Phishing classification models: issues and perspectives

Hiba Zuhair*

Department of Systems Engineering,
College of Information Engineering,
Al-Nahrain University,
Baghdad, Iraq
Email: hiba.zuhair.pcs2013@gmail.com
*Corresponding author

Ali Selamat

Centre for Information and Communication Technologies,
Software Engineering Department,
Faculty of Computing,
Universiti Teknologi Malaysia (UTM),
Johor, Malaysia
Email: aselamat@utm.my

Abstract: The never-ending threats of phishing to the cyberspace motivate researchers to develop more proficient phishing classification models for the survival of cyber-security with safe web services. However, these models remain variable in their reaction and incompetent in their performance against novel phishes at the real-time of application. This attributed to their partial or full deficiency of inductive factors including a rich set of decisive features, actively learned and adaptive machine learning based classification model. Upon this issue, our paper revisits the current machine learning-based phishing classification models and analyses their performance qualitatively and quantitatively across three benchmarking data sets. Empirical results and observations emphasised the causality between the models' limitations and their lack of inductive factors. Accordingly, future outlooks are recommended as a navigating taxonomy to serve the researchers at developing their upcoming achievements in both academia and industry.

Keywords: novel phish; phishing classification model; machine learning; feature-based classifier; FBC; ensemble feature-based classifier; EFBC; inductive factors; active learning; adaptive model.

Reference to this paper should be made as follows: Zuhair, H. and Selamat, A. (2019) 'Phishing classification models: issues and perspectives', *Int. J. Digital Enterprise Technology*, Vol. 1, No. 3, pp.219–240.

Biographical notes: Hiba Zuhair currently works as a Senior Lecturer and Researcher in Dept. of Systems Engineering at College of Information Engineering, Al-Nahrain University, Baghdad, Iraq. Prior to this, she is awarded her PhD from Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia with a high distinction as the Best Postgraduate Student Award. Before, she received her MSc and BSc with a high distinction from Dept. of Computer Science at College of Science, Al-Nahrain University,

Baghdad, Iraq. Her recent research interests and publications are cyber-crimes, intrusion detection systems, ethical hacking, digital forensics, big data science and analytics, smart cities, machine learning and deep learning as well as other fields in computer networks and security.

Ali Selamat is currently the Chief Information Officer and Director of the Centre for Information and Communication Technologies, Universiti Teknologi Malaysia (UTM), Malaysia. He is also a Professor in Software Engineering Department at the Faculty of Computing, UTM. He is nominated as the Chair of IEEE Computer Society Malaysia since 2014. He is a co-Editor-in-Chief of *International Journal of Software Engineering and Technology (IJSET)*, and a member of the journal editorial boards*: Knowledge Based Systems*, Elsevier, *International Journal of Information and Database Systems*, Inderscience Publications, and *Vietnam Journal of Computer Science*, Springer Verlag. His research interests and publications include software engineering, software process improvement, software agents, web engineering, information retrievals, pattern recognition, genetic algorithms, neural networks and soft computing, computational collective intelligence.

This paper is a revised and expanded version of a paper entitled 'Phishing classification models: issues and perspectives' presented at 2017 IEEE Conference on Open Systems (ICOS), Miri, Sarawak, Malaysia, 13–14 November 2017.

1 Introduction

Motivating by more illegal gains, phishers have threaten the users' digital identity and the industries' reputation on the cyberspace by evolving phish web pages (Khonji et al., 2013). To mitigate phishing threats, many efforts have been made by researchers in both academia and industry for obtaining effective anti-phishing schemes (Khonji et al., 2013; Zeydan et al., 2014b). Almost all anti-phishing schemes have adopted client-side filtering to detect phish web pages towards more safe online communication (Zeydan et al., 2014b). Amongst them, are machine learning-based phishing classification models that assisted by a set of features and machine learning algorithms to tackle phish web pages effectively (Khonji et al., 2013; Zeydan et al., 2014b; Shahriar and Zulkernine, 2012; Whittaker et al., 2010). However, late phish tackle, somewhat faulty classification, variable and adverse performance has been reported along with long elapsed time, complex computations, and heavy use of external resources (Zeydan et al., 2014a; Wardman et al., 2014). Such heavy-weight and static machine learning based anti-phishing models still provide good opportunities to the phishers with to evolve new phishing pattern (novel phishes) that exploit advanced deceptions and tricks to bypass anti-phishing schemes. Then again, more thefts to the users' identities are caused besides more monetary losses to the enterprises and disastrous consequences to the cyber-security (Zeydan et al., 2014a; Wardman et al., 2014; Abbasi and Chen, 2009; Islam and Abawajy, 2013).

As time progresses, researchers have attempted to perform more proficient mitigation of phishing threats for ideal cyber-security and optimal users' safety. To assist in achieving this goal, this paper critically studies the most salient machine learning-based phishing classification models in an analytical context. In the analytical context, the most salient models are revisited, characterised according to their merits, and categorised according to their classifiers' design like single feature-based classifier (FBC), FBCs assisted by feature selection mechanism (FSFBC), and ensemble FBCs (EFBC). This is followed by an empirical analysis of their performance in terms of inductive factors including rich set of features, handling big web data, active learning of the constructed FBC model, and the adaptable modelling for real-time detection (Kumar et al., 2010; Bishop, 2006). An empirical workflow is devoted to extract features from the fetched batches of web pages (datasets), formulate the required feature space, and divide the feature space into training data and testing data for learning and testing the classification model. Based on the empirical results, classification performance is evaluated to justify the causality between the models' performance and their deficiency to the inductive factors. Intellectual and empirical observations are discussed in-depth to restate what research facets need to boost in the future towards obtaining an efficient phishing classification model. Lookouts like exploring new features, chronological aggregation of web pages as evolving dataset, designing an adaptive assembly and maintaining a deep learning mechanism; are highly recommended as the best possible solution that can be undertaken for real-time phishing detection.

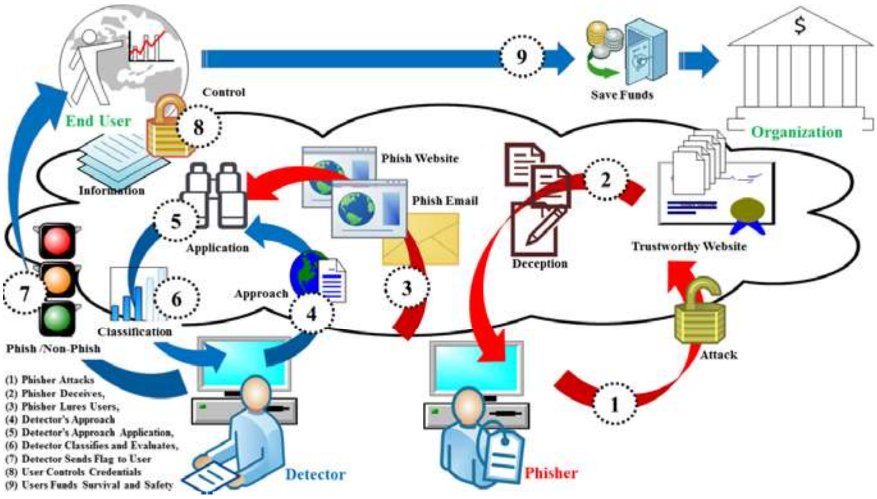
To demonstrate all the aforesaid issues, the rest of this paper is organised as follows: Section 2 introduces phishing activities along with the types of existing anti-phishing schemes. Whereas, Section 3 presents the preliminaries of FBC and the state-of-the-art machine learning algorithms that highly applied in anti-phishing domain. On this topic, Section 4 surveys the recently published works and appraises them critically. Consequently, Section 5 analyses the revisited works empirically their inductive factors. Further, Section 6 restates the main issue of inductive factors that need to boost and recommends the research facets that still open to further study. Altogether, are concluded in Section 7 along with several remarks.

2 Phishing and anti-phishing

Although, the web is a huge communication channel between users and enterprises which provides many services and applications including e-business, online banking, and retails, etc. It causes many losses to the users and enterprises annually due to the insecure web-based applications and the vulnerable web services that put both users and industries at the risk of credentials' theft, malware distribution, industrial espionage, and then big financial losses. Such consequences often occur when phishers imitate the look of the trustworthy web pages of publically known organisations to mislead victim users by inserting spoofed links and using social engineering technologies. As illustrated in Figure 1, victims catch the bait and submit their own credentials via online transactions. Then, the phishers acquire these credentials for their own illegal gains (Khonji et al., 2013; Zeydan et al., 2014b).

To mitigate phishing threats, many anti-phishing schemes have been developed by using either a whitelist of legitimate web pages or a blacklist of prominent phish web pages. Also, some anti-phishing schemes have used heuristics and algorithms for phishing characterisation and phishing classification respectively (Shahriar and Zulkernine, 2012; Whittaker et al., 2010; Zeydan et al., 2014a). Even though some anti-phishing schemes have performed well at phishing detection, they have been circumvented by phishers who usually advance their deceptions and spread their activities day by day (Zeydan et al., 2014a). Amongst the defeated anti-phishing schemes, are the machine learning-based classification models which still perform sub-optimally at detecting novel phish web pages as their competitors did (Whittaker et al., 2010; Zeydan et al., 2014a; Wardman et al., 2014; Abbasi and Chen, 2009; Islam and Abawajy, 2013). To address their defeatism against novel phishes, this paper studies the recently published achievements in the domain of machine learning-based anti-phishing models and it states what open problems need to solve.

Figure 1 Phishing and anti-phishing as adopted in Zuhair et al. (2016b) (see online version for colours)



3 Applied machine learning algorithms

Due to their classification purposes, machine learning algorithms still play prominent roles in the development of many research domains including anti-phishing. In anti-phishing domain, machine learning algorithms are usually used to construct a FBC that can identify phishes and legitimate web pages on a batch of web pages (Zeydan et al., 2014a; Wardman et al., 2014; Abbasi and Chen, 2009; Islam and Abawajy, 2013). As yet, such algorithms vary in their classification performance across increasing web streams due to their diverse specifics, decision settings, and the induction factors (Zeydan et al., 2014a; Wardman et al., 2014; Abbasi and Chen, 2009; Islam and Abawajy, 2013). For example, Naïve Bayes (NB), logistic regression (LR), sequential minimal optimisation (SMO), support vector machine (SVM), and transductive SVM (TSVM) as they are described briefly in Table 1. Typically, the existing phishing classification models are constructed as single FBC model, or ensemble FBC (EFBC) model, or single FBC model assisted by feature selection method (FSFBC). In Figure 2, a FBC maps the input feature vector to the output classes by attributing the input feature vector $V = (v_1, \dots, v_n)$ and inducts its relevance to either phish or not phish classes with $Y = f(V, \gamma)$. All input feature vectors that extracted from the m-dimensional training dataset $(V_1, V_2, \dots,$

V_m) are induced in the training phase to classify the incoming instance V_{new} in the testing phase into either phish or legitimate label (Kumar et al., 2010; Bishop, 2006; Nguyen and Armitage, 2008).

Table 1 Examples of machine learning algorithms adopted in Kumar et al. (2010), Bishop (2006), Nguyen and Armitage (2008), Galar et al. (2012), and Toolan and Carthy (2010)

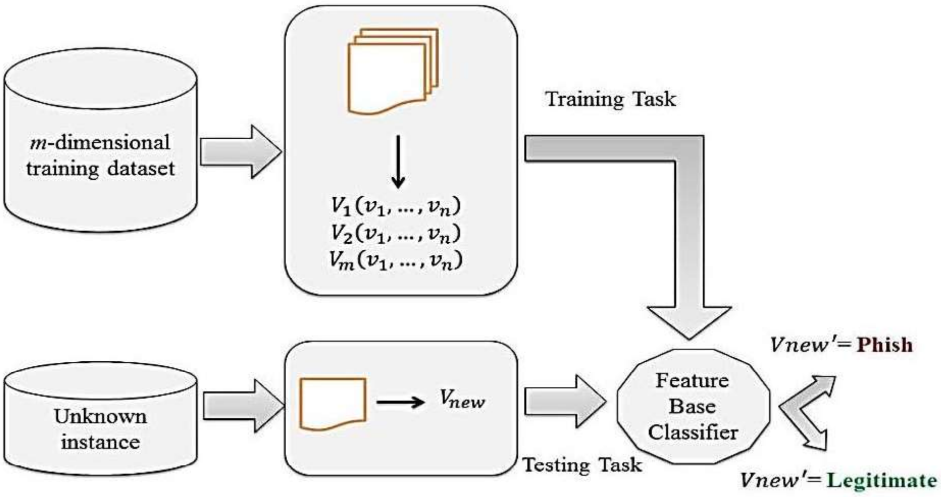
Algorithm	*Description*
C4.5	It depends on decision tree hypothesis that traces the node paths, their branches until terminating leafs.
Decision tree (DT)	It models the unknown instances as nods in a rooted tree, and the feature values as edges. Induction starts from the root node approaching to leaf and passing through edges. Test is applied at each node to re-order feature values which determine the next edge to go. Final decision found at the end-up leaf node.
NB	A probabilistic judgment done conditionally with independent attributes of all instances belonging to a given class: $P(C\|X) = P\left(C\|x_1, \dots, x_n\right) = \frac{P(C)P\left(x_1, \dots, x_n\|C\right)}{P\left(x_1, \dots, x_n\right)}$ (1) Where X is an instance with a vector of n features (x_1, ..., x_n), C is the class label that the classifier seeks for.
SVM	A separating hyper-plane maximises the margins between closest points of two classes to estimate the induction function: $\min \frac{1}{2} w^T w + C \sum_i \xi_i$ (2) That subjects to: $y_i\left(\left(w^T \cdot x_i\right) + b\right) \geq 1 - \xi_i, \xi \geq 0, i = 1, 2, \dots, m$ (3) $\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} y_i y_j \alpha_i \alpha_j K\left(x_i, x_j\right)$ (4) Which is subject to: $0 \leq \alpha_i \leq C, i = 1, 2, \dots, m$ and $\sum_{i=1} \alpha_i, y_i = 0$ (5) Where: x_i is m-dimensional data vector $x_i \in R^m$ with samples belong to either one of two classes labelled as $y \in \{-1, +1\}$ that it is separated by a hyper-plane of $(w \cdot x) + b = 0$, α_i denotes the lagrange multipliers for each vector in the training dataset.
TSVM	It separates positive and negative samples of training dataset with a maximal margin of SVM hyper-plane, such that it minimises over $\left(y_1^*, \dots, y_k^*, w, b, \xi_1, \dots, \xi_n, \xi_1^*, \dots, \xi_k^*\right)$ into $: \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_1 + C' \sum_{j=1}^{k} \xi_j^*$ (6) Which subjects to: $\forall_{i=1}^{n} : y_i\left[w v_i + b\right] \geq 1 - \xi_i$ (7) $\forall_{j=1}^{k} : y_i\left[w v_i^* + b\right] \geq 1 - \xi_j^*, \forall_{i=1}^{n} : \xi_j^* \geq 0$ (8)

Table 1 Examples of machine learning algorithms adopted in Kumar et al. (2010), Bishop (2006), Nguyen and Armitage (2008), Galar et al. (2012), and Toolan and Carthy (2010) (continued)

Algorithm	*Description*
LR	Use probabilistic induction that evaluates relationship between a categorical dependent variable and a continuous independent variable(s): $\pi(x) = \frac{e\left(\beta_0 + \beta_1 x\right)}{e^{(\beta_0 + \beta_1 x)} + 1} = \frac{1}{e^{-(\beta_0 + \beta_1 x)} + 1}$ (9) $g(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_i x$ (10) $\frac{\pi(x)}{1 - \pi(x)} = e^{(\beta_0 + \beta_1 x)}$ (11) Where: $g(x)$ is the logistic function of a given predictor X, ln and, $\pi(x)$ denote natural logarithm and case probability, β_0 and β_1 denote criterion of X, and $\beta_1 x$ is the regression coefficient.
Random forests (RF)	Forest constructed for randomly selected set of instances on training dataset. It comprises of many combined tree predictors that are distributed similarly. Each tree predictor is learned on feature vector belongs to independent sample.
K-nearest neighbour (K-NN)	Nonparametric classifier estimates class conditional densities by using a discriminant function $g_i(x) = P\left(x \mid C_i\right) P\left(C_i\right)$ (12) $P\left(x \mid C_i\right) = \frac{k_i}{\left(N_i V^k(x)\right)}$ (13) Where $P(x\|C_i)$, k_i and $V^k(x)$ are the class conditional densities, the number of nearest neighbors that belong to C_i, and the volume of n-dimensional hyper-sphere centred at x with radius of $r = \\|x - x_k\\|$ and x_k is the nearest observation to x.
SMO	It solves the optimisation problem caused during classification iteratively and analytically: $\max_\alpha \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K\left(x_i, x_j\right) \alpha_i, \alpha_j$ (14) Where $0 \leq \alpha_i \leq C$, for $i = 1, 2, \ldots, n$ and $\Sigma_{i=1}^{n} y_i \alpha_i = 0$. C is the classifier's hyper-sphere, $K(x_i, x_j)$ refers to the kernel function provided by user, and α_i is the Lagrange multiplier.
Neural network (NN)	It updates the individual weights of different inputs during the training task according to the examples of network receives to reduce the error rates: $f(x) = g\left[\sum_i v_i g\left(\sum_j w_{ij} x_j + b_i + b_0\right)\right]$ (15) where x, v_i, g, w_{ij} and $b_{i,\,o}$ are the input vector, the weight of output neuron, the activation function, the weight of hidden neuron and the bias respectively.

Unlike FBC, EFBC promotes an assembly classifier which integrates multiple machine learning algorithms of different induction settings (Nguyen and Armitage, 2008; Galar et al., 2012). It learns the input feature vectors by its entire constituent machine learning algorithms that may vary in their induction boundaries and decision outcomes (Nguyen and Armitage, 2008; Galar et al., 2012). Whereas, FSFBC maintains learning task with a feature selection method to predict phishness in a high-dimensional feature space (Toolan and Carthy, 2010). In practice, the final judgement of FBC depends on classifying the features extracted from the input web page to predict its class as either phish web page or legitimate. While, EFBC leverages the average of all its constituent classifiers' predictions on a fetched web page to set its final decision on its phishness or its legitimacy (Nguyen and Armitage, 2008; Galar et al., 2012; Toolan and Carthy, 2010; Xiang, 2013). On the other hand, FSFBC analyses the most informative features of the fetched web page with the aid of features selection, or features ranking, or features weighting method. Then, it classifies the web page by applying a machine learning algorithm assisted by the selective features (Toolan and Carthy, 2010).

Figure 2 Work flow of FBC model (see online version for colours)



Source: Kumar et al. (2010), Bishop (2006), Nguyen and Armitage (2008), Galar et al. (2012), and Toolan and Carthy (2010)

4 Machine learning-based phishing classification models

In anti-phishing domain, researchers have attempted to perform their achievements at nominal computational cost by using machine learning algorithms. As depicted in Table 2, this section provides a bird's-eye view on the most salient machine learning-based phishing classification models along with their regressing frontiers.

As depicted in Table 2, among the most salient EFBCs in anti-phishing domain was that the developed model of (CANTINA$^+$) (Xiang, 2013). CANTINA$^+$ was constructed with many machine learning algorithms such as NB, SVM, and LR etc. Also, it was learned with 15 textual and structural features that were derived from the fetched web page and it's URL along with its online features. CANTINA$^+$ devoted to perform accurate classification on many phish exploits. It reported 92% of true positive rate (TPR), and 1.4% of false positive rate (FPR). The examined exploits included redirecting web page, login form handlers, and web pages hosting in English. However, CANTINA^{+} encountered a trade-off in leveraging up-to date phish web pages due to the use of limited feature space to English textual features as well as re-learning on defaults settings (Xiang, 2013). While, the authors of Gowtham and Krishnamurthi (2014a, 2014b) leveraged 17 features to examine login form phish web pages via a FBC by using SVM classifier. Their model achieved a rationale performance with (99.6%) of TPR and (0.44%) of FPR. However, it was computationally intensive and time-consuming due to the use of external resources and less adaptive to present training datasets.

Table 2 Categorisation of machine learning-based phishing classification models

Model category	*Year*	*Related work*	*Machine learning algorithms*	*Related limitations*
FBC	2014	Gowtham and Krishnamurthi (2014a, 2014b)	SVM	• Generic features • Limited size of datasets
	2014	Marchal et al. (2014a, 2014b)	k-NN	• Limited maintain to web page exploits • Use of external resources for data query • No features selection mechanism • Inactive learner • Not adaptive model
FSFBC	2014	Zhang et al. (2014)	SMO, LR, RF, NB	• Generic features • Imbalanced datasets • Irrelevance and redundancy problems • Features heterogeneity • Inactive learner • Not adaptive model
EFBC	2011, 2013	Xiang (2013)	NB, SVM, LR	• Generic features
	2014	Mohammad et al. (2014)	SVM, RF, JRip	• Imbalanced datasets
	2015	Marchal (2015)	SVM, C4.5, RF, JRip	• No selected set of features • External sources like search engines and blacklists were used for data query. • Inactive learner • In-adaptive classification model

Notes: SVM = support vector machine; LR = logistic regression; BN = Bayesian; DT, C4.5, and JRip are types of decision tree classifier; RF = random forest; NN = neural network; SMO = sequential minimal optimisation, NB = Naïve Bayes. FBC: feature-based classifier; FSFBC: FBC assisted by features selection method; EFBC: ensemble FBC.

To degrade the trade-offs of extracting features and to upgrade the classification performance with less false classifications, some researchers like Zhang et al., identified phishing on (2,878) Chinese e-business websites via phishing Chinese web page detection model. They selected 15 language independent features by weighting and ranking them with Chi-Squared (χ^2) statistic criterion. The selected set of features were utilised exclusively to identify Chinese web pages. Four machine learning algorithms like SMO, LR, NB, and random forests (RF) were applied individually in an FBC to learn the selected set of features on the collected dataset. Even Experimentally, their model performed (95.83%) of accuracy rate on Chinese e-business web pages solely, even though; it was not reliable to classify other types of phish web page due to its exclusive set of features and datasets (Zhang et al., 2014).

As time progresses, EFBC were developed such as that adopted in Mohammad et al. (2014) to learn 12 URL features by applying multiple machine learning algorithms like SVM and RF in addition to C4.5, and JRip as types of decision tree algorithm Such EFBC achieved (94.91%) and (1.44%) as classification accuracy and classification faults. In spite of using big training and testing datasets, the used dataset were imbalanced in their classes of phishing and not phishing as well as including e-commerce websites exclusively.

On the other hand, an FBC was developed by authors of Marchal et al. (2014a, 2014b) to catch phishing in e-commerce, login form, and English and French web pages by using 17 generic features with NN classifier. Even though, this achievement yielded up to 94.07% accuracy rates, high misclassification rates were reported. However, these models scarcely detected novel phish websites due to learning big datasets and extracting many features to characterise phishing. Whereas, another version of this phishing classification model was optimised in Marchal (2015) with the aid of an EFBC construction and functionality. The optimised version attained active learning by deploying a hybrid set of 212 typical features. Thus, it performed an effective classification on (96,018) web pages aggregated during (2012–2015). However, deploying typical (generic) features on large and imbalanced datasets revealed notable misclassification rates versus novel phishes. Moreover, long execution time, complex computations are encountered due to data query from several external resources like GoogleTrends and YahooClues. Inactive learning on the up-to-date data also observed and it caused limited adaptation in real-time practice.

As a matter of fact, the web contains millions of the strongly associated web pages belonging to many types of cyber-attacks like spam, ham, scam, malware as well as phishing attacks (Uzun et al., 2013). Such web pages of cyber-attacks may share millions of features and embedded links and objects generically (Zuhair et al., 2016c). Additionally, the web involves many exploits of web pages including login forms, pharming, homepages, and ending up pages along with the web page hosting language like English, French, Chinese and others (Zuhair et al., 2016a). As such, web page processing and extraction of heterogeneous features from these web page exploitations were somewhat insignificant in the FBCs adopted in Gowtham and Krishnamurthi (2014a, 2014b) and Marchal et al. (2014a, 2014b). Such models required more extensive computation to bare holistic phishing characterisation and robust mechanisms to select the most informative features and feature subsets. Therefore, altogether produced divergent accuracies of classification with false detections, more and complex computations, memory and processing footprints, and long execution time.

Even though, the models in Zhang et al. (2014) and Mohammad et al. (2014) were assisted by feature selection methods, they encountered several shortcomings attributed to the feature selection strategies that they employed. These selection methods relied on ranking the best features by weighting them individually with respect to their interdependencies amongst the others (Zuhair et al.,2015a, 2015b, 2016b). Therefore, they fall short in leveraging:

1 the heterogeneity of features' values (categorical/continuous/mixed values) which could vary among attack classes across the training and the testing datasets

2 the irrelevant and the redundant features involving in the selected subsets of features which is caused by the overlapping of generic features among the examined classes (phish and non-phish) (Zuhair et al.,2015a, 2015b, 2016b). Altogether, cause inefficient phishing classification particularly on a high-dimensional training and testing dataset with class imbalance problem.

Whilst, the EFBC models such as those developed in Xiang (2013), Mohammad et al. (2014) and Marchal (2015) outperformed their competitors in phishing classification, they still unaware of new phish patterns (novel phishes) that have been evolved periodically by the phishers. This is attributed to the deficiency of:

1 actively learnt classifiers that supposed to be able to expect the future error and pick up the best batch of web pages which will reduce that error iteratively (Galar et al., 2012; Huang et al., 2015)

2 adaptive modelling along with updating mechanism which inspects any change (an unknown pattern) in the fetched web stream, identifies the new features, and adjusts the default induction function for the future learning of the classifier over the time (Huang et al., 2015; Tsai et al., 2009; Shabtai et al., 2009).

Recently, several published works have studied the performance of some existing anti-phishing schemes (Whittaker et al., 2010; Abbasi and Chen, 2009; Vink and de Haan, 2015; Abu-Nimeh et al., 2007; Miyamoto et al., 2008) however, they rarely analysed their lack of induction factors and their disastrous consequences on novel phish web page classification in the real-time experience. Therefore, a necessary depth of analysis in the term of inductive factors is required to assert whether the aforementioned machine learning based classification model do or do not their best with maximal accuracy, minimal false detections, zero misclassifications, simple computations, and low execution time as well as the least amount of external resources in the realistic mode. To do so, this paper pays much attention to study phishing classification performance in terms of the deficiency of inductive factors quantitatively and qualitatively. Analysis, observations and standpoints will provide researchers the right direction on how to undertake their further achievements which is the main goal of this paper. To attain this goal, the next section focuses on the empirical analysis of the revisited models which is followed by an extensive discussion and justification of the observations in the following sections.

5 Empirical analysis

This section analyses the aforementioned phishing classification models in a practical context to assess their classification performances and discuss their related limitations. The empirical analysis is pursued with three recently published, publically available and highly used datasets for benchmarking in anti-phishing domain. Consequently, the findings of the computerised simulation have been evaluated with a set of standard metrics that usually used for validation in anti-phishing domain. Altogether, the benchmarking datasets, the evaluation metrics, the empirical workflow and results will be discussed in the following sub-sections.

5.1 Benchmarking datasets and evaluation criteria

As described in Table 3, three benchmarking datasets with low and high range of phish and legitimate samples are used. All the datasets differ in terms of their size, the imbalance and/or balance of phishing class distribution, the number of phish web pages. Also, they differ in the number of legitimate web pages, and the source of datasets (i.e., data archives). In addition, they vary in terms of web page's functionalities like homepage, login form page, e-business web page, retail, etc. Moreover, they encompass the web page's hosting languages such as English, Chinese, French, Italian, German, Spanish, etc. All benchmarking datasets are retrieved from three recently published works in which they were aggregated periodically during 2010 and 2015. For instance, the *dataset 1* used in Shahriar and Zulkernine (2012) included 52 phishes exploited login forms and they are exclusively belonging to a broad range of brands and industries. Whereas, *dataset 2* used in Zhang et al. (2014) contains (2,878) samples of e-business Chinese web pages to as phish and legitimate samples. *dataset 3* consists of (96,018) web pages and it is a large scale dataset retrieved from Marchal et al. (2014a, 2014b) and Marchal (2015). Also, it varies in the web page functionalities, hosting languages and targeting industries. Moreover, it was used to detect DNS poisoning and login form handler phishes hosting different natural languages.

Table 3 Description of the benchmarking datasets

Merits	*dataset 1*	*dataset 2*	*dataset 3*
Size	52	2878	96,018
Phishes	36	1382	48009
Legitimates	16	1496	48009
Data archives	PhishTank/Alexa	Chinese e-business	PhishTank/DMOZ
Training split (2/3)nd	34	1918	64012
Testing split (1/3)rd	18	960	32006
Data source	Shahriar and Zulkernine (2012)	Zhang et al. (2014)	Marchal et al. (2014a, 2014b), Marchal (2015)
Aggregation time	25–31/7/2010	2014	2012–2015
Web page functionality	Login forms/bank web pages/e-business web pages/retailing web pages	E-business web pages	E-business/web pages/homepages/login forms/social networking/web pages
Hosting language	English/French/ German	Chinese	English/French/German/ Italian/Spanish etc.

For performance evaluation, each benchmarking dataset is split up into $\frac{2}{3}^{nd}$ and $\frac{1}{3}^{rd}$ splits for both training and testing task, respectively. Throughout experiments, each classification model is tested and evaluated across the splits of every benchmarking dataset individually. Benchmark results are averaged to estimate the overall performance outcomes and overhead of each tested classification model. To do so, typical evaluation metrics are used including: TPR that indicates the rate of correctly classified phish samples, and FPR refers to mistakenly classified legitimate samples as phishes, whereas; false negative rate (FNR) refers to mistakenly labelled phish samples as legitimate ones which implies misclassification cost [25]. Furthermore, Elapsed Time is used to compute the amount of time spent by the tested classification model from its start-up to its ending-up. Elapsed time quantifies how long the tested classification model takes to detect phishing on a batch of web page stream in practice (Huang et al., 2015).

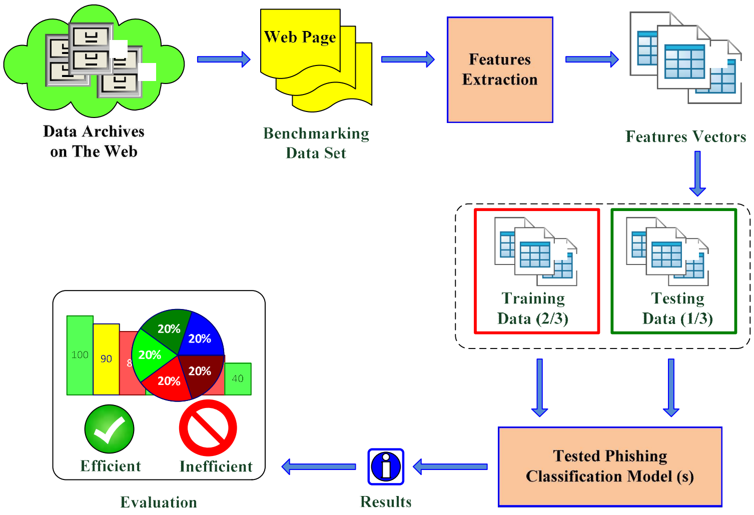
5.2 Empirical workflow

As illustrated in Figure 3, the empirical workflow was pursued via three steps: features extraction, implementing the classification model, and then evaluation. In features extraction step, the source code and URLs of the fetched web pages were parsed to extract a three categories of features that were discussed in our previously published papers: URL features, cross site scripting (XSS) features, and HTML features. URL features such as certain patterns, terms, irregularities, and indicators are widely used by potential phishes to impersonate legitimate web pages (Zuhair et al., 2016a, 2016c). The XSS features are suspicious java scripts injected in the source code of the web page by phishers for malware damage. HTML features are the embedded objects and attributes structured by the tree of document object model (DOM) (Zuhair et al., 2016a, 2016c). Thus, the previously revisited classification models are tested on different sets of generic and new features. Furthermore, the used sets of features -as they have been employed by the their corresponding works- are of either a hybrid set, or URL features set, or web page content features set.

Accordingly, each dataset was formulated into feature vectors to be manipulated by the tested classification model. Then, the formulated dataset (the group of extracted feature vectors) was split up into $\frac{2^{nd}}{3}$ and $\frac{1^{rd}}{3}$ splits as training and testing dataset to learn and test the chosen classification model (Abbasi and Chen, 2009; Islam and Abawajy, 2013; Xiang, 2013). These splits were fed to the classification model and to obtain the detection results, see Figure 3. In the evaluation step, the obtained results of each classification model that was tested on every benchmarking dataset individually were averaged to estimate the overall performance outcomes and overhead.

It is worthy to mention that evaluation was calculated in terms of average TPR, average FPR and average FNR, respectively. Also, 27 computerised simulations (27 repetitions of the conducted experiment across three benchmarking datasets) were implemented for the comparable classification models and benchmarking datasets. Their implementation was carried out via a highly used tool for data mining that is ‘WEKA 3.5.7 – Waikato environment for knowledge analysis’ which is developed by some researchers at the University of Waikato.

Figure 3 Workflow of empirical analysis (see online version for colours)



5.3 Results and discussions

The obtained classification performance outcomes as they plotted in Figure 4 demonstrate the performance and manifest the detection capability of the tested phishing classification models across all the benchmarking datasets.

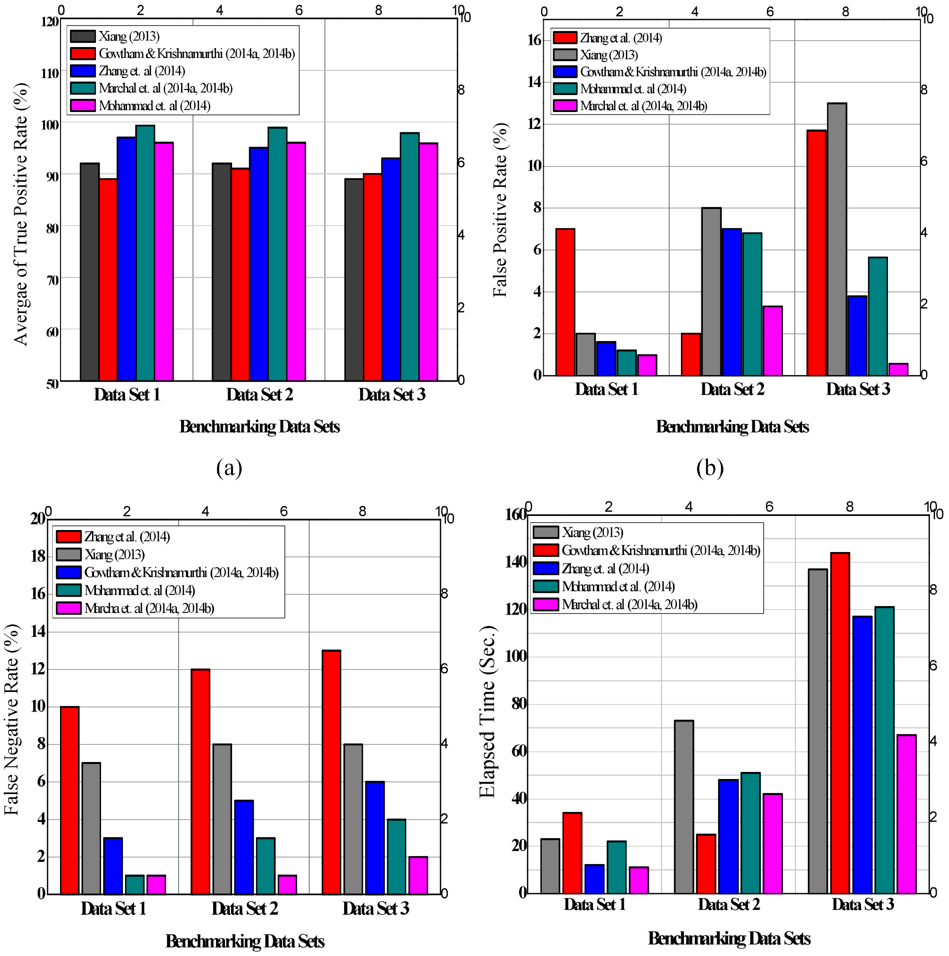
As shown in Figure 4(a), the phishing classification adopted in Marchal et al. (2014a, 2014b) and Marchal (2015) showed that the competitive models could achieve high levels of TPR on all benchmarking datasets. Amongst them, the tested model adopted by Marchal et al. (2014a, 2014b) and Marchal (2015) achieved the best scores of TPRs, FPRs and FNRs among its competitors across benchmarking datasets. This is attributed to its related feature set which was a large scale and a hybrid set of feature. It consisted of 212 generic features including those previously used by its competitors and/or other anti-phishing schemes. In practice, such set of features provided an optimal configuration of features' variety and quantity. Therefore, the tested model of Marchal et al. (2014a, 2014b) and Marchal (2015) could predict phishness in the chronologically collected datasets: *datasets 1*, *2* and *3*. Whereas, the other classification models showed less performance (lower TPRs, higher FPRs and FNRs). This is attributed to those models deployed relatively less number of features along with newer features than those adopted by model in Marchal et al. (2014a, 2014b) and Marchal (2015). As previously depicted in Table 4, almost comparable models deployed hybrid sets of features rather than features of the same type (i.e., a combination of URL, HTML, and XSS features). Except that set of feature used by the models developed by Zhang et al. (2014) and Mohammad et al. (2014). It was of a unique type either HTML or URL features. Hence, the variation in empirical results manifest how the hybrid set of features is the most significant factor of the best and holistic phishing induction.

Table 4 Issues of the tested models

Work *Merits*	*Xiang (2013)*	*Gowtham and Krishnamurthi (2014a, 2014b)*	*Zhang et al. (2014)*	*Mohammad et al. (2014)*	*Marchal et al. (2014a, 2014b), Marchal (2015)*
Machine learning algorithm(s)	SVM, LR, BN, DT, Adaboost	SVM	SMO, LR, RF, NB	NN	SVM, C4.5, RF, JRip
Classifier's design	Ensemble	Single	Single	Ensemble	Single
Number of features	15	17	15	17	212
Feature category	Hybrid features	Hybrid features	URL features	Web page content features	Hybrid features
Generic features	7	14	10	17	200
New features	8	3	5	None	None
Feature selection mechanism	Not	Not	χ^2	CFS, IG, χ^2	Not
Active learning	Not	Not	Not	Not	Active
Adaptive model	Not	Not	Not	Not	Not
Exploited web pages	Redirecting, login form, e-business, social networking, English	E-commerce login forms, redirecting, English	E-business, Chinese	E-commerce, login forms, redirecting, English, French	E-commerce, English, French, German, Italian, Spanish, Portuguese
Avg. TPR (%)	91%	90%	96%	98.6%	97.53%
Avg. FPR (%)	7.6%	6.3%	5.9%	4.48%	2.85%
Avg. FNR (%)	6.7%	11.6%	4.6%	3.4%	1.3%
Avg. elapsed time (sec.)	77 sec.	76.8 sec.	59 sec.	64.6 sec.	43 sec.

Furthermore, the chart plotted in Figure 4(a), pays a close attention onto the discriminating power of the deployed features. It is found that not all the employed sets of features were contributing enough to characterise all types of phish web pages and their exploits holistically. Indeed, the set of features should be informative to induct phishing class among all other remaining classes like legitimate and suspicious. More precisely, the set of features must identify phishing attacks decisively among the other cyber-attacks that might share the same features. It is observed from Figures 4(b) and 4(c) that the high rates of FPR and FNR demonstrated the need to use a set of decisive features which could crucially classify phishes on small and/or large data streams. To do so, a robust feature selection mechanism is required which chooses the best subset of features frequently to induct phishing accurately. Subsequently, using the best selective subset of features could solve the problems of class imbalance, size and variety of data, evolution of web page streams.

Figure 4 Performance outcomes of the empirical analysis, (a) avg. true positive rate (b) avg. of false positive rate (c) avg. of false negative rate (d) elapsed time (see online version for colours)



(a) (b)

(c) (d)

In spite of using a gradual scale of the benchmarking datasets along with their diversity in aggregation time that offered a suitable test-bed to assess the studied phishing classification models chronologically. Yet, the learning the tested classification model across the past and the present datasets was passive and questionable, see Figures 4(a), 4(b) and 4(c). This is due to the variable time of data aggregation as well as their variable and scalable size. For example, *dataset 1* was the smallest in size and oldest in aging among other datasets. Whilst, *dataset 2* was bigger in size and younger in aggregation time than *dataset 1* but it convolved exploits of Chinese e-business websites of URL features exclusively. Whereas, *dataset 3* was the biggest in size, the most present in aging, the most balanced in the distribution of phish/not phish classes, and the most variant in web page functionalities and hosting languages among its competitors. Therefore, almost all tested models fall short to classify phishes on *dataset 3* accurately.

That indicates their inability to identify the phish web pages (w) which might be emerged at time ($T + \Omega$) on a dataset (W) which might be fetched at time (T). As yet, they need a period of time (Ω) to learn (W) and to set decision boundaries for the newly fetched and unknown web page on the incoming web page stream. In this concern, the escalating classification accuracies in Figure 4(a) implied that the emerged phish web page (w) might be short-living and taken down by its phisher during time (Ω). More precisely, the results plotted in Figure 4(a) pointed out that the time of emerging (Ω) was a long time horizon that could mislead classification accuracies of the tested models against novel phishes. That, in turn, makes the sense to validate the factors of big and evolving data as well as the active learning straightforwardly crosswise all the tested classification models.

As shown in Figure 4(c), a variation of FNRs among the tested models was reported from low to high rates across all the benchmarking datasets that, in turn arises a crucial problem of real-time detection. Such variation was attributed to the inabilities of the tested models in identifying the aging of web page streams and the evolutionary features of phishing. Typically, the tested models trained the datasets and then set their decision boundaries for phishing classification. As such, they detected the generic features exploited by the prevalent phishes in the incoming web page stream. Other reason is that phishing patterns change very fast as long as the web page streams evolve. Indeed, the distribution of phish and legitimate classes also will change after each web page stream evolves. That, in turn, requires the classification model to update its previous decision rules which will not be to tackle the new changes. From observations in Figure 4(c), it is clear that all tested models lack of re-learning their classifiers, updating their decision rules frequently, and adapting to the up-to-date emerging novel phishes throughout their test across benchmarking datasets.

On the other hand, the plots in Figure 4(d) indicate another concern including the time of response (phish tackle) and execution along with the nominal cost of computations in phishing classification. In this concern, almost all tested models reported approximately long elapsed time to fetch the batch of dataset, extract the vectors of features, characterise phishing features (generic and/or new features), setup the classifiers' decision settings, classify phish and/or legitimate web pages, and then testify the remaining data to give a precise prediction with less faults in the future. It can be seen in Figure 4(d) that all the tested models performed well with *dataset 1* due to its small size and limited phish exploits. Whereas, the elapsed time escalated to higher rates on *dataset 2* because *dataset 2* was relatively different in its web page exploits than the others and it had a larger size than that of *dataset 1*, as presented in Table 3. Additionally, *dataset 2* involved Chinese e-business web page exploits which needed more complicated computational algorithms to extract and characterise their features. Further, most tested models learnt *dataset 3* during a very long elapsed time due to its big size, various web page exploits, and different distribution in phish/not phish class that need much computation cost. Therefore, all tested models except that adopted by Marchal et al. (2014a, 2014b) and Marchal (2015) could not attain the best case of TPRs, FPRs, FNRs, and elapsed time with *dataset 3*. In short, the conducted analysis with the obtained classification outcomes highlighted open problems of the competitive models in the term of inductive factors that will be discussed in the next section.

6 Issues and perspectives

This section addresses the main contributions of this paper via highlighting the deficiency of inductive factors and discussing the important research tendencies of future work.

6.1 Main issue: deficiency of inductive factors

The aforesaid empirical analysis showcases why none of the tested models is ideal for novel phish classification. This intricate issue, as it is described briefly in Table 4, points out this question "what inductive factors do the tested models lack?" Those inductive factors could include the following:

- Rich set of features. Day after day, phishers emerge new patterns of phishes along with the prevalent ones (Zeydan et al., 2014a, 2014b; Zuhair et al., 2016a, 2016c). In these patterns, new, numerous, and various features are exploited to bypass existing anti-phishing schemes and cause more damage to users' computer systems (Zuhair et al., 2016a, 2016c). Most of these features are more sophisticated than those already adopted by the tested models. Such new features may include hidden links for redirecting users to fake pages, obfuscated scripts of JavaScript, PHP, and ASP for malware insertion and further damage, deceptive cookies and fake advertisements in the web banners, modified source code in the terms of Applets, Flash, and ActiveX controls and other embedded objects (; Zuhair et al., 2016a, 2016c). As such, the model is at the risk of substantial misclassifications due to the partial characterisation of all types of phish variants. That, in turn, could degrade its performance in thwarting novel phishes. In this light, most tested classification models relied on generic (typical) features rather than new features, and utilised particular attributes to classify zero-hour and novel phish attacks and to characterise all the types of web exploits (Zeydan et al., 2014a; Zuhair et al., 2016b). Accordingly, they scored low to moderate rates of classification accuracy on the benchmarking datasets. On the other hand, not all the adopted features are contributing and informative to phishing class induction. They were almost limited at selecting a subset of features of minimal redundancy and/or maximal relevance to the phishing class (Zuhair et al., 2015b, 2016b). Such subset of features can be the best for phishing induction with trivial false classifications. For this reason, the tested classification models vary in their classification outcomes due to the diversity of their employed and selected features upon different datasets. In the light of this factor 'rich set of features', this paper introduces a continuation study of previously published studies addressed the importance of new features and the effects of robust feature selection (Zuhair et al., 2015a, 2015b, 2016c).
- Big data and data collection. Low in dimension, short in life span, limited in variety, imbalanced in class distribution, and divergent in the time of aggregation; altogether are the most crucial part in training the classification model (Suthaharan, 2014). Small sets of data having substantial amount of instances belonging to a specific class rather than other classes, will not be reflective enough to the abundance of the competitive classes (Kwon and Sim, 2013). Also, the big web is abundant in data but imbalanced in classes of both phish and not-phish (Huang et al., 2015). Besides phishes, the web may contain various attacks like ham and spam which share the

same features of phishing (Toolan and Carthy, 2010). Altogether, yield substantial false classifications along with complex computation and long processing time (Toolan and Carthy, 2010). On the other hands, the web offers a variety of valid phish and suspicious web pages (attacks that not yet identified as phishes by the existing anti-phishing schemes) (Shahriar and Zulkernine, 2012; Zeydan et al., 2014a; Tsai et al., 2009; Shabtai et al., 2009). Suspicious web pages could share the same web functionalities and hosting languages that usually exploited by phishers (Zeydan et al., 2014a; Zuhair et al., 2016a). Such web variety contributes a complementary factor of holistic phish characterisation; however, it may yield substantial phish misclassifications. Furthermore, retrieving phish web pages is hard to implement due to the short life span of phishes on the web, inaccessible and null phish web pages, and phish web pages leads to the same website but they are hosted in different domains (Zeydan et al., 2014a; Uzun et al., 2013; Zuhair et al., 2016a, 2016c). Thus, dataset's integrity and availability need to boost as a complementary factor of solid datasets to work with (Suthaharan, 2014; Kwon and Sim, 2013; Huang et al., 2015). To do so, a continual aggregation of data via training task leads to an implicit identification of the continuously evolved patterns of phishing, updating the induction settings, resampling the training data, and re-learning the previously generated classification model that will minimise both false detections and misclassifications (Tsai et al., 2009).

- Active learning. The task of active learning is to label instances of the most informative batch of the training data artificially (Tsai et al., 2009). The abundant instances are qualified with a particular function of phishing induction which is frequently updated throughout training every fetched batch of data (Kumar et al., 2010; Galar et al., 2012; Shabtai et al., 2009). By default, the classification model learns a batch of few labeled instances to generate the prototype prediction during the first iteration of the training phase. When an unlabeled (unknown) datum is acquired via the testing phase, the classification model uses the prototype prediction to classify the unlabeled instance. Then, it adds the classified datum to the default training data for further iteration of training. Iteratively, this procedure is repeated along with the update of induction settings and fetching unlabeled datum (Vink and de Haan, 2015). By active learning, the classification model such as FBC, FSFBC and EFBC can expect the future error with the aid of its own regulations and then pick up the most informative batch of feature vectors (web pages) which is closest to minimise that error while handling the forthcoming web stream in the testing phase (Tsai et al., 2009; Shabtai et al., 2009). Accordingly, the classification model needs to maintain a particular regulation to cope with the notable change between the initially classified feature vectors and the newly classified one (Shabtai et al., 2009; Vink and de Haan, 2015). Simultaneously, it needs to set the new margins of induction artificially and to remove the unwanted feature vectors from the old training dataset progressively (Tsai et al., 2009; Shabtai et al., 2009; Vink and de Haan, 2015). Altogether, should be repeated at every iteration of classification to tackle the unidentified phish (novel phish). Considering the empirical results and performance evaluation, all tested FBCs, FSFBCs and EFBCs lacked the factor of active learning. Additionally, they tend to use the default induction settings at every classification of unidentified web page via testing phase.

- Adaptive classification model. Almost, the tested models are of static detection model due to their inability to verify the reported predictions with respect to their rates of false positives (FPRs) (Abu-Nimeh et al., 2007; Miyamoto et al., 2008). Predictions of high FPRs could be used as available data to update the default induction boundaries of the FBCs, FSFBCs and EFBCs in order to minimise the future case of false positive classifications (Abu-Nimeh et al., 2007; Miyamoto et al., 2008). Tested classification models lacked the frequent verification mechanism which inspects the predictions of FPRs and report any change of phishing class abundance in the learning batch of data over time. As such, they lacked updating the learnt induction settings eventually and then left the unknown patterns of phishes to bypass as legitimate web pages after a period of execution time. Thus, the tested models were heavy to tackle new patterns of phishes (novel phishes) on the sequentially increasing batches of datasets. Accordingly, their performance implied a non-zero misclassification rates. Overall, dynamically updating the former predictions by a new prediction, which is made from inspecting the unlabeled web page, becomes an intricate challenge to the tested models in real-time experience. As time progresses, they are restricted at detecting phishes online with zero misclassification cost. That, in turn, requires developing an adaptive algorithm which detects any changes in the fetched web flows and accordingly updates its induction boundaries to learn them accurately.

6.2 Future perspectives

From the above highlighted issue and the obtained findings of the conducted analysis, a research question is raised to be solved: "How the inductive factors could be boosted to obtain an efficient and real-time phishing classification model?" To answer the highlighted question, it is highly recommended to consider the following perspectives for future work:

- New features. A great care must be put on exploring new features and enrich the currently used ones. That will provide holistic characterisation of both novel and prevalent phishes. Based on rich features, the induction settings of the classification models will be promoted. Specifically, if the explored features belong to the recently observed feature categories including XSS features, embedded objects, language independent features, and hybrid features (Uzun et al., 2013; Zuhair et al., 2016a, 2016c). Where, hybrid feature category involve numerous features of different categories; i.e. typically being formed from the former categories (Uzun et al., 2013; Zuhair et al., 2016a, 2016c). Additionally, variant outcomes of classification will be avoided with the heavy dependence on the best and robust selected subset of features on any dataset that escalating in its size and its age. Therefore, more optimised mechanisms should be involved in building a phishing classification model to select the best contributing features.
- Chronological training. Based on the assumption that a classification model is actively learned with the training dataset at a certain time T. A given web page w at T could be predicted as a phish in the future ($T + \check{T}$). Nevertheless, the revisited machine learning-based phishing classification models rarely estimate their prediction across the benchmarking datasets that are scaling in their dimensions and

chronological in their aggregation time. To control the compromise between the consumption of external resources and the real-time classification of phishing, aggregating datasets periodically every *T* (*time interval*) of minutes, for example; will be a complementary induction factor in less biased training and testing. In addition, chronological aggregation yields cost-sensitive classification of novel phishes on evolving web flow in the realistic-mode of practice due to its scalability versus the implicit and explicit class balance problem.

- Adaptive modelling. In the real-time application, the mechanism of building an anti-phishing scheme along with its default inductive settings falls short to detect novel phishes on up-to-date web page stream. Adaptive mechanism assesses the unknown fetched data frequently to select the most informative datum for updating its default inductive settings as well as training dataset (Kwon and Sim, 2013). As such, it would be able to detect the modest phish patterns on the upcoming data flows. To do so, an adaptive assembly could be devoted with three functionally inter-related modules working together in a synchronised mechanism: prediction module, validation module and detection module. Initially, the prediction module classifies the training web page stream and learns the FBC or FSFBC or EFBC offline. Whilst, the detection module fetches a new web page from the evolving web page stream and tests it online by the default inductive settings. Whenever a change is observed (unknown phish pattern is tackled), the detection module will exclude it and send it back to the validation module. Then, the validation module will validate the change and reconfigure the default settings (features). To update the previously trained web page stream at the prediction module, the validation module will feedback the prediction module with the excluding datum (feature vector).

7 Concluding remarks

By revisiting the current achievements in machine learning-based anti-phishing domain, it is observed that they affirmed to be computationally effective but in-adaptive to accomplish real-time phishing detection. That is due to their full or partial deficiency of inductive factors such as rich set of features, big web data and its class imbalance, actively learned FBC, and adaptable modelling. By restating the causality between their limitations and their inductive deficiency throughout an empirical analysis; future outlooks are suggested to promote their induction power. Furthermore, a phishing classification model could be extended in the future via a high level assembly integrating functionally inter-relating and synchronously working modules to adapt novel phish patterns on the evolving web page stream. Regarding the issues stated in this paper at building any machine learning-based classification model; effectiveness of classification could be elevated along with reducing the misclassification and computational cost. Additionally, this paper with the underlined perspectives are hoped to serve as a navigating taxonomy to the researchers for their future efforts.

References

Abbasi, A. and Chen, H. (2009) 'A comparison of fraud cues and classification methods for fake escrow website detection', *Information Technology and Management*, Vol. 10, Nos. 2–3, pp.83–101.

Abu-Nimeh, S., Nappa, D., Wang, X. and Nair, S. (2007) 'A comparison of machine learning techniques for phishing detection', in *Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit*, October, pp.60–69, ACM.

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*, 4th Ed., Springer, New York.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F. (2012) 'A review on ensembles for the class imbalance problem: bagging, boosting, and hybrid-based approaches', *IEEE Transactions on Systems, Man And Cybernetics, Part C, Applications and Reviews*, Vol. 42, No. 4, pp.463–484.

Gowtham, R. and Krishnamurthi, I. (2014a) 'A comprehensive and efficacious architecture for detecting phishing webpages', *Computers and Security*, Vol. 40, No. 1, pp.23–37.

Gowtham, R. and Krishnamurthi, I. (2014b) 'PhishTackle – a web services architecture for anti-phishing', *Cluster Computing*, Vol. 17, No. 3, pp.1051–1068.

Huang, G., Huang, G.B., Song, S. and You, K. (2015) 'Trends in extreme learning machines: a review', *Neural Networks*, Vol. 61, No. 1, pp.32–48.

Islam, R. and Abawajy, J. (2013) 'A multi-tier phishing detection and filtering approach', *Journal of Network and Computer Applications*, Vol. 36, No. 1, pp.324–335.

Khonji, M., Iraqi, Y. and Jones, A. (2013) 'Phishing detection: a literature survey', *Communications Surveys and Tutorials*, Vol. 15, No. 4, pp.2091–2121, IEEE.

Kumar, G., Kumar, K. and Sachdeva, M. (2010) 'The use of artificial intelligence based techniques for intrusion detection: a review', *Artificial Intelligence Review*, Vol. 34, No. 4, pp.369–387.

Kwon, O. and Sim, J.M. (2013) 'Effects of dataset features on the performances of classification algorithms', *Expert Systems with Applications*, Vol. 40, No. 5, pp.1847–1857.

Marchal, S. (2015) *DNS and Semantic Analysis for Phishing Detection*, Doctoral Dissertation, University of Luxembourg.

Marchal, S., François, J., State, R. and Engel, T. (2014a) 'PhishScore: hacking phishers' minds', *Proceedings of 10th International Conference on Network and Service Management (CNSM2014)*, IEEE, Rio de Janeiro, pp.46–54.

Marchal, S., François, J., State, R. and Engel, T. (2014b) 'PhishStorm: detecting phishing with streaming analytics', *IEEE Transactions on Network and Service Management*, Vol. 11, No. 4, pp.458–471.

Miyamoto, D., Hazeyama, H. and Kadobayashi, Y. (2008) 'An evaluation of machine learning-based methods for detection of phishing sites', in *International Conference on Neural Information Processing*, Springer, Berlin, Heidelberg, November, pp.539–546.

Mohammad, R.M., Thabtah, F. and McCluskey, L. (2014) 'Predicting phishing websites based on self-structuring neural network', *Neural Computing and Applications*, Vol. 25, No. 2, pp.443–458.

Nguyen, T.T. and Armitage, G.A. (2008) 'Survey of techniques for internet traffic classification using machine learning', *IEEE Communications Surveys and Tutorials*, Vol. 10, No. 4, pp.56–76.

Shabtai, A., Moskovitch, R., Elovici, Y. and Glezer, C. (2009) 'Detection of malicious code by applying machine learning classifiers on static features: a state-of-the-art survey', *Information Security Technical Report*, Vol. 14, No. 1, pp.16–29.

Shahriar, H. and Zulkernine, M. (2012) 'Trustworthiness testing of phishing websites: a behavior model-based approach', *Future Generation Computer Systems*, Vol. 8, No. 28, pp.1258–1271.

Suthaharan, S. (2014) 'Big data classification: Problems and challenges in network intrusion prediction with machine learning', *ACM Sigmetrics Performance Evaluation Review*, Vol. 41, No. 4, pp.70–73.

Toolan, F. and Carthy, J. (2010) 'Feature selection for spam and phishing detection', paper presented at *eCrime Researchers Summit (eCrime)*, Dallas, TX, pp.1–9.

Tsai, C-F., Hsu, Y-F., Lin, C-Y. and Lin, W-Y. (2009) 'Intrusion detection by machine learning: a review', *Expert Systems with Applications*, Vol. 36, No. 10, pp.11994–12000.

Uzun, E., Agun, H.V. and Yerlikaya, T. (2013) 'A hybrid approach for extracting informative content from web pages', *Information Processing and Management: an International Journal*, Vol. 49, No. 4, pp.928–944.

Vink, J.P. and de Haan, G. (2015) 'Comparison of machine learning techniques for target detection', *Artificial Intelligence Review*, Vol. 43, No. 1, pp.125–139.

Wardman, B., Britt, J. and Warner, G. (2014) 'New tackle to catch a phisher', *International Journal of Electronic Security and Digital Forensics*, Vol. 6, No. 1, pp.62–80.

Whittaker, C., Ryner, B. and Nazif, M. (2010) 'Large-scale automatic classification of phishing pages', paper presented in *17th Annual Networks and Distributed System Security Symposium (NDSS2010)*, The Internet Society, March, San Diego, California, USA.

Xiang, G. (2013) *Towards a Phish Free World: a Cascaded Learning Framework for Phishing Detection*, p.15213, Doctoral Dissertation, Carnegie Mellon University, Pittsburgh, PA.

Zeydan, H.Z., Selamat, A. and Salleh, M. (2014a) 'Current state of anti-phishing approaches and revealing competencies', *Journal of Theoretical and Applied Information Technology*, Vol. 70, No. 3, pp.507–515.

Zeydan, H.Z., Selamat, A. and Salleh, M. (2014b) 'Survey of anti-phishing tools with detection capabilities', in the *Proceedings of 14 Int. Symposium on Biometrics and Security Technologies (ISBAST'2014)*, Kuala Lumpur, Malaysia.

Zhang, D., Yan, Z., Jiang, H., and Kim, T. (2014) 'A domain-feature enhanced classification model for the detection of Chinese phishing e-business websites', *Information and Management*, Vol. 51, No. 7, pp.845–853.

Zuhair, H., Salleh, M. and Selamat, A. (2016a) 'Hybrid features-based prediction for novel phish website', *Jurnal Teknologi*, Vol. 78, Nos. 12–33, pp.95–109.

Zuhair, H., Selamat, A. and Salleh, M. (2016b) 'Feature selection for phishing detection: a review of research', *International Journal of Intelligent Systems Technologies and Applications*, Vol. 15, No. 2, pp.147–162.

Zuhair, H., Selamat, A. and Salleh, M. (2016c) 'New hybrid features for phish website prediction', *International Journal of Advances in Soft Computing and Its Applications*, Vol. 8, No. 1, pp.28–43.

Zuhair, H., Selamat, A. and Salleh, M. (2015a) 'Selection of robust feature subset for phish webpage prediction using maximum relevance and minimum redundancy criterion', *Journal of Theoretical and Applied Information Technology*, Vol. 81, No. 2, pp.188–205.

Zuhair, H., Selamat, A. and Salleh, M. (2015b) 'The effect of feature selection on phish website detection: an empirical study on robust feature subset selection for effective classification', *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 1, No. 6, pp.221–232.