Phishing Hybrid Feature-Based Classifier by Using Recursive Features Subset Selection and Machine Learning Algorithms

Hiba Zuhair[1(✉)] and Ali Selamat[2]

[1] College of Information Engineering, Al-Nahrain University, Baghdad, Iraq
hiba.zuhair.pcs2013@gmail.com
[2] Faculty of Computing, Universiti Teknologi Malaysia (UTM), Johor, Malaysia
aselamat@utm.my

Abstract. Machine learning classifiers enriched the anti-phishing schemes with effective phishing classification models. However, they were constrained by their deficiency of inductive factors like learning on big and imbalanced data, deploying rich sets of features, and learning classifiers actively. That resulted in heavyweight phishing classifiers with massive misclassifications in real-time phishing detection. To diminish this deficiency, this paper proposed a new Phishing Hybrid Feature-Based Classifier (PHFBC) which hybridized two machine learning algorithms (Naïve Base) and (Decision Tree) with a statistical criterion of Phish Ratio. In conjunction, a Recursive Feature Subset Selection Algorithm (RFSSA) was also proposed to characterize phishing holistically with a robust selected subset of features. Outcomes of performance assessment via simulations, real-time validation, and comparative analysis demonstrated that PHFBC was highly distinctive among its competitors in terms of classification accuracy and minimal misclassification of novel phishes on the Web.

Keywords: Machine learning · Feature-based classifier
Feature subset selection · Maximal relevance · Minimal redundancy
Active learning

1 Introduction

Although, the web provides a huge communication channel and many services to the users and enterprises, it causes significant digital identity theft and monetary loss annually due to phishing attacks. Phishers evolve their attacks by impersonating the trustworthy web pages in phish web pages to deceive users [1, 2]. Therefore, researchers develop different anti-phishing schemes to tackle phish web pages and mitigate their consequences. Among them are those assisted by machine learning classifiers [1–3]. The developed machine learning classifiers deploy baseline machine learning algorithms such as *Naïve Bayes* (*NB*), *Support Vector Machine* (*SVM*), *Logic Regression* (*LR*), *Sequential Minimum Optimization* (*SMO*), *Random Forest* (*RF*), *C4.5*, and *JRip* [3]. In practice, they outperform their competitors because they rely on various features that distinguish phishing deceptions to give their overall decisions [2–4]. However, they still can be evaded by novel phish web pages due to their

F. Saeed et al. (Eds.): IRICT 2018, AISC 843, pp. 267–277, 2019.
https://doi.org/10.1007/978-3-319-99007-1_26

deficiency of inductive factors which led to partial characterization and inefficient classification in real-time application [3–5]. To hinder their drawbacks, this paper proposes *Phishing Hybrid Feature-Based Classifier* (*PHFBC)* which is assisted by features subset selection algorithm and hybridizes two most salient machine learning algorithms with a statistical induction criterion. The proposed *PHFBC* demonstrate its effectiveness and supremacy throughout an experiment on benchmarking data sets, a week of real-time practice, and a comparative analysis versus the most salient state-of-the-art classifiers. Next sections gives a bird's eye on PHFBC as follows: Sect. 2 analyzes the related works to address their main problems. Section 3 depicted the work flow of *PHFBC* in details. Section 4 discusses the *PHFBC* performance. Whereas, Sect. 5 presents the conclusions and future insights.

2 Background

Among the prominent phishing classifiers, was *CANTINA*$^{+}$ that learned *Naïve Bayes* (*NB*), *Support Vector Machine* (*SVM*), and *Logic Regression* (*LR*) with 15 features of web page content and URL to effectively classify phishing in pharming and login web pages. It achieved a *True Positive Rate of* 92% and a *False Positive Rate* of 1.4% [5]. However, it encountered a trade-off in classifying phishing on the up-to date web flows due to the limited number and genericity of the used features. In [6, 7] a phishing classifier with *Support Vector Machine* (*SVM*) algorithm is devoted to classify phish login forms by using 17 features. Despite its effectiveness (*True Positive Rate* and *False Positive Rate* of 99.6% and 0.44% respectively); it performed a heavyweight classification and used external resources for data inquiry. Then, a Chinese e-business phishing website classifier was developed in [8] to characterize phishing with the best ranked set of 15 URL features by using *Chi-Squared* (χ^2). Also, it relied on *Sequential Minimum Optimization* (*SMO*), *Logic Regression* (*LR*), *Naïve Bayes* (*NB*), and *Random Forests* (*RF*) as machine learning algorithms. It achieved an accuracy rate of 95.83% on Chinese e-business web pages execlusively. Later, an ensemble classifier learned *Support Vector Machine* (*SVM*), *Random Forest* (*RF*), *C4.5*, and *JRip* with 12 features was developed in [9]. It was assisted by three methods to select the best features such as *Correlation Features Based Selection* (*CFS)*, *Information Gain* (*IG*), and *Chi-Squared* (χ^2). However, it achieved high rate of classification faults (1.44%) with moderate rate of accuracy (94.91%) on big and imbalanced datasets. Other classifier was devoted in [10, 11] to tackle phishing in e-business, login webpages that hosted in English and French. Therefore, it utilized a set of typical and new features to learn *Neural Network* (*NN*) algorithm. It performed well with accuracy rate of (94.07%) but a high rate of misclassifications on big and imbalanced datasets. Then, such classifier was optimized and presented in [12] to learn a set of 212 typical and new features on a big dataset (96,018 samples). However, it achieved significant performance overhead versus high accuracy. Almost revisited classifiers encountered the problems of: (i) evasion by the novel phish web pages continually to cause more damage and illegal gain [1–3]; (ii) limited tolerance of features strongly relevance and redundancy in web pages [1–4, 13–16]; (iii) extensive crawling, processing, and induction for phishing characterization and classification on different web exploits [1–4, 13, 14]; (iv) learning on unreflective

datasets versus the vast web page streams of different size, class abundance, web page exploits, and aggregation time [3, 4, 15, 16], (v) inactive learning to adjust induction settings and inadaptability to classify novel phish on the fetched web page streams at any given time [3, 4, 17, 18], and then (v) ambiguous and ineffective phishing real-time detection. Altogether, were attributed to the deficiency of inductive factors that is the main concern to solve and the key contribution in this paper as it will be presented in the next sections.

3 PHFBC

3.1 Features Extraction and Selection

The training set of web pages (dataset) was formulated into a feature space of feature vectors. Each feature vector consisted of the values of 58 various features a they were extracted from web page source code and URLs. Such 58 features included ten *URL* features, 24 *Cross Site Scripting* (*XSS*) features, and 24 *HTML* features as they were proposed in our previously published work [13, 14]. Such features represented the advanced activities and newly explored deceptions of phishers such as imitation of trustworthy web page by embedding dynamic objects and Flash attributes, and injection of suspicious java scripts for malware damage. Then, an optimal selection of features was relied on nominating the most decisive and distinctive features without compromising their minimal exploitation in the input feature space by using a Recursive Features Subset Selection Algorithm (RFSSA) which is assisted by a sub-algorithm FSA and supportive specifics; that was proposed in our previous works [15, 16]. As illustrated in Fig. 1, *FSA* enables RFSSA to boost the mutual information of the targeting class and the mutual dependency among features in the same set of features for the best selection of the most distinctive features (Red_Relv*)* from the input features set (FSet) by using Maximum Relevance and Minimum Redundancy Criterion (mRMR) [17, 18] Then, the candidate (Red_Relv) is projected from (FSet) into (OutSet) to be fed to RFSSA for features subset prioritization so that RFSSA splits (OutSet) into N subsets to validate their goodness (Good Ratio), stability (Stab Ratio), and similarity (SimRatio) [4, 15, 16]. Those supportive specifics are presented in [15, 16] along with their mathematical modelling.

3.2 Phishing Classification

PHFBC boosted the deficiency of inductive factors that Naïve Bayes (NB) and Decision Tree (DT) suffering from. For more decisive classification, it hybridized NB and DT in a synchronized platform and it complemented their deficiency by pruning their induction settings using a statistical measure Phish Indication Ratio (or Phish Ratio), (see Figs. 2 and 3). Then, PHFBC manifested its induction setting iteratively by

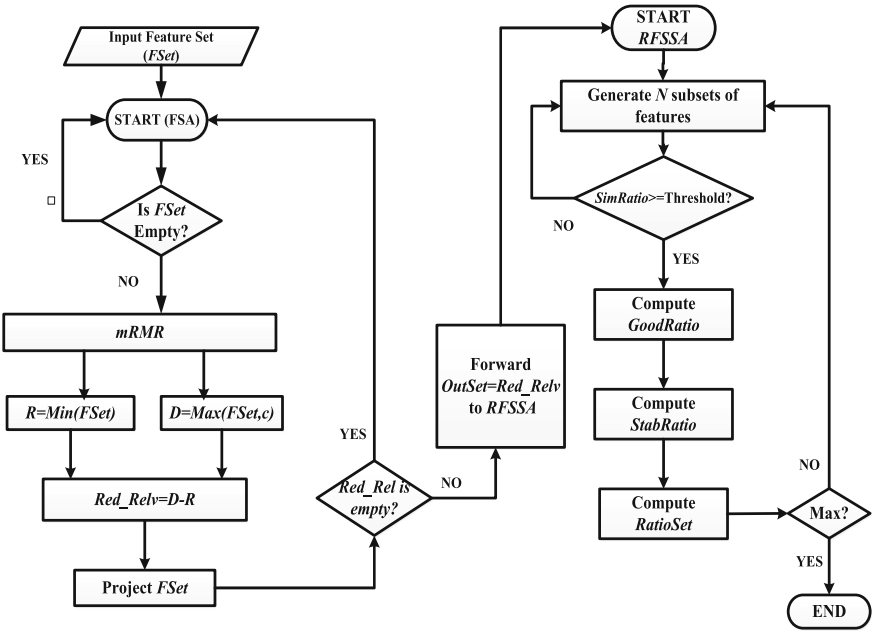


Fig. 1. Flowchart of RFSSA.

mapping the misclassified features in the illegible nodes into their true predictive classes and converged the unclassified feature vectors (overlooked by NB and/or DT) from the remaining feature space. Finally, PHFBC re-learned the feature space progressively and updated its induction settings as along as an unclassified feature vector was inspected. For effective classification, almost revisited phishing classifiers adopted two baseline machine learning algorithms, Naïve Bayesian (NB) and Decision Tree (DT) [5–7, 10–12], see Appendix Table 1. NB pursued Bayes' theorem probabilistically to classify samples on a small training data set to estimate the classifier's parameters correctly and artificially by assuming that all the used features in the training samples were independent from each other [5, 16–18]. Whereas, DT could be sketched in a particular tree structure of nodes, leafs, and branches for mining data statistically and generating the predictive labels effectively. The nodes represented all features in the input feature space (formulated data set), and the leaves denoted the predictive labels of those features while the branches conjunct the inspected features to their relevant predictive labels. Thus, DT set throughout two phases: tree building and tree pruning phases. During tree building phase, the training data set was split recursively to assign each included feature with its relevant predictive labels. In pruning phase, each subtree was pruned by traversing its relevant branches to inspect the minimal training error [5, 16–18]. In spite of their effective classification approach, both NB and DT might perform differently in the case of real-time detection. NB lacked of handling big and imbalanced data, and it fall short in tolerating features heterogeneity and then classifying the duplicate samples belonging to different classes

accurately. On the other hand, DT might overlook the unknown features due to the deployment of its default mapping of features and default induction margins via tree building and pruning phases respectively [3, 17, 18].

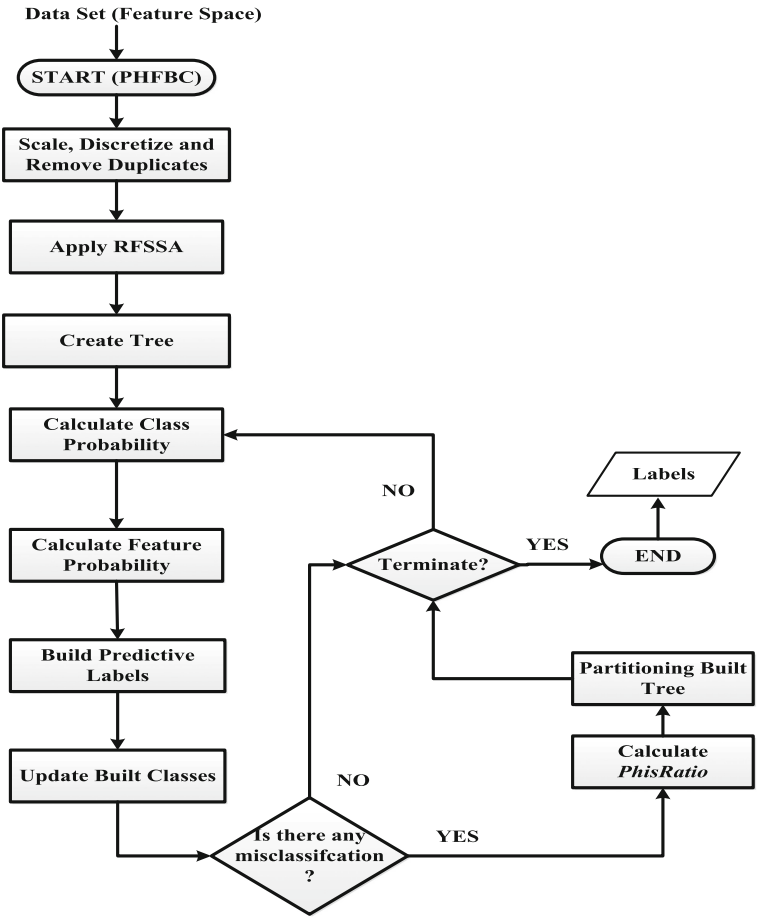
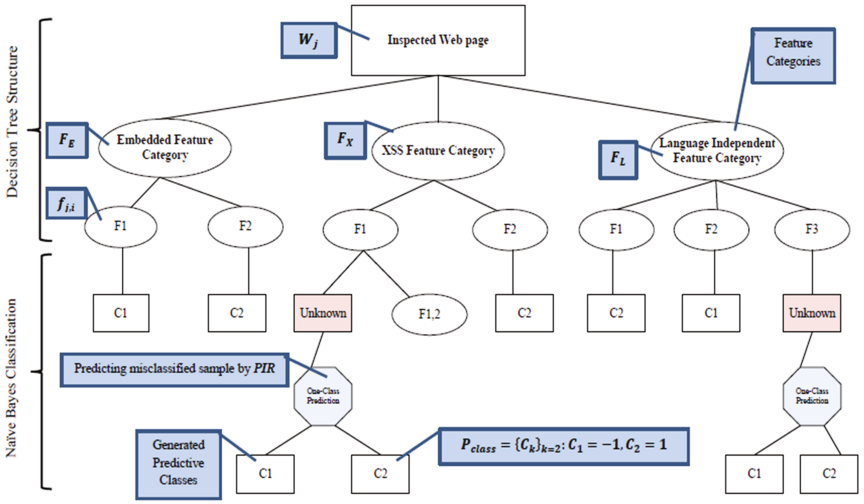


Fig. 2. Workflow of *PHFBC*

4 Performance Assessment

PHFBC's performance was analyzed to justify its distinction among its competitors throughout three strategies: a practice of PHFBC across three benchmarking data sets (see Table 1 in Appendix), a week of real-time validation of PHFBC, and a comparative analysis of PHFBC versus five baseline machine learning-based classifiers.



Note: W_j is the examined web page
P_{Class} represents the predictive class such that $P_{class} = \{C_1, C_2\}$;
$C_1 = 1$ denotes phish class;
$C_2 = -1$ denotes non-phish;
$f_{j,i}$ is a given feature in the examined web page;
F_H refers to hybrid feature category such that $F_H = (F_E, F_X, F_L)$;
F_E refers to embedded objects features category;
F_X refers to XSS based features category;
F_L refers to Language independent feature category.

Fig. 3. Classification process through PHFBC.

Related simulation done by using "WEKA 3.5.7-Waikato Environment for Knowledge Analysis" and analysis was pursued by using four typical evaluation metrics: True Positive Rate (TPR) which rationalized the correctly classified phishing over all phish samples, False Positive Rate (FPR) which rationalized the wrongly classified legitimate samples as phishing over the total number of legitimate samples, False Negative Rate (FNR) that rationalized the wrongly labeled phishing samples as legitimate samples overall samples, and AUC that rationalized the plots of TPR versus plots of FPR in the area under the ROC into a scalar value. Experimentally, PHFBC achieved the best rates of TPR (from 0.984 to 0.989), FPR (from 0.051 to 0.066), and FNR (from 0.014 to 0.0156) across the benchmarking data sets. That demonstrated its well performance in the conjunction with RFSSA and 58 hybrid set of features and attributed to the richness and robust compactness of features subset at any settings against the scalable and imbalanced datasets; i.e. RFSSA provided subsets of maximal relevant features to phishing class and minimal redundant features to phish class exploitation and distribution (see Fig. 4 in Appendix).

On the other hand, PHFBC showed progressive and effective phishing classification during 1-week of real-time validation (see Fig. 5 in Appendix). PHFBC reported an ideal classification outcome (AUC was equal to "1") in the last day. That manifested its decisive and effective classification due to the hybridization of two complementary machine learning algorithms (NB and DT) with a statistical criterion (Phish Ratio), update of its default induction margins, and actively learned at every fetched web flow. So far, PHFBC showed its superiority versus five of baseline machine learning algorithms like SMO, SVM, TSVM, NB and DT (see Fig. 6 and Table 1 in Appendix). This was disclosed to: (i) comparable classifiers rendered variations in TPRs and FPRs because they fall short in deploying rich features, tolerating big and imbalanced datasets, and learning inactively (see Fig. 6a and b in Appendix). Whilst, PHFBC achieved the highest TPRs and the least FPRs among its competitors that assured its decisive characterization of different phish exploits and effective classification of novel phishes over the scalable datasets; (ii) the active learning of PHFBC versus inactive learning of the comparable classifiers attained the minimal misclassification cost (FNRs of PHFBC were the closest scores to zero among those of its competitors). Because PHFBC could adjust its initial induction margins by hybridizing NB and DT with Phish Ratio. Then it could adapt to tackle novel phishes on scalable benchmarking datasets.

5 Conclusion and Future Work

This paper addressed inductive factors in phishing machine learning-based classifiers by proposing a Phishing Hybrid Feature-Based Classifier (PHFBC). Conceptually, PHFBC hybridized a Recursive Feature Subset Selection Algorithm (RFSSA) and two complementary machine learning algorithms Naïve Bayes and Decision Tree with a statistical criterion Phish Indication Ratio. Experimentally, PHFBC achieved (97%), (0.7%), (0%), and (98.07%) average rates of TPR, FPR FNR, and AUC respectively. That demonstrated its supremacy as a novel phish-aware and a real-time classifier among its competitors due to its abundance in inductive factors. It employed rich sets of 58 features to characterize Phish web page exploits holistically. It adjusted its margins of induction via the hybridity of statistical and machine learning-based algorithms. It actively learned on big and imbalanced web page streams to adapt novel phishes. Empirical and deductive proofs of its decisive and effective classification, assured that PHFBC can serve as a prospective approach in the future of anti-phishing.

Appendix

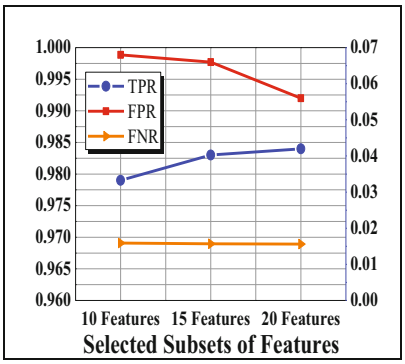
See Tables 1, 2 and Figs. 4, 5, and 6.

Table 1. Contributions of this work versus the related works

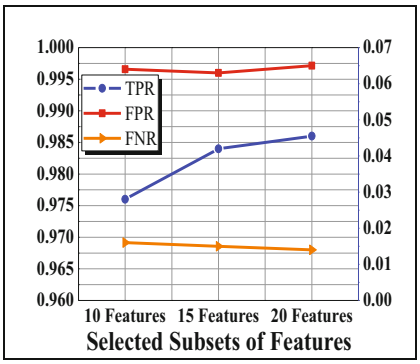
Merits	Classifier					
	[5]	[6, 7]	[8]	[9]	[12]	This work
Features	15	17	17	15	122	58
Classifier (s)	SVM, LR, BN, DT	SVM	NN, DT	SMO, LR, RF, NB	SVM, RF, C4.5, JRip	PHFBC
Features selection	No	No	Yes	Yes	No	Yes
Active learning	No	No	No	Yes	No	Yes
Data imbalance	No	Yes	No	No	Yes	No
Irrelevance	No	Yes	No	No	Yes	No
Redundancy	Yes	Yes	Yes	Yes	Yes	No
Outcomes	TPR (92%), FPR (1.4%)	TPR (99.6%), FPR (0.42%)	TPR 94.07% FPR: (2.2)	TPR 95.83% FNR: (1.05)	TPR (99%), FPR (0.37%)	TPR 99%, FPR 0.7%

Table 2. Benchmarking datasets

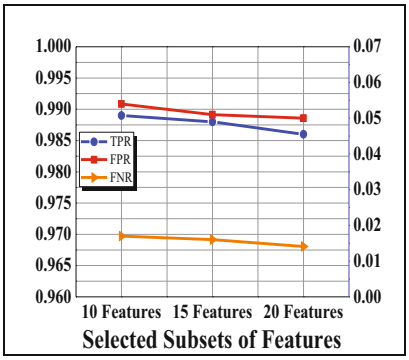
Merits	Data set 1	Data set 2	Data set 3
Size/phishes legitimates	52/36/16	2878/1382/1496	96,018/48009/48009
Data archive	PhishTank	Chinese e-Business	PhishTank/DMOZ
Training split ($2/3^{rd}$)	34	1918	64012
Testing split ($1/3^{rd}$)	18	960	32006
Data source	[19]	[8]	[10–12]
Aggregation time	2010	2014	2012–2015
Webpage exploits	Login form/e-Business/Pharming	e-Business	e-Business/homepage/ login form
Hosting language	English/French/German	Chinese	English/French/German



(a) Performance on Data Set1



(b) Performance on Data Set2



(c) Performance on Data Set3

Fig. 4. Performance assessment of *PHFBC* with respect to features selection

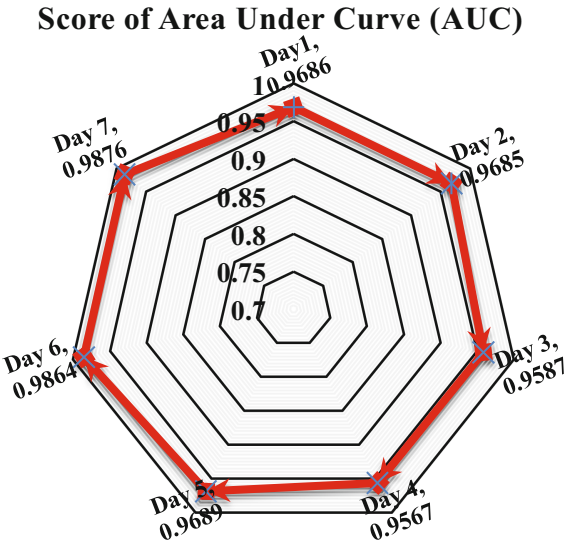
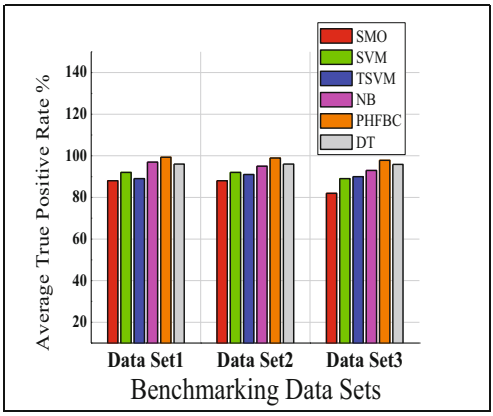
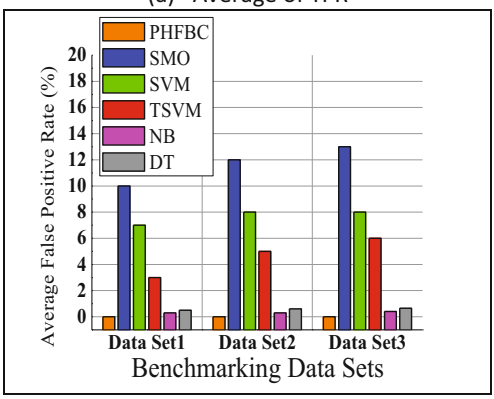


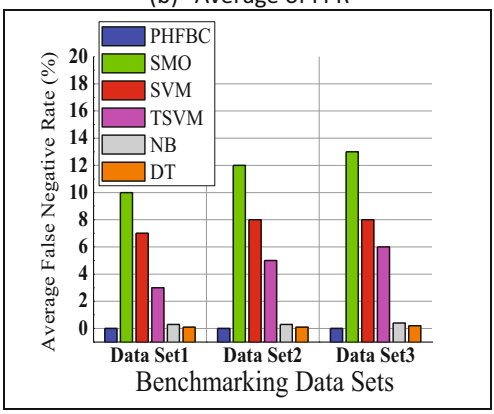
Fig. 5. Real-time validation of *PHFBC*



(a) Average of TPR



(b) Average of FPR



(c) Average of FNR

Fig. 6. Outcomes of comparative analysis

References

1. Zeydan, H.Z., Selamat, A., Salleh, M.: Survey of anti-phishing tools with detection capabilities. In: 14th International Symposium on Biometrics and Security Technologies (ISBAST 2014), pp. 46–54, Kuala Lumpur-Malaysia, August 2014
2. Zeydan, H.Z., Selamat, A., Salleh, M.: Current state of anti-phishing approaches and revealing competencies. J. Theor. Appl. Inf. Technol. **70**(3), 507–515 (2014)
3. Zuhair, H., Selamat, A.: Phishing classification models: issues and perspectives. In: IEEE Conference on Open Systems (ICOS 2017), pp. 26–31, IEEE, Miri-Sarawak (2017)
4. Zuhair, H., Selamat, A., Salleh, M.: Feature Selection for phishing detection: a review of research. Int. J. Intell. Syst. Technol. Appl. **15**(2), 147–162 (2016)
5. Xiang, G.: Towards a phish free world: a cascaded learning framework for phishing detection. Doctoral Dissertation, Carnegie Mellon University, Pittsburgh, PA 15213 (2013)
6. Gowtham, R., Krishnamurthi, I.: A comprehensive and efficacious architecture for detecting phishing webpages. Comput. Secur. **40**, 23–37 (2014)
7. Gowtham, R., Krishnamurthi, I.: PhishTackle-a web services architecture for anti-phishing. Clust. Comput. **17**(3), 1051–1068 (2014)
8. Zhang, D., Yan, Z., Jiang, H., Kim, T.: A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. Inf. Manag. **51**(7), 845–853 (2014)
9. Mohammad, R.M., Thabtah, F., McCluskey, L.: Predicting phishing websites based on self-structuring neural network. Neural Comput. Appl. **25**(2), 443–458 (2014)
10. Marchal, S., François, J., State, R., Engel, T.: PhishScore: hacking phishers' minds. In: 10th International Conference on Network and Service Management (CNSM2014), pp. 46–54, IEEE, Rio de Janeiro, 17–21 November 2014
11. Marchal, S., François, J., State, R., Engel, T.: PhishStorm: detecting phishing with streaming analytics. IEEE Trans. Netw. Serv. Manag. **11**(4), 458–471 (2014)
12. Marchal, S.: DNS and semantic analysis for phishing detection. Doctoral Dissertation. University of Luxembourg, 22 June 2015
13. Zuhair, H., Salleh, M., Selamat, A.: New hybrid features for phish website prediction. Int. J. Adv. Softcomput. Appl. **8**(1), 28–43 (2016)
14. Zuhair, H., Salleh, M., Selamat, A.: Hybrid features-based prediction for novel phish website. Jurnal Teknologi **78**(12–13), 95–109 (2016)
15. Zuhair, H., Selamat, A., Salleh, M.: Selection of robust feature subsets for phish webpage prediction using maximum relevance and minimum redundancy criterion. J. Theor. Appl. Inf. Technol. **81**(2), 188–205 (2015)
16. Zuhair, H., Selamat, A., Salleh, M.: The effect of feature selection on phish website detection: an empirical study on robust feature subset selection for effective classification. Int. J. Adv. Comput. Sci. Appl. **6**(10), 221–232 (2016)
17. Vink, J.P., Haan, G.: Comparison of machine learning techniques for target detection. Artif. Intell. Rev. **43**, 125–139 (2015)
18. Kumar, G., Kumar, K., Sachdeva, M.: The use of artificial intelligence based techniques for intrusion detection: a review. Artif. Intell. Rev. **34**(4), 369–387 (2010)
19. Shahriar, H., Zulkernine, M.: Trustworthiness testing of phishing websites: a behavior model-based approach. Future Gener. Comput. Syst. **8**(28), 1258–1271 (2012)