sets still rises as a key challenge. Motivated by this problem at hands, this study attempts to introduce as minimal and effective subsets of hybrid features as possible. It deployed a large set of hybrid features, it chose minimal subsets from them, and quantified their robustness and effectiveness over collected datasets. As experimentally demonstrated, the presented hybrid features varied from highly to low predictive features due to their prediction potentials and inter-dependencies. And they revealed several

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

compact combinations (i.e. selective subsets) by including the relevant features and excluding the redundant ones simultaneously using *mRMR* criterion. Additionally, the selective subsets were investigated to gain a deeper understanding of their qualities in terms of goodness, stability and similarity. Finally, they were applied on a machine learning classifier to adjust whether they were promoting enough to construct an effective classification model or not. The outcomes emphasized that *mRMR* can be handled as an adequate feature selection criterion to assist phishing detection approach. Further, the joint usage of *mRMR* and specific evaluation criteria gave a useful insight on how to provide robust and effective subsets of features. Such robust and effective feature subsets are strongly needed to construct an adaptive phishing classification model against more challenging datasets. Amongst selective subsets, several subsets peaked the best cases of goodness, stability and similarity. Thus, they can be considered as near optimal feature subsets to provide an increased robustness and effectiveness with less amounts of errors and misclassifications.

Besides, the reported findings are encouraging and promising to enhance phishing detection in terms of computational costs and performance. They restated that the joint use of hybrid features, feature selection criterion, and robustness evaluation measures are supportive to further optimization of realistic phishing detection. Moreover, we intend to develop this work in the future towards finding optimum solution to the problem at hands. Our future work aims to quantify the effects of different feature selection mechanisms along with *mRMR* criterion. And it aims to compare their outcomes across different classification models and datasets through more intensive experimentations. Thus, their contributions can be well optimized for better classification performance.

REFRENCES:

[1] M. Khonji, Y. Iraqi, and A. Jones, Phishing detection: a literature survey. Communications Surveys & Tutorials, IEEE , (15), 2013., 2091-2121.

[2] S. Purkait, Phishing counter measures and their effectiveness–literature review. Information Management & Computer Security, (20), 2012, 382-420.

[3] G. Ramesh, I. Krishnamurthi and K. Kumar, An efficacious method for detecting phishing webpages through target domain identification, Decision Support Systems, (61), 2014, 12-22.

[4] A. Almomani, B. B. Gupta, S. Atawneh, A., Meulenberg, and E. Almomani, A survey of phishing email filtering techniques, Communications Surveys & Tutorials, IEEE. 15(4), 2013, 2070-2090.

[5] V. Shreeram, M. Suban, P. Shanthi and K. Manjula, Anti-phishing detection of phishing attacks using genetic algorithm, In Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference, 2010, 447-450.

[6] M. He, S.-J. Horng, P. Fan, M. K. Khan, R.-S. Run, and J.-L. Lai, An efficient phishing webpage detector, Expert Systems with Applications, (38), 2011, 12018-12027.

[7] Y. Li, R. Xiao, J. Feng and L. Zhao, A semi-supervised learning approach for detection of phishing webpages, Optik-International Journal for Light and Electron Optics, (124), 2013, 6027-6033.

[8] W. Han, Y. Cao, E. Bertino and J. Yong, Using automated individual white-list to protect web digital identities, Expert Systems with Applications, (39), 2012, 11861-11869.

[9] H. Shahriar, and M. Zulkernine, Trustworthiness testing of phishing websites: A behavior model-based approach, Future Generation Computer Systems, (28), 2012, 1258-1271.

[10] R. Islam and J. Abawajy, A multi-tier phishing detection and filtering approach, Journal of Network and Computer Applications., (36), 2013, 324-335.

[11] P. Barraclough, M. Hossain, M. Tahir, G. Sexton and N. Aslam, Intelligent phishing detection and protection scheme for online transactions, Expert Systems with Applications, (40), . 2013, 4697-4706.

[12] M. Bhati and R. Khan, Prevention Approach of Phishing on Different Websites, International Journal of Engineering and Technology, (2), 2012.

[13] M. S. Sadi, M.M.R. Khan, M. M. Islam, S. B. Srijon and M. M. H. Mia, Towards detecting phishing web contents for secure internet surfing, 2012 International Conference on Informatics, Electronics & Vision (ICIEV), 2012, 237-241.

[14] S. Gastellier-Prevost, G. G. Granadillo and M. Laurent, Decisive heuristics to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

differentiate legitimate from phishing sites, 2011 Conference on Network and Information Systems Security (SAR-SSI), 2011, 1-9.

[15] H. Wang, B. Zhu and C. Wang, A Method of Detecting Phishing Web Pages Based on Feature Vectors Matching, Journal of Information and Computational Systems. (9), 2012, 4229-4235.

[16] P. Likarish, E. Jung, D. Dunbar, T.E. Hansen and J. P. Hourcade, B-apt: Bayesian anti-phishing toolbar, IEEE International Conference on Communications, ICC'08. 2008, 1745-1749.

[17] C. Whittaker, B. Ryner and M. Nazif, Large-Scale Automatic Classification of Phishing Pages, NDSS, 2010.

[18] A. Bergholz, J. De Beer, S. Glahn, M.-F. Moens, G. Paaß, G. and S. Strobel, New filtering approaches for phishing email, Journal of computer security, 18(1), 2010, 7-35.

[19] G. Xiang, J. Hong, C. P. Rose and L. Cranor, CANTINA+: a feature-rich machine learning framework for detecting phishing web sites, ACM Transactions on Information and System Security (TISSEC), (14), 2011, 21

[20] H. Zhang, G. Liu, T. W. Chow and W. Liu, Textual and visual content-based anti-phishing: a Bayesian approach, IEEE Transactions on Neural Networks, 22(10), 2011, 1532-1546.

[21] H. Kordestani and M. Shajari, An entice resistant automatic phishing detection, 2013 5th Conference on Information and Knowledge Technology (IKT), 2013, 134-139.

[22] G. Ramesh and I. Krishnamurthi, A comprehensive and efficacious architecture for detecting phishing webpages, Computers & Security, (40), 2014, 23-37.

[23] Y. Pan, and X. Ding, Anomaly based web phishing page detection, In 22nd Annual 2006 Computer Security Applications Conference, 2006 (ACSAC'06) IEEE.

[24] L. Ma, B. Ofoghi, P. Watters, and S. Brown, Detecting phishing emails using hybrid features, Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing (UIC-ATC'09), IEEE, 2009. 493-497.

[25] F. Toolan, and J. Carthy, Feature selection for Spam and Phishing detection, In eCrime Researchers Summit (eCrime), IEEE, 2010. 1-12.

[26] M. Khonji, A. Jones, and Y. Iraqi, A study of feature subset evaluators and feature subset searching methods for phishing classification, In Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, ACM, 2011, 135-144.

[27] R. B. Basnet, A. H. Sung, and Q. Liu, Feature selection for improved phishing detection, In Advanced Research in Applied Artificial Intelligence. Springer Berlin Heidelberg, 2012, 252-261.

[28] D. Zhang, Z. Yan, H. Jiang and T. Kim, A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites, Information & Management, 51(7), 2014, 845-853.

[29] M. Aburrous, M. Hossain, K. Dahal and F. Thabtah, Associative classification techniques for predicting e-Banking phishing websites, International Conference on Multimedia Computing and Information Technology (MCIT), 2010.

[30] W. Zhuang, Q. Jiang, and T. Xiong, An intelligent anti-phishing strategy model for phishing website detection, 32nd International Conference on Distributed Computing Systems Workshops (ICDCSW), IEEE, 2012, 51-56.

[31] I. R. A. Hamid, and J. H. Abawajy, An approach for profiling phishing activities. Computers & Security, (45), 2014, 27-41.

[32] M. Khonji, Y. Iraqi, and A. Jones, Lexical URL analysis for discriminating phishing and legitimate websites, Anti-Abuse and Spam Conference Proceedings of the 8th Annual Collaboration, Electronic Messaging, ACM, 2011.

[33] J. Zhang, and Y. Wang, A real-time automatic detection of phishing URLs, 2nd International Conference on Computer Science and Network Technology (ICCSNT), 2012, 1212-1216.

[34] H. Huang, L. Qian, and Y. Wang, A SVM-based technique to detect phishing URLs, Information Technology Journal, (11), 2012, 921-925.

[35] C. K. Olivo, A.O. Santin, and L.S. Oliveira, Obtaining the threat model for e-mail phishing, Applied Soft Computing, 2011.

[36] M. Rajalingam, S. A. Alomari, and P. Sumari, Prevention of Phishing Attacks Based on Discriminative Key Point Features of WebPages, International Journal of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Computer Science and Security (IJCSS), (6), 2012.
[37] M. G. Alkhozae, Phishing websites detection based on phishing characteristics in the webpage source code, International Journal of Information and Communication Technology Research, 2011.
[38] Y. Liu, and M. Zhang, Financial websites oriented heuristic anti-phishing research., 2012 IEEE 2nd International Conference on Cloud Computing and Intelligent Systems (CCIS), 2012, 614-618.
[39] K. Sanka, and B. A. Suresh, New Framework for Thwarting Phishing attacks based on Visual Cryptography.
[40] E. H. Chang, K. L. Chiew, S. N. Sze, and W. K. Tiong, Phishing Detection via Identification of Website Identity, 2013 International Conference on IT Convergence and Security (ICITCS), 2013, 1-4.
[41] E. Uzun, H. V. Agun, and T. Yerlikaya, A hybrid approach for extracting informative content from web pages. Information Processing & Management, (49), 2013, 928-944.
[42] L. Fu, Y. Meng, Y. Xia, and H. Yu, Web content extraction based on webpage layout analysis, 2010 Second International Conference on Information Technology and Computer Science (ITCS), 2010, 40-43.
[43] R. B. Basnet, and A. H. Sung, Mining Web to Detect Phishing URLs, 11th International Conference on Machine Learning and Applications (ICMLA), 2012, 568-573.
[44] L. Yu, and H. Liu, Efficient feature selection via analysis of relevance and redundancy, The Journal of Machine Learning Research, 5, 2004, 1205-1224.
[45] I. R. A. Hamid, and J. Abawajy, Hybrid feature selection for phishing email detection, Algorithms and Architectures for Parallel Processing, 2011, 266-275.
[46] R. Mohammad, F. Thabtah, and L. McCluskey, An assessment of features related to phishing websites using an automated technique, 2012 International Conference for Internet Technology And Secured Transactions, 2012, 492-497.
[47] S. Lee, Y. T. Park, and B. J. Auriol, A novel feature selection method based on normalized mutual information, Applied Intelligence, 37(1), 2012, 100-120.
[48] C. M. Chen, H. M. Lee, and Y. J. Chang, Two novel feature selection approaches for web page classification, Expert systems with Applications, 36(1), 2009, 260-272.
[49] H. Peng, F. Long, and C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence. 27(8), 2009, 1226-1238.
[50] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, S., A. Anand, and H. Liu, Advancing feature selection research., ASU feature selection repository, 2010
[51] A. Fahad, Z. Tari, I. Khalil, I. Habib, and H. Alnuweiri, Toward an efficient and scalable feature selection approach for internet traffic classification, Computer Networks, 57(9), 2013, 2040-2057.
[52] Z. He, and W. Yu, Stable feature selection for biomarker discovery, Computational Biology and Chemistry, 34(4), 2010, 215-225.
[53] L. C. Molina, L. Belanche, and A. Nebot, Feature selection algorithms: A survey and experimental evaluation, IEEE International Conference on Data Mining (ICDM), 2002, 306-313
[54] Y. Chen, Y. Li, X. Q. Cheng, and L. Guo, Survey and taxonomy of feature selection algorithms in intrusion detection system, In Information Security and Cryptology, Springer Berlin Heidelberg, 2006, 153-167.
[55] V. S. Lakshmi, and M. Vijaya, Efficient prediction of phishing websites using supervised learning algorithms, Procedia Engineering, (30), 2012, 798-805.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Survey Of Prior Works With Their Relative Merits

Related Citation	Classifiers	Classifier Design	Feature Selection	Detected Classes
Pan & Ding, 2006 [23]	SVM	Single	χ^2	Phish/legitimate Website
Likarish et al., 2008 [16]	NB	Single	N/A	Phish/legitimate Website
Ma et al., 2009 [24]	SVM, RF	Single	IG	Phish/legitimate Emails
Aburrous et al. , 2010 [29]	K-NN, SVM	Ensemble	N/A	Phish/legitimate Emails
Bergholz et al., 2010 [18]	SVM	Single	N/A	Phish/legitimate Emails
Toolan & Carthy, 2010 [25]	C4.5	Single	IG	Phish/Spam/legitimate Emails
Whittaker, Ryner & Nazif, 2010 [17]	LR, NB	Single	N/A	Phish/legitimate Website
He et al., 2011 [6]	SVM	Single	N/A	Phish/legitimate Website
Khonji, Jones & Iraqi, 2011 [26]	RF	Single	IG, CFS, WFS	Phish/legitimate Emails
Xiang et al., 2011 [19]	SVM, LR, NB	Single	N/A	Phish/legitimate Website
Zhang et al., 2011 [20]	NB	Single	N/A	Phish/legitimate Website
Basnet, Sung & Liu, 2012 [27]	LR, RF	Single	CFS, WFS	Phish/legitimate Website
Huang, Qian & Wang, 2012 [34]	SVM	Single	N/A	Phish/legitimate Website
Zhuang, Jiang & Xiong, 2012 [30]	SVM	Ensemble	Max Relevance	Phish/legitimate Website
Islam & Abawajy, 2013 [10]	AdaBoost, SVM, NB	Ensemble	N/A	Phish/legitimate Emails
Kordestani & Shajari, 2013 [21]	SVM, RF, NB	Single	N/A	Phish/legitimate Website
Li et al., 2012 [7]	SVM, TSVM	Single	N/A	Phish/legitimate Website
Gotham & Krishnamurthi, 2014 [22]	SVM	Single	N/A	Phish/legitimate Website
Hamid and Abawajy, 2014 [31]	AdaBoost, SMO	Ensemble	IG	Phish/legitimate Emails
Zhang et al., 2014 [28]	SMO, LR, NB	Single	χ^2	Phish/legitimate Website

Table 2: Comparison Of Different Feature Categories

Feature Category	Advantage	Disadvantage
Webpage Content features [20], [36]-[41], [44]	Comprehensiveness & widely usage	Challenge of obfuscation, code coverage, malicious code injection and delivery. The symptom of phishing susceptibility by loading references to fake media, libraries, actions, cookies and hyperlinks.
URL features [8], [23], [27], [32], [33]	Easy extraction & widely considered in the literature.	Totally can be controlled and modified by phishers to easily misguide users' attention and lure them harder. With their usage, a challenge of phish detection with high sensitivity is encountered.
Online features [3],[9], [11]	Easy extraction	Limited usage and requirement of external resources.
Hybrid features [5], [6], [12]-[15], [19], [21], [22], [28]-[30]	High comprehensiveness and not easy to mislead them totally by the phishers.	Complex extraction process and time overhead.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3: Prediction Susceptibility Of The Prior Works Against The Most Striking Features Crafted By Phishers

Related Citation	Phishing Features		
	XSS Features	Embedded Objects Features	Language Independent Features
Pan & Ding, 2006 [23]	No	No	Yes
Likarish et al., 2008 [16]	No	No	Yes
Ma et al., 2009 [24]	Yes	No	Yes
Aburrous et al. , 2010 [29]	Yes	No	Yes
Bergholz et al., 2010 [18]	No	No	Yes
Toolan & Carthy, 2010 [25]	Yes	No	Yes
Whittaker, Ryner & Nazif, 2010 [17]	No	No	No
He et al., 2011 [6]	No	Yes	Yes
Khonji, Jones & Iraqi, 2011 [26]	No	No	Yes
Xiang et al., 2011 [19]	Yes	No	No
Zhang et al., 2011 [20]	No	No	Yes
Basnet, Sung & Liu, 2012 [27]	No	No	Yes
Huang, Qian & Wang, 2012 [34]	No	No	Yes
Zhuang, Jiang & Xiong, 2012 [30]	No	No	Yes
Islam & Abawajy, 2013 [10]	Yes	No	Yes
Kordestani & Shajari, 2013 [21]	No	No	Yes
Li et al., 2012 [7]	No	Yes	No
Gotham & Krishnamurthi, 2014 [22]	Yes	No	Yes
Hamid and Abawajy, 2014 [31]	No	Yes	Yes
Zhang et al., 2014 [28]	No	Yes	No

Table 4: Prior Works In Terms Of Their Limited Feature Selection Methods

Related Citations	Feature Selection Method (S)	Limitations
Ma et al., 2009 [24]	IG	▪ Heterogeneous features in their values ▪ Low dimensional feature space (7 features) ▪ Redundancy problem
Toolan et al., 2010 [25]	IG	▪ High dimensional feature space (40 features) ▪ High computational cost
Whittaker et al., 2010 [17]	TF-IDF	▪ Noisy & redundant features ▪ High computational time and cost
Basnet et al., 2011 [27]	CFS, WFS	▪ High computational time and cost ▪ High dimensional hybrid feature space (177 features) ▪ Redundant and irrelevant features
Khonji et al., 2011[26]	IG, WFS, CFS	▪ High dimensional hybrid feature space (47 features) ▪ Relevance and redundancy problems
Zhuang et.al., 2012 [20]	Max Relevance	▪ Complex computation ▪ Problem of redundant features
Hamid and Abwajy, 2014 [31]	IG	▪ Heterogeneous values of features ▪ Time and resources consumption ▪ Relevance and redundancy problems

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5: Evaluation Metrics

Metrics	Definition	Calculation
Precision [1], [4]	The percentage of correct positive predictions	$\frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FP\rvert}$ (1)
Recall [1], [4]	Recall Sensitivity refers to the percentage of positively predicted positive instances (TPs).	$\frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FN\rvert}$ (2)
F-score [1], [4]	It refers to the test's accuracy score.	$\frac{2\cdot P\cdot R}{P+R}$ (3)
ROC [1], [4],[35]	TP plotted against FP.	*ROC curve*
AUC [35]	It denotes the weight of features' prediction susceptibilities upon the classifier's efficiency plotted by ROC curve.	$AUC = \frac{1}{MN}\sum_{j=1}^{M}(S_j - j)$ (4) Where S_j is the rank of each j^{th} feature in each group and $S_j - J$ is the number of positive features before the negative features in weight.
Goodness [51]-[55]	It measures the classification accuracy of the selective feature subset upon extremely imbalanced datasets.	$Goodness(S_i) = \frac{1}{Y}\sum_{i=1}^{Y}\frac{N_i^{tp}}{N_i}$ (5) Where Y, N_i^{tp} and N_i are the number of classes in the dataset, the number of true positive of each class and the total number of instances for class i respectively
Stability [51]-[55]	It quantifiably approves the stability of the selective subsets of features against varied datasets over a period of time in real-world application.	$Stab(S) = \sum_{f_i \in X}\frac{F_{f_i}}{N}\times\frac{F_{f_i}-1}{\lvert D\rvert-1}$ (6) Where $f_i \in X$ and $\frac{F_{f_i}}{N}$ are all features in a collection dataset S and the relative frequency of each feature in a subset. If all subsets are identical then *Stab(S)* is close to 1; otherwise is close to 0.
Similarity [51]-[55]	It compares the behavior of multiple feature subsets on the same dataset.	$Sim(t_1,t_2) = 1 - \frac{1}{z}\sum\left\lvert\frac{F_{f_i}^{t_1}}{N^{t_1}} - \frac{F_{f_i}^{t_2}}{N^{t_2}}\right\rvert$ (7) Where $F_{f_i}^{t_1}$ and $F_{f_i}^{t_2}$ denoting the number of frequencies of feature f_i in two candidate feature subsets t_1 and t_2 respectively. Similarity takes values within $[0,1]$.

Table 6: Proposed Hybrid Features With Their Values

Webpage Content Features Category					
Index	**Feature**	**Value**	**Index**	**Feature**	**Value**
F1	Number of Scripting.FileSystemObject	{0~1}	F24	Number <input> in java scripts	{0~1}
F2	Number of Excel.Application	{0~1}	F25	JavaScript scripts length	{0~1}
F3	Presence of WScript.shell	{0, 1}	F26	Number of functions' calls in java scripts	{0~1}
F4	Presence of Adodb.Stream	{0, 1}	F27	Number of script lines in java scripts	{0~1}
F5	Presence of Microsoft.XMLDOM	{0, 1}	F28	Script line length in java scripts	{0~1}
F6	Number of <embed>	{0~1}	F29	Existence of long variable names in java scripts	{0, 1}
F7	Number of <applet>	{0~1}	F30	Existence of long function names in java scripts	{0, 1}
F8	Number of Word.Application	{0~1}	F31	Number of fromCharCode()	{0~1}
F9	link length in <embed>	{0~1}	F32	Number attachEvent()	{0~1}

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

F10	Number of <iframe>	{0~1}	F33	Number of eval()	{0~1}
F11	Number of <frame>	{0~1}	F34	Number of escap()	{0~1}
F12	Out-of-place tags	{0, 1}	F35	Number of dispacthEvent()	{0~1}
F13	Number of <form>	{0~1}	F36	Number of SetTimeout()	{0~1}
F14	Number <input>	{0~1}	F37	Number of exec()	{0~1}
F15	Number of MSXML2.XMLHTTP	{0~1}	F38	Number of pop()	{0~1}
F16	Frequent <head>, <title>, <body>	{0, 1}	F39	Number of replaceNode()	{0~1}
F17	<meta index.php?Sp1=>	{0, 1}	F40	Number of onerror()	{0~1}
F18	"Codebase" attribute in <object>	{0, 1}	F41	Number of onload()	{0~1}
F19	"Codebase" attribute in <applet>	{0, 1}	F42	Number of onunload()	{0~1}
F20	"href" attribute of <link>	{0, 1}	F43	Number of <script>	{0~1}
F21	Number of void links in <form>	{0~1}	F44	frequent<div onClick=window.open()">	{0, 1}
F22	Number of out links	{0~1}	F47	Number of onerror()in javascripts	{0~1}
F23	Number of <form> in java scripts	{0~1}	F48	Number of SetInterval()	{0~1}
Url Features Category					
F49	Multiple TLD	{0, 1}	F54	Typos in Base name	{0, 1}
F50	Brandname in hostname	{0, 1}	F55	Long domain name	{0, 1}
F51	Special symbols in URL	{0, 1}	F56	Misleading subdomain	{0, 1}
F52	Coded URL	{0, 1}	F57	Number of dots in URL	{0~1}
F53	IP address instead of domain name	{0, 1}	F58	Path domain length	{0~1}

Table 7: Performance Analysis Of Selective Subsets Of Features On Training Dataset

Features Subsets	**Features**	**Goodness**	**Similarity**	**Stability**	**AUC**
1	{F1, F3, F49, F16, F3, F2, F20, F36, F57, F52}	0.9996	0.4064	0.9593	0.9884
2	{F49, F2, F57, F1, F52, F32, F21, F29, F30, F3}	0.9956	0.3693	0.9631	0.9885
3	{F1, F49, F2, F32, F30, F17, F52, F57, F21, F15}	0.9842	0.2132	0.9868	0.9879
4	{F3, F49, F15, F3, F20, F23, F57, F30, F2, F1}	0.9914	0.3432	0.9873	0.9887
5	{F8, F36, F15, F1, F2, F20, F23, F57, F30, F52}	0.9824	0.4284	0.9716	0.9889