The Effect of Feature Selection on Phish Website Detection

An Empirical Study on Robust Feature Subset Selection for Effective Classification

Hiba Zuhair[a]
Dept. of Computer Science
Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM, Johor Bahru, Johor, Malaysia;
Al-Nahrain University, Baghdad, Iraq

Ali Selmat
UTM-IRDA Center of Excellence
Universiti Teknologi Malaysia and
Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM, Johor Bahru,
Johor, Malaysia

Mazleena Salleh
Dept. of Computer Science
Faculty of Computing, Universiti Teknologi Malaysia,
81310 UTM, Johor Bahru, Johor, Malaysia

***Abstract*—Recently, limited anti-phishing campaigns have given phishers more possibilities to bypass through their advanced deceptions. Moreover, failure to devise appropriate classification techniques to effectively identify these deceptions has degraded the detection of phishing websites. Consequently, exploiting as new; few; predictive; and effective features as possible has emerged as a key challenge to keep the detection resilient. Thus, some prior works had been carried out to investigate and apply certain selected methods to develop their own classification techniques. However, no study had generally agreed on which feature selection method that could be employed as the best assistant to enhance the classification performance. Hence, this study empirically examined these methods and their effects on classification performance. Furthermore, it recommends some promoting criteria to assess their outcomes and offers contribution on the problem at hand. Hybrid features, low and high dimensional datasets, different feature selection methods, and classification models were examined in this study. As a result, the findings displayed notably improved detection precision with low latency, as well as noteworthy gains in robustness and prediction susceptibilities. Although selecting an ideal feature subset was a challenging task, the findings retrieved from this study had provided the most advantageous feature subset as possible for robust selection and effective classification in the phishing detection domain.**

Keywords*—phish website; phishing detection; feature selection; classification model

I. INTRODUCTION

Phishers impersonate trustworthy websites of financial organizations through online transactions. Many efforts have been made to overcome the phishing attacks through numerous phishing detecting approaches. Nevertheless, phishing has caused enormous money loss in the cyberspace over the past years, which has motivated researchers to seek effective phishing detection techniques that protect users' digital identity [1-3]. In general, phishing detection techniques fall into several categories due to the deployed scenarios of detection. In the literature, Islam & Abawajy [4] roughly categorized them into non-classification and classification techniques. Specifically, white lists of famous trustworthy URLs; black lists of valid phish URLs; heuristics; and information flow techniques were categorized as non-classification techniques. In contrary, classification techniques involved those relied on machine learning classifiers and data mining based scenarios. They differ in terms of classification accuracies, rates of classification errors, and demands on external resources [1-5]. However, they commonly have deployed features as the key factor for classification task, such as hybrid features. Besides, classification task mostly rely on extracting a set of features from tested instances (i.e. emails and websites) and deploy them to distinguish phish instances from the legitimate ones [1-5]. Thus, classification techniques outperformed their competitors by intuitively detecting phishing that exploits the web to protect clients [3, 6]. Moreover, they could automatically extract features from webpage content; URL of websites, hosting information, and classifying their phishness [7 and 8]. Besides, the usage of hybrid features supported the generality of the classification techniques to classify phishing variations and such techniques reported high rates of detection accuracy than those provided by their competitors [4, 6 and 9]. However, constraints like high-dimensionality of feature set, hybridity of features, their irrelevance to the corresponding classes (i.e. phish and legitimate), their dependency on each other, their redundancy on the examined feature space, and heterogeneity of their values (i.e. discrete and continuous values) might degrade detection accuracy. In addition, they might have increased the false detection errors and computational costs. Then, they would limit the overall effectiveness of classification techniques in the real-world experience along with their scalability to the enormous web data and the evolving phish exploits [5, 9].

Hence, to tolerate with the aforesaid issues, researchers had looked into their constructed classification models via feature selection methods that played an important role in data analysis during the classification task. Such methods typically refined the extracted set of features into a minimal and effective subset for the classification task. Besides, they eliminated the least representative features by applying the lowest discrimination on the tested data. However, these assisted methods yielded different outputs of feature selection. Meanwhile, as for the existing researches; specifically in phishing websites detection, the direct comparison of such differences had been neglected. In their evaluations, they

underlined the differences with respect to the detection accuracy and overall performance [4-12]. They rarely quantified feature selection methods in terms of (i) the measure of feature's prediction susceptibility that they had utilized, (ii) their scalability under different feature sets' dimensions, (iii) the goodness of their output in the presence of different classification models, (iv) the stability of their output against evolving data and phishing variations, and (v) the similarity between the outputs of multiple feature selection methods.

Besides, the causality between the aforesaid issues and the optimum choice of feature selection subset had been highlighted. It quantified the highest quality of selected feature subset that yielded the best case of detection accuracy with least error rate as possible. Moreover, this contribution is extended by testing the selected feature subset across multiple classification models. Apart from that, this study promotes its contribution by handling a proposed set of hybrid features. Hence, it is hoped that the proposed features, the characterized literatures, the highlighted issues, and the empirical tests would offer a global picture on phishing detection assisted by feature selection. Moreover, they could be regarded as the baselines for future works to appropriately choose the feature selection methods for their classification models.

In this context, this study characterizes the prior works, and critically appraises them with respect to their frontiers in feature selection as presented in Section II. Then, Section III recommends certain criteria and depicts their relevant terminologies to assess both resilience and effectiveness of selective feature subsets. Section IV, practically appraises feature selection exploits and testifies their outcomes in the presence of the recommended criteria. Based on the stated findings, Section V deduces the present work on hand and gives an outlook to the future implications.

II. BACKGROUND

A. Feature Selection Methods

All feature selection methods aim at reducing the dimensionality of the feature space and in enhancing the compactness of the features. Meanwhile, in data processing, specifically data mining and machine learning approaches; a large number of features may cause problems of high dimensionality, irrelevance, and redundancy [13]. Therefore, in order to reduce the dimensionality and to obtain the most representative features that could effectively predict instances over a given dataset, data pre-processing is needed [13 and 14]. Mainly, feature selection has been considered as a data pre-processing technique that chooses a minimum subset of *m* features from an original set of *n* features. Accordingly, the selection involves: a search procedure for feature subset generation, and an evaluation criterion for iterative feature selection [13, 14]. Furthermore, the search procedure often discards or adds one feature based on its evaluation outcome, whereas the evaluation criterion compares that feature with the previously selected one regarding to either its information, or dependency, or consistency, or distance or its transformation. However, feature selection methods differ in specifics and parameters that can be tuned for both the search procedure and the evaluation criterion [13, 14]. *Table I* enlists four feature selection methods that had been adopted for phishing detection in the reviewed literature, which were characterized by search procedure, as well as evaluation specifics and criteria.

TABLE I. CHARACTERIZATION OF FEATURE SELECTION METHODS (ADOPTED FROM [13-15])

Feature Selection Method	Search Procedure	Specifics	Evaluation Criterion
Information Gain (IG)	Filter	Information	$IG(S,a) = Entropy - \sum_{V \in a} \frac{\lvert S_V \rvert}{\lvert S \rvert} * Entropy(S_V)$ (1) "Where *S*, S_v, *V* and *a* are the collection of instances, a subset of instances with *V* of a, a relevant value and an attribute, respectively."
Correlation Based Feature Selection (CFS)	Filter	Consistency	$p(C = c \lvert V_i = v_i) \neq p(C = c)$ (2) "Where, V_i is said to be relevant if there exists some v_i and *c* for which $p(V_i = v_i) > 0$."
Chi-squared (χ^2)	Filter	Transformation	$\chi^2 = \frac{N \times (AD-CB)^2}{(A+C') \times (B+D) \times (A+B) \times (C+D)}$ (3) Where $A = \#(t,c)$, $B = \#(t,\neg c)$, $N = A + B + C + D$, $= \#(\neg t,c)$, $D = \#(\neg t,\neg c)$, and *t* and *r*are independent parameters
Wrapper Feature Selection (WFS)	Embedded with Classifier	Accuracy	Greedy search for feature subset in a forward selection and backward elimination of features

B. Related Works

At present, vast literature is available on the merits and demerits of phishing detection campaign. Towards devising anti-phishing solutions for the specific problem at hand (i.e. phishing websites), many proposals have been introduced and experiments conducted by using different machine learning-based approaches combined without features extraction and features selection. For instance, Likarish et al. [15] developed a Bayesian filter to identify phish websites based on retrieved tokens obtained from the HTML document and constructing DOM (Document Object Model) with the aid of DOM parser. Then, researchers at Google Inc., Whittaker, Ryner & Nazif [16]; worked on the up-gradation of Google's phishing blacklist integrated with a classifier. In addition, another anti-phishing technique was developed by Bergholz et al. [17] to phish email filtering by analyzing several extracted features related to body, external, and model based on examined emails. The developed techniques involved two training phases; one for model-based features and the other was for the rest of the features. Later, CANTINA$^+$ was proposed by Xiang, Hong, Rose, and Cranor [18] with three classifiers and ten features derived from the URLs and the contents of webpages, as well as some online features for highly accurate results of phishing detection. Meanwhile, Zhang Liu, Chow,

and Liu [19] 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 likelihood of the suspect website as being phished or legitimate. Likewise, a *Supervised Machine Learning* (*SVM*) classifier was developed by He et al. [8] to predict phishness on examined webpage by exploiting webpage identity and some textual features. The textual features were extracted by using a well-known information retrieval method to be deployed for classification process. Contrarily, a phish webpage detector was proposed by Li, Xiao, Feng, and Zhao [20] based on visual features and DOM objects of the webpage content that learned and tested over datasets by using Semi-Supervised Machine Learning (*TSVM*) classifier. Furthermore, Kordestani and Shajari [21] applied three classifiers, including Naïve Bayes (*NB*), *Supervised Machine Learning* (*SVM*), and *Random Forest* (*RF*), on a randomly selected dataset to predict phishness in suspected websites. They were deployed for phishness prediction with the presence of URL and online features. Then, Gowtham and Krishnamurthi [22] extracted fifteen, which were trained by using *Supportive Vector Machine* (*SVM*) classifier and a whitelist through two modules. The first module involved checking the identity features of the examined website against a pre-defined white list of legitimate ones, whereas the second module predicted phishness of the examined webpage based on 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 datasets, the extraction of hybrid features, the analysis of their heterogeneity, increasing storage requirements and processing time, as well as some costly miss-classifications.

Moreover, it is worthy to mention that final decisions of phishing detection relied potentially on predictive features against phishing susceptibility. More precisely, phishing detection in the presence of predictive features should yield minute amounts of both valid phish misclassifications and losses of valid legitimate instances. Thus, researchers were motivated to maintain some feature selection methods as those briefly described in *Table II* to cope with the aforesaid factors. In the literature, Pan and Ding [23] proposed phishing detector based on applying *Supportive Vector Machine* (*SVM*) classifier and extracting both textual and Document Object Model (*DOM*) features from the examined webpages. They employed two major components for 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 Ofoghi, Watters, and Brown [24] experimentally analyzed seven webpages and pages to rank the features with the aid of a filter-based feature selection method, *Information Gain (IG)*, to phish website classification and deploy two classifiers that varied in their classification accuracy due to the selected features. On top of that, Khonji, Jones, and Iraqi [25] enhanced classification performance by selecting the most effective subset of the most commonly used 47 features. Both filter-based and Wrapper-based feature selection methods, such as *Information Gain* (*IG*), *Correlation Based Feature Selection (CFS)*, and *Wrapper Feature Based Selection (WFS)*, were developed with machine learning classifiers to predict phish emails. The classification results differed due to the employed feature selection method and the number of selected features. On the other hand, Basnet, Sung, and Liu [26] analyzed high dimensional feature space, including 177 features extracted from both the content and URL of websites to select the best feature subset. In fact, several subsets were considered for application of *Wrapper Feature Based Selection* (*WFS*) and *Correlation Based Feature Selection* (*CFS*). They were trained over a dataset with the aid of *Logistic Regression* (*RF*) classifiers. Nevertheless, they varied in selecting the most contributing features such that classifiers caused variation on detection accuracies. Later, Zhang, Jiang, and Kim [27] developed automatic detection approach for Chinese e-business websites by incorporating the unique features extracted from URL and contents of website. Alongside, Hamid and Abawajy [28] proposed a multi-tier detector to phish emails filtering with the aid of Adaboost and SMO classifiers in an ensemble design. Moreover, they used *Information Gain (IG)* and clustering strategy to quantify the best predictive features of phish emails and also tested the outcomes over three large scale datasets. However, large size dataset, imbalanced datasets, redundancy, the limit of cluster size, and error rates emerged as the key issues in their work.

C. Shortages

In order to offer a global view on feature selection for exploitation in phishing detection domain, *Table II* characterizes the previous works with respect to their deployed feature selection methods and their limitations.

TABLE II. RELATED WORKS WITH LIMITED FEATURE SELECTION METHODS

Citation	Feature Selection Method (S)	Classifier (S)	Related Limitations
Pan and Ding, 2006 [23]	χ^2	SVM	▪ Heterogeneity of features values ▪ Dissimilarity of selection outputs ▪ Computational cost ▪ Redundancy and irrelevance
Ma et al., 2009 [24]	IG	C4.5	▪ Heterogeneity of features values
Khonji, Jones and Iraqi, 2011[25]	IG, WFS, CFS	RF	▪ Dissimilarity of selection outputs ▪ Imbalanced Data ▪ No scalability
Basnet., 2011 [26]	CFS, WFS	LR, RF, C4.5	▪ Computational cost ▪ Dissimilarity of selection outputs ▪ Heterogeneity of features values ▪ Redundancy and irrelevance
Zhang, Yan and Jiang, 2014 [27]	χ^2	SMO, LR, NB	▪ Redundancy and irrelevance
Hamid and Abwajy, 2014 [28]	IG	Adaboost, SMO	▪ Heterogeneity of features values ▪ Non-scalability ▪ Computational cost ▪ Dissimilarity of selection outputs

As depicted in *Table II*, the surveyed works often deployed sub-optimal feature subsets for phishing detection due to some limitations. Such limitations include: the dependency of feature selection outcomes on a given dataset, different feature selection outcomes across different classification models, heterogeneity of features values, and un-scalable feature selection method to more challenging datasets [23-28]. Furthermore, most of the dedicated efforts focused on discarding the relevant features rather than the redundant ones during feature selection [23-28]. Besides, since they are mutually dependent on other features belonging to the same targeting class; the redundant features might distort the classification task and then degrade its accuracy by producing high error rates [29 and 30]. Consequently, *Table III* underlines some striking issues like non-scalability, heterogeneity, non-robustness, irrelevance, and redundancy that must be considered to deal with feature selection limits [29-31].

TABLE III. STRIKING ISSUES OF FEATURE SELECTION, ADOPTED FROM [29-31]

Striking Issues	Description
Non-scalable Feature Subset [29]	The deployed features rarely raise the classification accuracy to the best case as possible under different selection scenarios and over different datasets.
Redundant Features [29, 30]	Since the high-dimensional data have a substantial amount of irrelevant features which require high computational cost selection strategy to reduce. Such strategy potentially causes inefficient classifier. Irrelevant features, in turn, may contain redundant and non-redundant features which require a robust feature selection strategy capable to handle their redundancy.
Irrelevant Features [29, 30]	Large scaled and realistic datasets like that involved in anti-phishing techniques the may contain high fraction of irrelevant features. Because of the exponential growth of more sophisticated and deceptive phishing features, the resultant irrelevant features highly degrade the classifier's performance.
Feature Values Heterogeneity [31]	Websites are inconsistent datasets with various hybrid features that have different values - discrete, categorical and continuous values. For any collected dataset, the extracted hybrid feature space is heterogeneous in values and huge in size. That is, in the presence of any extracted or selected subset of features, the machine learning classifier should be able to categorize them for both training and testing purposes with a minimum loss of feature values.
Non-robust Feature Subset [31]	When applying feature selection for knowledge discovery, robustness of the feature selection result is a desirable characteristic, especially if subsequent analyses or validations of selected feature subsets are costly. Modification of the dataset can be considered at different levels: perturbation at the instance level (e.g. by removing or adding samples), at the feature level (e.g. by adding noise to features), or a combination of both.

III. ASSESSMENT MEASURES

Other than that, as for the problems at hand (*Table III*), the outcomes of selective feature subset must be quantified on its scalability, goodness, stability, and similarity over multiple datasets [29-33]. In addition, the assessment of outcomes prediction susceptibility against phishing over different datasets is a noteworthy issue to be highlighted towards obtaining the most advantageous features [34]. Thus, specific measures adopted by prior researchers in different fields have been recommended in this work (*Table IV*) to test and to assess the outcomes of feature selection methods [31-37]. Such measure can be considered as comparison baselines for any further study on feature selection effects.

TABLE IV. RECOMMENDED EVALUATION MEASURES FOR FEATURE SELECTION [32-38]

Metrics	Advantage	Evaluation Criterion
Goodness [31]	It measures how well the selected feature subset can accurately classify extremely imbalanced datasets.	$Goodness(S_i) = \frac{1}{Y}\sum_{i=1}^{Y}\frac{N_i^{tp}}{N_i}$ (4) 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 [31, 32]	It quantifiably proves whether the selected features are relatively stable against variations of real world datasets over a period of time.	$Stab(S) = \sum_{f_i \in X}\frac{F_{f_i}}{N} \times \frac{F_{f_i}-1}{\lvert D\rvert-1}$ (5) 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 [31, 33]	It compares the behaviour of multiple feature selection methods and their selected features on the same data.	$Sim(t_1, t_2) = 1 - \frac{1}{2}\sum\left\lvert\frac{F_{f_i}^{t_1}}{N^{t_1}} - \frac{F_{f_i}^{t_2}}{N^{t_2}}\right\rvert$ (6) 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 selection methods t_1 and t_2 respectively. Similarity takes values within[0,1].
Prediction Susceptibility Or Phishness Ratio [34]	A ***phishness ratio*** restates the prediction susceptibility of selective feature set to phishing upon each instance in the dataset. The probability $P_r(P \mid t_i)$ of estimated phishness along with a feature t_i is computed across all instances in the dataset. Then, the instance's phishness is computed by averaging the probability of all its related features.	$P_r(P \mid t_i) = \frac{N_{t_i \to P}}{N_{t_i \to P} + N_{t_i \to L}}$ (7) $Phishness(S) = \frac{\sum_{i=1}^{n}\Pr(P\mid t_i)}{n}$ (8) Where S is the examined webpage, *Phishness (S)* is the prediction of phishing susceptibility, t_i is the feature in *S*, $N_{t_i \to P}$ is the number of occurrences of t_i in phish instance, $N_{t_i \to L}$ is the number of occurrences for t_i in legitimate instance. and *n* is the number of features in *S*.
Minimal Redundancy [35, 36]	It eliminates duplicate features that having another one replicate them in the dataset.	$Min\ R(S) = \frac{1}{\lvert S\rvert^2}\sum_{x_i, x_j \in S} I(x_i, x_j)$ (9) Where *R(S)* is the set of highest mutually exclusive features that selected between x_i and x_j.
Maximal Relevance [35, 36]	It selects most relevant features to the target class and highly affecting the classification output.	$\mathrm{Max}\, D(S, c) = \frac{1}{\lvert S\rvert}\sum_{x_i \in S} I(x_i, c)$ (10) Where *D(S,c)* is the mean value of all mutually informative features x_i with respect to class *c*.

mRMR* $\Phi(D,R)$ *[37]	This criterion selects a subset feature compactness composed of the most relevant and least redundant features from the original set simultaneously.	$Max\,\Phi(D,R), \Phi = D - R.$ (11) Where, *D and R* indicate the dependency between a feature x_i and its class, and the highest relevance between features x_i and x_j in the same feature set.

IV. Empirical Test and Discussion

Based on the recommended measures presented in *Section III.C*, the empirical test was conducted to state not only the variations of assisted feature selection methods on *Prediction Susceptibility*, *Goodness*, *Stability*, *Similarity*, and *Scalability*, but also it assessed outputs of the simultaneous discarding criterion of redundant and irrelevant features (*mRMR*). To the best of our knowledge, this type of empirical test with the aid of the recommended criteria is scarcely underscored in the literature of phishing detection despite of its significance for feature selection. Hence, an empirical test was implemented on a specific test-bed that was set to extract a large number of hybrid features. Then, a comparison was made on the effectiveness of the best chosen feature subset across different classification models. Test-bed is described, results are reported, and discussion is summarized in the following:

A. Test-Bed and Features

A wide range of aggregated phish and legitimate webpages were considered as test-bed for this study. Mostly they are reported in public archives such as *PhishTank*, *CastleCops*, and *Alexa*. Both *PhishTank* and *CastleCops* are phishing data archives that volunteers frequently update them with valid living phish webpages. While, *Alexa* archive is publicly used to retrieve valid legitimate webpages. We chose such archives because they were commonly used by prior researchers in the literature of phishing detection [15-28]. *Fig. 1* illustrates the aforesaid test-bed in terms of dimension, the number of phish webpages and the number of legitimate webpages. In *Fig. 1*, the test-bed consists of three multiple datasets: Dataset1, Dataset2 and Dataset3. Dataset1 composed of 1000 webpages, Dataset2 composed of 5000 webpages but Dataset3 consists of 10000 webpages. Multi-dimensional test-bed helped to empirically assess the outcomes of the reviewed feature selection methods towards demonstrating the most suitable one among them for phish website detection. Indeed, the webpage content and URL can be used to characterize each instance included in the aforesaid datasets such that they can be categorized accordingly to a specific class either phish or legitimate.

Consequently, the characterized datasets with their features and corresponding classes helped to generate the required feature space. *Fig. 2(a)* illustrates the structure of the generated feature space in terms of class label, feature index, the feature itself, and its value. Furthermore, *Fig. 2(b)* shows a part of the database schema to provide a global view on how raw data could be generated. Moreover, a set of web development tools, such as Fireburg, Jsoup, and Import.Io, had been helpful in implementing this task. Besides, several publicly used tools, such as KNIME and WEKA -the Waikato Environment for Knowledge Analysis, were employed for feature selection implementation and tests.

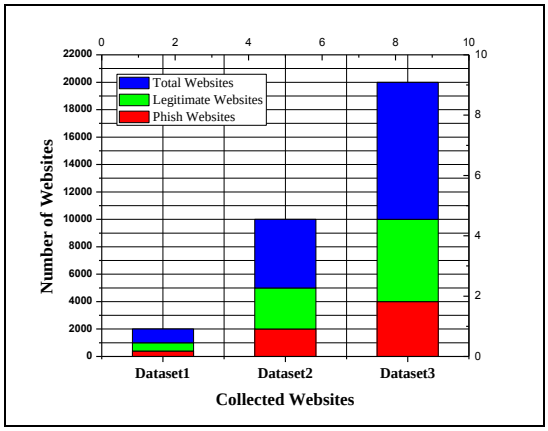


Fig. 1. Description of collected datasets in terms of the total number of instances, legitimate websites and phish websites

m Samples	*Feature Vectors*	*Class Label*	*Features*				
	i	C_j	$f_{j,1}$	$f_{j,2}$	$f_{j,i}$	…	$f_{j,n}$
	j		$v_{j,1}$	$v_{j,2}$	$v_{j,i}$	…	$v_{j,n}$
	W_1	C_1	$v_{1,1}$	$v_{1,2}$	$v_{1,i}$		$v_{1,n}$
	W_2	C_2	$v_{2,1}$	$v_{2,2}$	$v_{2,i}$		$v_{2,n}$
	W_j	C_j	$v_{j,1}$	$v_{j,2}$	$v_{j,i}$		$v_{j,n}$
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	W_m	C_m	$v_{m,1}$	$v_{m,2}$	$v_{m,i}$		$v_{m,n}$

(a)

Index	C	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	33	17	0	0	1	232	124	54	6	10	1	10	69	16	14	0	1	1
2	1	61	15	1	0	0	91	12	10	10	1	0	6	34	16	2	1	1	0
3	1	15	55	1	1	1	56	79	87	12	23	0	9	7	8	9	1	1	1
4	1	20	9	0	1	0	32	43	35	9	10	0	3	13	12	11	0	0	1
5	1	12	17	1	1	1	45	19	58	5	8	0	7	20	51	11	1	0	1

(b)

Fig. 2. Illustrative example of (a) generated dataset structure and (b) database schema

In *Fig. 2(a)*, the j^{th} webpage is characterized as a vector of features W_j. Then, all feature vectors extracted from *m*-dimensional set of webpages are represented as combined together in a feature matrix M such that $M = \{W_1 \;\; W_2 \;\; W_m\}$; where m indicates the number of feature vectors included in *M*. Each entry vector W_j in *M* consists of its features' indexes and their corresponding values along its corresponding class label as the first column, i.e. $W_j = \{C_j, (f_{j,1}, v_{j,1}), \;\; (f_{j,2}, v_{j,2}), \ldots \;\; (f_{j,n}, v_{j,n})\}$; where n is the number of features, $f_{j,i}$ is the index of each i^{th} feature of j^{th} feature vector W_j, where $0 \leq f_{j,i} \leq 1$, $i = 1, 2, 3, \ldots, n$ and $j = 1,2,3, \ldots, m$. Whereas C_j is the label of the class such that $C_j \in \{1, 0\}$ with $C_j = 1$ and $C_j = 0$, which indicates the membership of W_j in the phish class or in the legitimate class based on its corresponding features [38, 39] as portrayed in *Fig. 2(b)*. Further, features of Boolean values are mapped into either 0 or 1, and features of Continuous quantities are represented as numeric quantities. *Appendix I* enlists the original set of features extracted from all webpages included in the test-bed. Totally 58 features were included in the original feature set. 48 features were extracted from specific parts, tags and scripts in the webpage source code. Besides,

ten features were extracted from the indicators of webpage URLs. This high-dimensional set of features will be refined later to a subset of selected features using several feature selection methods as it will be presented in the next subsection.

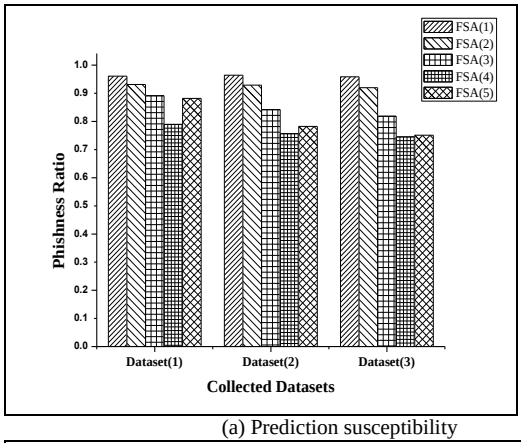
B. Comparison Across Feature Selection Methods

In this section, all the details and discussions of the first empirical test and the related findings are presented. The test was conducted on four *feature selection algorithms* (FSAs); namely *CBF, WFS, χ^2,* and *IG;* which had been previously adopted in the surveyed works. Besides, the *mRMR* feature selection method was also involved in the comparison to qualify if it could be recommended as an alternative FSA for the problems at hand (i.e. features' redundancy and irrelevance). Among its competitors those mentioned in *Table I*, *mRMR* discards redundant and irrelevant features in parallel and yields a selective subset of the most relevant and least redundant features together in a compact combination. Hence, both test and comparison were achieved in the presence of three datasets with different sizes and collections of phish and legitimate instances, as presented in *Fig. 3* and *Fig. 4.*

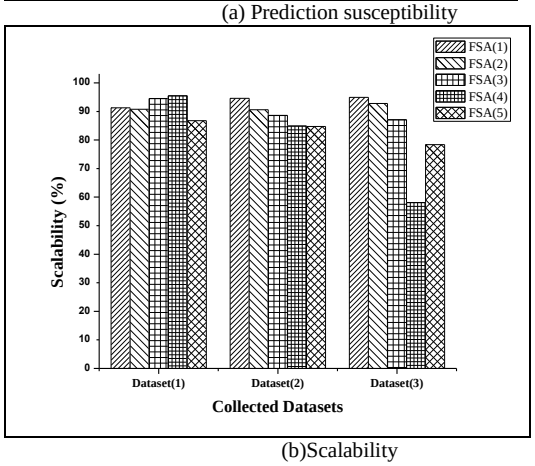
In this comparison, all the tested FSAs were practically appraised on prediction susceptibility (*Fig. 3(a)*), and scalability (*Fig. 3(b)*), goodness (*Fig. 4(a)*), stability (*Fig. 4(b)*), and similarity (*Fig. 4(c)*) to show their variations and likelihood. In *Fig. 3* and *Fig. 4*, FSAs 1, 2, 3, 4, and 5 are referring to *mRMR, CBF, WFS,* and *$\chi 2$*and *IG* respectively.

From *Fig. 3* and *Fig. 4* , the overall results are very encouraging towards deploying all the selective hybrid features as predictive ones on phishing websites. The only difference is the variation of their compactness by using different FSAs. Findings of this test are summarized as follows:

- In *Fig. 3(a),* the evaluation and comparison of their prediction susceptibilities were done by using the measure of *Phishness Ratio* (*Table IV.*). *Phishness Ratio scores* showed that the selected feature subsets chosen by using FSAs 1, 2 and 3 (i.e. *mRMR, CBF,* and *WFS*) reached the highest peak among their competitors over all datasets; whereas the feature subsets of FSAs 4 and 5 (i.e. *χ^2* and *IG*) produced the lowest peaks. Such findings point out the significance of features' mutual information for selecting the best feature subsets. More importantly, they demonstrate that the discarded redundant and irrelevant features were least predictive features among the others. And such generated compactness of most relevant and least redundant features increases their *Phishness Ratio*. Further, this test pointed out that *mRMR* criterion can be considered as a promoting technique to improve the overall discriminating behavior of the classification model websites. FSA 1 (i.e. *mRMR*) reduced both redundant and noisy features that are the prime objective of feature selection.



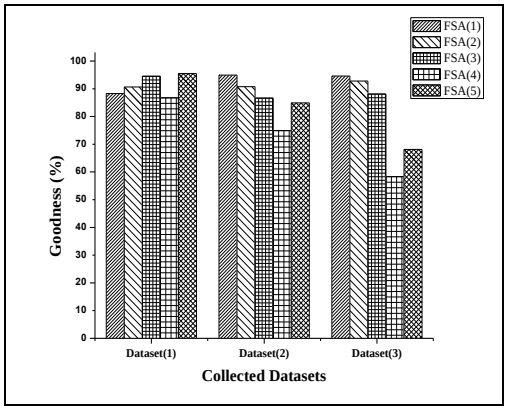
(a) Prediction susceptibility



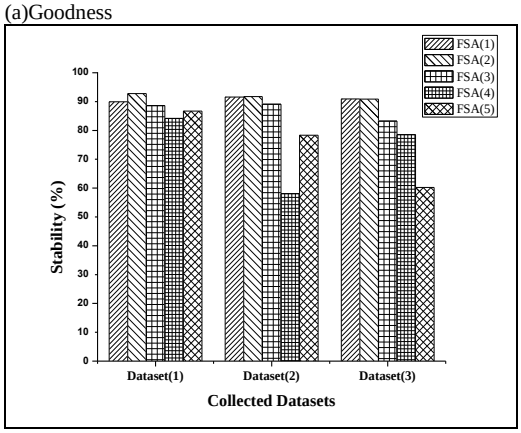
(b)Scalability

Fig. 3. Illustration of empirical test across four five feature selection methods. Each of FSA 1, 2, 3, 4, and 5 refers to mRMR, CBF, WFS, χ^2, and IG respectively

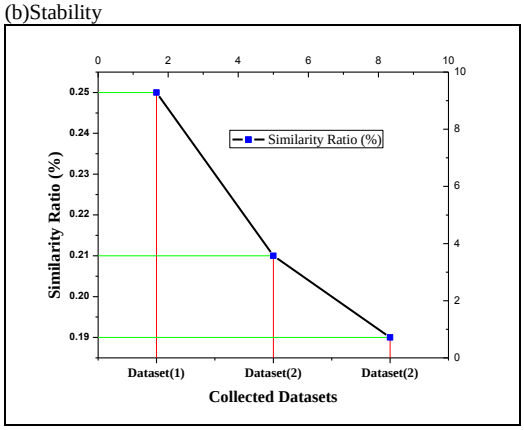
- *Fig. 3(b)* portrays the outcomes of scalability comparison. It shows that the feature subset chosen by using FSA 1 (i.e. *mRMR*) could successfully rise the score of prediction from the typical case to the best one over datasets having different sizes. This, in turn, restates that *mRMR* can be considered as the most scalable FSA among the others because it could preserve its prediction rate as close to the best case as possible.
- *Fig. 4(a)* qualified the goodness of the selected subsets over the three different datasets. It is clearly shown that FSA 1 (i.e. *mRMR*) still preserves the best case of goodness (i.e. quality) among the others despite of the volume variations of the utilized test-bed. But both of FSAs 4 and 5 (i.e. *χ^2* and *IG*) have the worst case of quality among the others. This implies that the significance of reducing feature set's dimensionality, and removing both redundant and noisy features to define the best features subset. Indeed, such feature subset will help the classification model to well perform over all datasets. More interestingly, such feature subset is needed to effectively detect phishing websites in realistic applications.



(a)Goodness



(b)Stability



(c)Similarity

Fig. 4. Illustration of empirical test across five feature selection methods; where: FSAs 1, 2, 3, 4, and 5 refer to mRMR, CBF, WFS, χ^2, and IG respectively

- *Fig. 4(b)* outlines how the feature subsets chosen by FSAs 1and 2 (i.e. *mRMR* and *CBF*) are notably more stable over all datasets than their competitors. Further, it emphasizes the significance of the inter-dependencies between the features in the same chosen feature subset. Features chosen on their inter-dependencies can compose a stable subset under different detection scenarios and datasets. In contrast, those subsets chosen with respect to the topmost ranking of their constituents like FSA 5 (i.e. *IG*) may vary in their discriminating power against vast dataset and different detection approach.

- In the context of overall outputs' similarity (*Fig. 4(c)*), it can be observed that FSAs' outputs are notably dissimilar over all the datasets. The reported similarity scores are lower than (0.3) which point out that the selected subsets overlap partially and they are complementary to each other's. Interestingly, such dissimilarity implies that feature subset composed of hybrid and diversely predictive features could be a promising avenue to improve the classification performance. Moreover, FSAs produce dissimilar feature subsets can be effectively integrated and exploited for a specific phishing detection approach. Despite this, it is clearly observed that the optimal feature subset chosen by specific FSA, it may be considered as sub-optimal choice regarding to another FSA. Hence, both likelihood and difference of FSAs outputs are crucial issue in a machine learning based detection approaches.

- Based on the overall results, we obtained a useful insight into the crucial importance of feature selection method for the problem domain at hands. This, in turn, enables us to improve the detection performance in the context of using as few, predictive and robust features as possible. In general, looking at the aforesaid test and its overall findings highlights the significance of selective feature subset in terms of prediction susceptibility, scalability, goodness, and stability. In particular, feature subset chosen by FSA 1 (i.e. *mRMR*) always has the first best scores in terms of the aforesaid perspectives among the others. Whilst, FSAs 2 and 3 (i.e. *CBF* and *WFS*) reveal the second and third best cases among the others. Contrarily, both FSAs 4 and 5 (i.e. χ^2 and *IG*) yield the worst cases across all the aforesaid perspectives.

In summary, this empirical test restates that several selection methods reach a quite bit similar peaks of prediction susceptibility and robustness. Therefore, they can be considered as the baseline methods for feature selection in phishing website detection. More importantly, if the feature selection method is carefully chosen, i.e. on the basis of its prediction susceptibility and robustness; the performance of the classification model could be highly improved with low latency and errors. However, there is still no exact answer for the perfect FSA among all the tested ones unless they assessed in terms of detection accuracy, specificity and sensitivity across several classification models and different datasets. This issue will be considered in the next subsection.

C. Comparison Across Classification Models

Herewith, we turn to qualify how the aforesaid selective subsets of features can shift detection accuracy, specificity and sensitivity of the classification model to the best rates as possible. The qualification is determined through two comparisons. First, the outputs obtained from the previously tested FSAs are compared on detection accuracy, detection sensitivity and specificity over training and testing datasets dedicated for this purpose. To accomplish the performance test and get findings for comparison, a specific machine learning classifier was applied; namely, *C4.5* as can be seen in

Fig. 5. Meanwhile, several supportive metrics are deployed for the performance evaluation as presented in *Table V*.

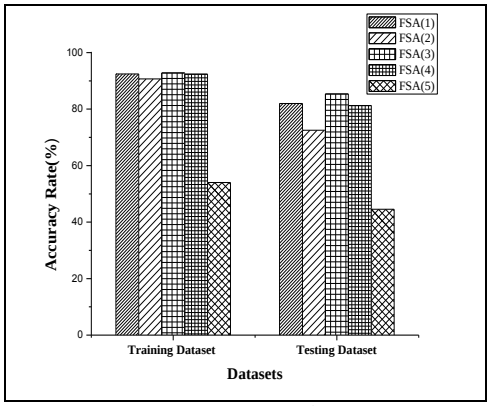
To qualify the discriminating behavior, four machine learning classifiers are involved in the second comparison. Those classifiers are described with their related calculations in *Table VI.* Such classifiers are chosen because of their wide use in the literature of phishing detection. Consequently, this comparison highlights how the best selective feature subset could classify phishing websites not only across different datasets (i.e. training and testing datasets) but also across different classification models as illustrated in *Fig. 6.*

Both comparisons are applied over two datasets: training and testing datasets that generated from a collection of phishing and legitimate webpages specifically aggregated for this purpose. The datasets are generated through extracting the features space from the aggregated webpages (i.e. data pre-processing) and dividing it into a training dataset (70% of the main dataset) and a testing dataset (30% of the main dataset).

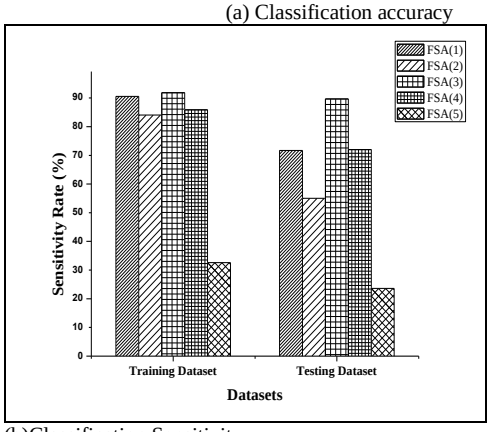
TABLE V. PERFORMANCE EVALUATION MEASURES [1, 3]

Metrics	Evaluation Criterion	Mathematical Formula
TP	True Positive indicates the rate of correctly classified phishing instances.	$\frac{N_{P\to P}}{(N_{P\to P}+N_{P\to L})}$ (12)
FP	False Positive refers to the rate of wrongly classified legitimate instance s as phishing ones.	$\frac{N_{L\to P}}{(N_{L\to L}+N_{L\to P})}$ (13)
TN	True Negative refers to the rate of correctly identified legitimate instances.	$\frac{N_{L\to L}}{(N_{L\to L}+N_{L\to P})}$ (14)
FN	False Negative indicates the wrongly labeled phishing instances as legitimate ones.	$\frac{N_{P\to L}}{(N_{P\to P}+N_{P\to L})}$ (15)
Specificity	The percentage of correctly positive predictions	$\frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FP\rvert}$ (16)
Sensitivity	It refers to the percentage of correctly predicted positive instances (TPs).	$\frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FN\rvert}$ (17)
Accuracy	It indicates the overall rate of correctly detected phishing and legitimate instances (the rate of correct predictions).	$\frac{\lvert TP\rvert+\lvert TN\rvert}{\lvert TP\rvert+\lvert TN\rvert+\lvert FP\rvert+\lvert FN\rvert}$ (18)

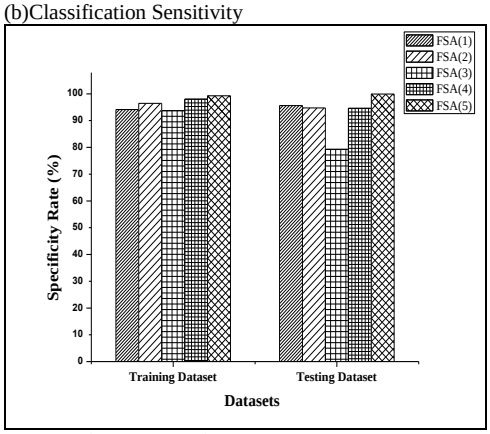
Where: $N_{P\to P,}$ $N_{L\to P,}$ $N_{P\to L,}$ $N_{L\to L}$ denote the number of correctly labeled phishing instances, the number of wrongly labeled legitimate instances, the number of phishing instances that are incorrectly recognized as legitimate, and the number of legitimate instances that are identified correctly as legitimate respectively [1, 3].



(a) Classification accuracy



(b)Classification Sensitivity



(c) Classification specificity

Fig. 5. Outcomes on classification performance with the aid of C4.5 classifier and all tested feature selection methods. Each of FSAs 1, 2, 3, 4, and 5 refers to mRMR, CBF, WFS, χ^2, and IG respectively

TABLE VI. EXAMPLS OF MACHIENE LEARNING CLASSIFIERS PREVIOUSLY ADOPTED IN PHISHING DETECTION [20, 40-44]

Machine Learning Classifier	Description	Related Calculation (s)
C4.5	It is a Decision Tree hypothesis that depends on a tree structure to construct a classification model. Its nodes represent features, its branches denote the features values whereas the leaf nodes denoting the final class decision.	The final decision of an instance to be classified relies on tracing the path of nodes and their branches to the terminating leaf nodes.
Decision Tree (DT)	It models the data with a rooted tree that contains: nodes, edges and leaves. Nodes are labeled corresponding to features, edges are labeled with the feature values and leaves are labeled with classes.	Instances of unknown class are classified by ordering them according to their feature values in the rooted tree such that features are denoted by nodes and their values are represented by branches that the node assumes. The classification of unknown instance is started at the root node and then passed through the tree. The test at each node along the path is applied to the sorted feature values that determine the next edge until ending up at the leaf nodes. The label of the ended up leaf node is the final decision of classification.
Naïve Bayes (NB)	A probabilistic classifier with assumption of conditionally independent attributes of each other given class of instances.	$P(C\|X) = P(C\|x_1, \dots, x_n) = \frac{P(C)P(x_1,\dots,x_n\|C)}{P(x_1,\dots,x_n)}$ (19) Where X is a given sample with a vector of *n* features$(x_1, \dots, x_n)$, C is the class label that the classifier seeks for maximizing the likelihood.
Support Vector Machine (SVM)	It is an optimistic separating hyper-plane that maximizes the margin between closest points of two classes to estimate the decision function.	$min\ \frac{1}{2}w^T w + C\sum_i \xi_i$ (20) Subject to: $y_i\big((w^T \cdot x_i) + b\big) \geq 1 - \xi_i, \xi \geq 0, i = 1, 2, \dots, m,$ (21) $max\ \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j K\big(x_i, x_j\big)$ (22) Subject to:$0 \leq \alpha_i \leq C, i = 1, 2, \dots, m\ and\ \sum_{i=1} \alpha_i y_i = 0.$ (23) Where x_i is M-dimensional data vector $x_i \in R^m$ with samples belong to either one of two classes labeled as $y \in \{-1, +1\}$that it is separated by a hyper-plane of $(w \cdot x) + b = 0$. $\alpha_i$denotes the Lagrange multipliers for each vector in the training dataset and it is used to transform the original input space to higher in dimension space.
Transductive Support Vector Machine (TSVM)	It separates the positive and negative samples included in the training dataset with a maximal margin by using SVM hyper-plane. It outperforms SVM with good generalization accuracy.	Minimize over $(y_1^*, \dots, y_k^*, w, b, \xi_1, \dots, \xi_n, \xi_1^*, \dots, \xi_k^*)$, $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_1 + C'\sum_{j=1}^{k}\xi_j^*$ (24) Subject to: $\forall_{i=1}^{n}: y_i[wv_i + b] \geq 1 - \xi_i$, $\forall_{j=1}^{k}: y_i[wv_i^* + b] \geq 1 - \xi_j^*, \forall_{i=1}^{n}: \xi \geq 0$, and $\forall_{j=1}^{k}: \xi_j^* \geq 0.$ (25) Where x_i is an *m*-dimensional vector such that$x_i \in R^m$ with independent labeled samples belong to either one of two classes labeled as$y \in \{-1, +1\}$, ξ and ξ^* are the slack variables of training and testing datasets, respectively. C and C' denote the influencing parameters determined by the user. The effect term of the j[th] unlabeled sample is denoted by $c^*\xi_j^*$.

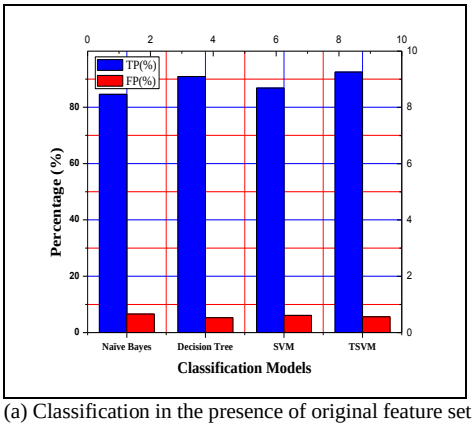
Regarding *Tables V and VI* as well as the statistics plotted in *Fig. 5* and *Fig. 6*, the following standpoints are inferred:

- The significant differences between classification models assisted by the tested FSAs (*Fig. 5*) point out the major or minor contribution that the assisted feature selection method can provide. Variations in accuracy, sensitivity, and specificity demonstrate that not all the tested feature selection method yield promising outcomes on phish website detection. This is because of (i) variations on specifics and evaluation criteria of FSAs themselves, (ii) the chosen features themselves due to their varied prediction susceptibilities and robustness, (iii) the inter-dependency of detection performance on the deployed classification model itself, (iv) the type of exploited features (i.e. webpage's URL and /or webpage's content) and (v) the dimension of the selected feature subset (i.e. the number of features included in the selected subset).
- Consequently, different outcomes of performance test (*Fig. 5*) show that certain classification model may sensibly being influenced by the training and testing datasets, and the suitability of machine learning classifier as well as the chosen feature selection method. This implies that the diversity and pre-processing of the collected dataset likely influence the overall classification performance because the dataset may encompass imbalanced data. More precisely, the imbalanced data indicate the divergent abundance of features corresponding to the classes of phishing or legitimate over the collected test-bed. Since the collected test-bed is quite bit different in dataset size and it consists of a dozen of labelled and unlabeled instances having a variety of features (i.e. hybridity), and a heterogeneity of features values. Therefore, k-fold validation and chronological assessment must be attained to come up with such diversity.

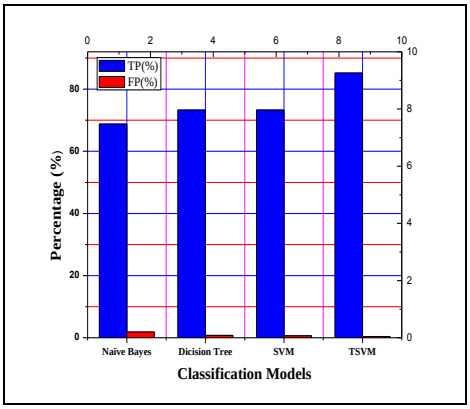
- The classification performance is likely to be influenced by the set of many features (*Fig. 6 (a)*). For instance, 58 extracted hybrid features may encompass irrelevance, redundancy and noisy data; therefore, eliminating the worst features and selecting the best ones (i.e. the most representative ones) are important inductive factors for well-performed classification as can be recognized in *Fig. 6(b)*.
- Also, the feature set's dimensionality is an important factor for the classification performance (*Fig. 6(a) and Fig. 6(b)*). As more features are being processed as more computational cost is being consumed. Moreover, the feature set's dimensionality interacts with the dataset's dimensionality.
- Selected feature subset chosen by the *mRMR* promotes the overall performance of classification models. Classification models assisted by *mRMR* outperform those baseline models in terms of classification accuracy and error rates (*Fig. 6(a) and Fig. 6(b)*).

V. Conclusions and Future Work

In the light of selecting a minimal and effective feature subset for well-performed phish website detection technique, this paper critically and practically appraised the exploitation of the feature selection via classification-based techniques. In this appraisal, those techniques assisted by machine learning classifiers and feature selection methods were involved, as well as a review of prior works with their related issues.



(a) Classification in the presence of original feature set



(b)Classification in the presence of selective feature subset chosen by mRMR

Fig. 6. Outcomes on classification performance across different classifiers

Further, empirical tests are conducted over 58 new hybrid features, five different datasets and five different classification models. Promoting measures are introduced to assess the outcomes of applied feature selection methods and then qualify the most suitable one among them for the problem at hands. Deeper understanding to their effects and significant gains on their outcomes' prediction susceptibility, scalability, goodness, stability and similarity are obtained respectively. Moreover, feature selection outcomes are compared on how they can notably improve the overall classification performance towards finding an optimal anti-phishing solution.

As a result, the findings displayed that some feature selection methods significantly outperformed their competitors by exhibiting better robustness, prediction, and performance. Between, other methods diverted from the best and the worst cases in relation to the aforesaid quantified factors. This was caused by the variations in dataset sizes and their constituent instances, the compactness of the chosen features and the features themselves, the evaluation criteria of the selected methods, and the discriminating behavior of the applied classifiers on training and testing instances. Moreover, the empirical tests addressed that the appropriately chosen set of features outperformed the original set of extracted features and/or the individual features themselves with least latency. However, the notably powerful selection method (i.e. *mRMR*) failed to provide an ideal subset of features; it could only produce as minimal and effective feature subset as possible. Nonetheless, *mRMR* could deal with the problematic features of redundancy and irrelevance at once. However, it is worthy to mention that no precise feature selection method existed in this study to cope with all the classification models. Hence, the forthcoming work will quantify feature selection outcomes concerning the processing time and misclassification costs. With that, more classification models will be involved in a remedial framework for feature selection towards rational phish website detection.

References

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

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

[3] 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).2070-2090, 2013.

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

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

[6] 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).4697-4706, 2013..

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

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

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

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

[11] S. Gastellier-Prevost, G. G. Granadillo, and M. Laurent, "Decisive heuristics to differentiate legitimate from phishing sites", 2011 Conference on Network and Information Systems Security (SAR-SSI), 1-9, 2011.

[12] 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).4229-4235, 2012.

[13] 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, 153-167, 2006.

[14] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research", ASU feature selection repository, 2010.

[15] 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, 1745-1749, 2008.

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

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

[18] 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). 21, 2011.

[19] 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), 1532-1546, 2011.

[20] 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), 6027-6033, 2013.

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

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

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

[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., 493-497, 2009.

[25] 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, 135-144, 2011.

[26] 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, 252-261, 2012.

[27] 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), 845-853, 2014.

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

[29] 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), 260-272, 2009.

[30] L. Yu, and H. Liu, "Efficient feature selection via analysis of relevance and redundancy", The Journal of Machine Learning Research, 5, 1205-1224, 2004.

[31] 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), 2040-2057, 2013.

[32] Z. He, and W. Yu, "Stable feature selection for biomarker discovery", Computational Biology and Chemistry, 34(4), 215-225, 2010.

[33] N. Dessì, and B. Pes, "Similarity of feature selection methods: An empirical study across data intensive classification tasks", Expert Systems with Applications, 42(10), 4632-4642, 2015.

[34] 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.

[35] S. Lee, Y. T. Park, and B. J. Auriol, "A novel feature selection method based on normalized mutual information", Applied Intelligence, 37(1), 100-120, 2012.

[36] S. Tabakhi, and P. Moradi, "Relevance–redundancy feature selection based on ant colony optimization", Pattern Recognition, 48(9), 2798-2811, 2015.

[37] 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), 1226-1238, 2009.

[38] E. Uzun, H. V. Agun, and T. A. Yerlikaya, "A hybrid approach for extracting informative content from web pages", Information Processing & Management, (49), 928-944, 2013.

[39] 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), 40-43, 2010.

[40] A. Ben-Hur, and J. Weston, "A user's guide to support vector machines", Data mining techniques for the life sciences, pp. 223-239, Springer, 2010.

[41] G. Kumar, K. Kumar, and M. Sachdeva, "The use of artificial intelligence based techniques for intrusion detection: a review", Artificial Intelligence Review, 34(4), 369-387, 2010.

[42] D. Miyamoto, H. Hazeyama, and Y. Kadobayashi, "An evaluation of machine learning-based methods for detection of phishing sites", Advances in Neuro-Information Processing, pp. 539-546, Springer, 2009.

[43] H. Patel, and J. Sarvakar, "Analysis of data mining algorithm in intrusion detection", International Journal of Emerging Technology and Advanced Engineering (IJETAE), 1(1), 2011.

[44] C.-F Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review", Expert Systems with Applications, 36(10), 11994-12000, 2009.

APPENDIX I. THE ORIGINAL FEATURE SET CONSISTS OF 58 HYBRID FEATURES

Webpage Content Features					
Index	***Feature***	***Type***	***Index***	***Feature***	***Type***
F1	Number of Scripting.FileSystemObj	Continuous	F24	Number <input> in java scripts	Continuous
F2	Number of Excel.Application	Continuous	F25	JavaScript scripts length	Continuous
F3	Presence of WScript.shell	Discrete	F26	Number of functions' calls in java scripts	Continuous
F4	Presence of Adodb.Stream	Discrete	F27	Number of script lines in java scripts	Continuous
	Presence of Microsoft.XMLDOM	Discrete	F28	Script line length in java scripts	Continuous
	Number of <embed>	Continuous	F29	Existence of long variable names in java scripts	Discrete
	Number of <applet>	Continuous	F30	Existence of long function names in java scripts	Discrete
	Number of Word.Application	Continuous	F31	Number of fromCharCode()	Continuous
	link length in <embed>	Continuous	F32	Number attachEvent()	Continuous
	Number of <iframe>	Continuous	F33	Number of eval()	Continuous
	Number of <frame>	Continuous	F34	Number of escap()	Continuous
	Out-of-place tags	Discrete	F35	Number of dispacthEvent()	Continuous
	Number of <form>	Continuous	F36	Number of SetTimeout()	Continuous
	Number <input>	Continuous	F37	Number of exec()	Continuous
	Number of MSXML2.XMLHTTP	Continuous	F38	Number of pop()	Continuous
	Frequent <head>, <title>, <body>	Discrete	F39	Number of replaceNode()	Continuous
	<meta index.php?Sp1=>	Discrete	F40	Number of onerror()	Continuous
	"Codebase" attribute in <object>	Discrete	F41	Number of onload()	Continuous
	"Codebase" attribute in <applet>	Discrete	F42	Number of onunload()	Continuous
	"href" attribute of <link>	Discrete	F43	Number of <script>	Continuous
	Number of void links in <form>	Continuous	F44	frequent<div onClick=window.open()">	Discrete
	Number of out links	Continuous	F47	Number of onerror()in javascripts	Continuous
	Number of <form> in java scripts	Continuous	F48	Number of SetInterval()	Continuous
URL Features					
Index	***Feature***	***Type***	***Index***	***Feature***	***Type***
	Multiple TLD	Discrete	F54	Typos in Base name	Discrete
	Brandname in hostname	Discrete	F55	Long domain name	Discrete
	Special symbols in URL	Discrete	F56	Misleading subdomain	Discrete
	Coded URL	Discrete	F57	Number of dots in URL	Continuous
	IP address instead of domain name	Discrete	F58	Path domain length	Continuous