A Framework of New Hybrid Features for Intelligent Detection of Zero Hour Phishing Websites

Thomas Nagunwa, Syed Naqvi, Shereen Fouad and Hanifa Shah

School of Computing and Digital Technology, Birmingham City University, UK
thomas.nagunwa@mail.bcu.ac.uk

Abstract. Existing machine learning based approaches for detecting zero hour phishing websites have moderate accuracy and false alarm rates and rely heavily on limited types of features. Phishers are constantly learning their features and use sophisticated tools to adopt the features in phishing websites to evade detections. Therefore, there is a need for continuous discovery of new, robust and more diverse types of prediction features to improve resilience against detection evasions. This paper proposes a framework for predicting zero hour phishing websites by introducing new hybrid features with high prediction performances. Prediction performance of the features was investigated using eight machine learning algorithms in which Random Forest algorithm performed the best with accuracy and false negative rates of 98.45% and 0.73% respectively. It was found that domain registration information and webpage reputation types of features were strong predictors when compared to other feature types. On individual features, webpage reputation features were highly ranked in terms of feature importance weights. The prediction runtime per webpage measured at 7.63s suggest that our approach has a potential for real time applications. Our framework is able to detect phishing websites hosted in either compromised or dedicated phishing domains.

Keywords: Phishing, Phishing webpage detection, Zero hour phishing website, Webpage features, Machine learning.

1 Introduction

Phishing websites mimic their targeted legitimate websites to trick users to collect their Personal Identification Information (PII) for online frauds [1]. Today, many phishers use easily available sophisticated tools including phishing toolkits and fast flux networks to create and host large number of highly dynamic and quality phishing websites [2, 3]. Consequently, there has been a rapid growth of new and unknown (zero hour) phishing websites in recent years [4]. As 91% of all global security breaches begin with phishing attacks and phishing websites being the key player [5], effective detection of the websites is inevitable towards making cyber space safe.

Phishing website detection solutions are mainly based on blacklists and heuristics techniques. Blacklists extensively use human skills in maintaining records of the databases therefore they lack real time intelligence to detect zero hour phishing websites [6,

7]. Heuristics solutions analyze distinctive webpage features using various algorithmic approaches to detect phishing webpages. Many of them have reported moderate accuracy and false alarm rates [7]. Phishers also have been learning their prediction features and adopt corresponding obfuscations in their phishing webpages to enhance detection evasion [7]. This is facilitated with the limited number and diversity of features used by most of the solutions. Therefore continuous discovery and adoption of new, robust and highly diversified features is vital in maintaining effective detection.

This paper proposes a framework of new hybrid features for real time prediction of zero hour phishing webpages using machine learning. A total of 31 features, of which 26 are novel, from five different types of features, the most diverse compared to previous works, are proposed. Features related to different URL components (FQDN[1], domain and path) are introduced to enable the framework to detect phishing websites hosted in either compromised or dedicated phishing domains. The framework for the implementation of the prediction process is designed and presented, in which three modules are introduced to pre-process webpages to improve accuracy and efficiency. Our webpage pre-processors include JavaScript form detector and URL redirections check modules which have never been used before. Eight machine learning classification algorithms are applied to evaluate the extracted features' data to develop a best performing prediction model. We also evaluate overheads of the prediction process by computing an average prediction runtime per a webpage.

The paper is organized as follows; section 2 discusses related works, section 3 introduces the framework's design while section 4 describes the conducted experiments and their results. Finally, sections 5 and 6 provide discussions and conclusion respectively.

2 Related Works

Several studies have applied machine learning (ML) approaches for predicting zero hour phishing webpages. Generally, they have used different diversity of feature types, typically between one and four types of features. For instance, [8] extracted 1701 word and 40 natural language processing based features, all from the URL, for the prediction. By evaluating the features using seven classifiers, Random Forest produced the highest accuracy of 97.98%. Zuhair et al. [9] investigated and designed 58 predictive features in which 48 and 10 were webpage structure and URL related features respectively. Using SVM classifier, they evaluated the features and the resulting model produced false positives (FP) of 1.17% and false negatives (FN) of 0.03%. Li et al. [10] also developed 20 features of the same two types of features to create a fast real time prediction model. By combining Gradient Boosting DT, XGBoost and LightGBM algorithms, the model obtained an accuracy, misclassification rate and FN of 97.3%, 4.46% and 1.61% respectively. Jain et al. [11], similar to [10], did not use third party features to avoid network overheads so as to improve efficiency.

Studies including [12] used a hybrid of three types of features while others such as [1], [13], [14] and [15] used four types of features. Mohammad et al. [12], for instance,

[1] Full Qualified Domain Name, also known as hostname of a webpage.

developed 5, 9 and 3 features related to webpage structure, URL and domain registration information respectively. They applied deep neural network to develop a prediction model of an accuracy of 92.18%. Feng et al. [14] extracted 30 features related to webpage structure, URL, domain registration and webpage reputation and evaluated them against eight classifiers to develop the prediction model. A deep neural network algorithm obtained an optimal accuracy of 97.71%.

Generally, most of the reviewed studies scored accuracy and error rates of between 81% and 98.5%, and 0.43% and 18% respectively. We observed that only [13] and [16] deployed webpage pre-processor in which a HTML form detector was implemented to filter out webpages without the HTML forms. The work [16] also filtered known blacklisted webpages using a computing intensive SHA1 hash value comparison method.

3 Framework Design

3.1 Architecture

We have designed a machine learning based framework of new hybrid of features to predict whether the user requested webpage, prompting for PII, is phishing or not. Figure 1 demonstrates the framework's architecture. Paths 1-3 and 4-7 represent modelling of the classifier and the prediction processes respectively. The framework consists of the following six modules;

PII webpage filtering. A webpage requested by a user is checked if it prompts PII by examining if the webpage contains at least one of the PII webpage phrases and a HTML form or a JavaScript popup. Most of the webpages use the form or popup to prompt and collect PII. The PII webpage phrase is a word such as *sign in* and *login* contained in a webpage that is related to purpose of the webpage in collecting specific PII. We identified 43 PII phrases (in English) that we found were commonly used by over 50 samples of English based legitimate webpages prompting PII that we collected before. We used simple lookups, for instance, searching for a HTML form tag *<form…>…</form>* in combination with at least one of the phrases in a webpage, to reduce overheads. The role of the module is to filter out webpages which do not collect PII and therefore can never be phishing. This avoids misuse of computer and network resources for analyzing irrelevant webpages, thus optimizing users' web browsing experience as well as avoiding false positive errors due to positive prediction for webpages which do not prompt for PII.

Phishing blacklist check. A webpage's URL is checked if it exists in a PhishTank's phishing URL blacklist, one of the most reputable online databases of blacklisted phishing URLs. The module's role is to filter out webpages which have already been confirmed to be phishing, thus enhancing detection efficiency and reducing false negatives. We use a simple lookup of a URL in the blacklist to reduce overheads.

URL redirections check. We observed that a significant number of both phishing and legitimate webpages have their first visible URLs being redirected to other URLs when downloading the webpages. Redirections in phishing webpages may be for the reason

of hiding the true identity of the actual URLs hosting the webpages. Our interest was to learn features of the final redirected URLs, as actual addresses of the hosts. The module's role is to detect existence of all common redirections and extract their final URLs. Types of URL redirections detected were client-end redirections (implemented in HTML Meta and JavaScript tags), server-end redirections and short to long URL conversions. By determining the final redirected URLs, we ensured that we have collected relevant URL feature data to improve accuracy and error rates.

Feature data extraction. In this module, data about the webpage features are extracted from the webpage as well as from third party services such as search engines.

Training a classifier. A classifier builds a prediction model from the training dataset.

Prediction analysis. The prediction model analyses features' data extracted from a new webpage and generates a prediction result.

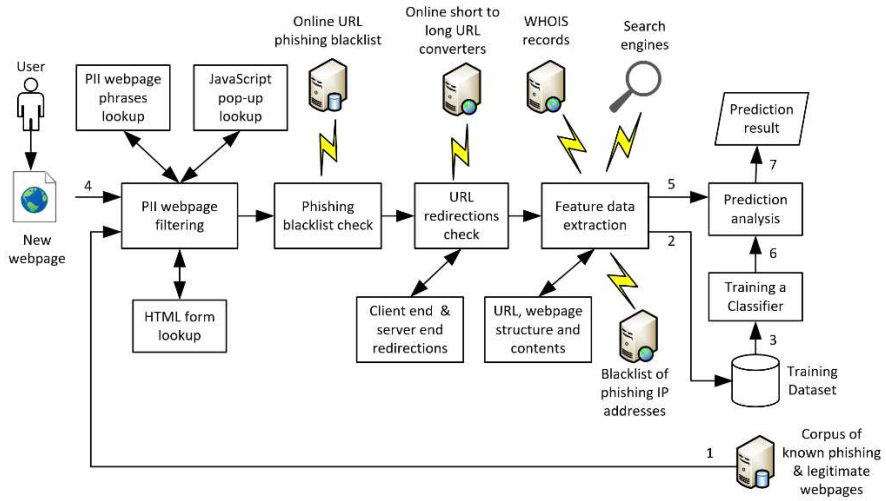

Fig. 1. An architecture of the proposed framework for predicting zero hour phishing websites.

3.2 Phishing Webpage Predictive Features

We have developed 31 webpage features, as listed in table 1, to model the classifier to predict phishing webpages. The features are categorized into five different types as described below.

Webpage structure and contents. The features (F.1 - F.7) are related to information contained in a webpage as content or part of its HTML/script structure.

URL structure. The features (F.8 – F.20) define specific decomposition characteristics of a webpage's URL. The features are related to the uses, positions and counts of special characters as well as the uses of third party services to host or form a URL.

Domain registration information. The features (F.21 – F.24) are related to domain registration information kept by domain registrars. Such information is retrieved from registrars' online WHOIS databases.

SSL certificate information. The features (F.25 and F.26) are related to the information contained in a SSL certificate of the webpage.

Webpage reputation. The features (F.27 - F.31) measure reputation of a webpage in both Google and Bing search engines and in a blacklist of IP addresses of PhishTank's phishing websites.

Table 1. The proposed phishing webpage predictive features.

F.1	Domain identity[2] in a webpage	F.17	Number of characters in FQDN
F.2	Domain identity in copyright	F.18	Number of characters in URL path
F.3	Domain name in canonical URL	F.19	Shortened URLs
F.4	Domain name in alternate URL	F.20	Free subdomain services
F.5	Foreign domains in hyperlinks	F.21	Domain name's validity
F.6	Void hyperlinks ratio	F.22	Domain age
F.7	Foreign form handler	F.23	Form handler's domain name validity
F.8	URL path encoding	F.24	Form handler's domain age
F.9	Use of '@' character in a URL	F.25	Type of a SSL certificate
F.10	Domain out positioning	F.26	Domain, certificate and geolocation country matching
F.11	Number of dots in FQDN	F.27	URL search engine ranking
F.12	Number of dots in the URL path	F.28	Domain search engine ranking
F.13	Unconventional port numbers	F.29	FQDN search engine ranking
F.14	Obfuscation characters in FQDN	F.30	FQDN blacklist IP counts
F.15	Obfuscation characters in URL path	F.31	Domain blacklist IP counts
F.16	Number of forward slashes		

We have also designed features related to different URL components (FQDN, domain and path), for instance, F.17, F.18, and F.27 – 31, to detect phishing websites hosted in either compromised or dedicated phishing domains. This is because phishing websites hosted in compromised domains have similarities with their hosts' legitimate websites in many features, including F.1, F.2, F.21 and F.22. For instance, if F.28 flags 'No' then the website is hosted in a dedicated phishing domain, and if F.28, F.29 and F.27 flag 'Yes', 'Yes' and 'No' respectively, the website is likely to be hosted in a compromised domain.

Of all the features, 18 features were based on the webpage's structure and contents while the other 13 features were based on third party services containing information related to a webpage. Along with the use of five different types of features, such diversity enhances resiliency against current obfuscation techniques deployed by phishers to

[2] Domain identity refers to a second level or third level domain label that represent an identity of the website owner. For instance, for a URL https://accounts.google.com/ServiceLogin, *google* is the domain identity.

circumvent detection. Third party services such as WHOIS records can hardly be obfuscated by phishers.

Value of each feature was computed by one of the three approaches; matching of feature's conditions to generate 'Yes', 'No' or 'Unknown' values (example F.2, F.9); identifying the string value answering the feature's question (example F.25); and counting of feature's condition to produce a numeric answer (example F.1, F.11).

As our contributions, we have proposed 12 new features (F.2 – F.4, F.18, F.20, F.21, F.23, F.25, F.27 – F.30) and modified 14 features from previous works (F.1, F.5 – F.7, F.9 – F.12, F.14, F.15, F.22, F.24, F.26 and F.31). The other five features (F.8, F.13, F.16, F.17 and F.19) were adopted from previous studies to improve the overall performances.

4 Experiments and Results

Experiments were designed and conducted using eight common ML classification algorithms (classifiers) to determine optimal performances and the best performing classifier for the prediction. Also, they were aimed at evaluating the overall framework's prediction runtime per webpage to determine its overheads. We used Python v3.4 and Scikit-learn v.019 library to build an application for the experiments in a 64x Windows Home environment.

We collected 9,019 phishing URLs from an online repository of PhishTank, a blacklist of phishing URLs. We also collected 1,733 legitimate URLs from Google and Bing search engines by querying the engines using search keywords such as *sign in* and *login*. For each collected phishing and legitimate URL, we confirmed if it was prompting PII by passing it through the PII webpage filtering module. We also filtered each legitimate URL against the PhishTank's blacklist. For each URL, we downloaded its webpage and extracted features' data to create a training dataset.

Missing values in continuous features were replaced with their respective mean values. *One hot encoding* method was used to convert all 17 categorical data into numeric to ensure that linear functions based classifiers are trained smoothly. All features' data was re-scaled to a mean of 0 and standard deviation of 1 to optimize the classifiers. We oversampled legitimate (minority) class by SMOTE technique to a 1:1 balanced dataset to ensure we get accurate predictions.

One ML algorithm from each of the eight common classes of binary classifiers was evaluated for the selection of the final classifier. These are Logistic Regression (LR), k-Nearest Neighbour (k-NN), Decision Tree (DT), Gaussian Naïve Bayes (GNB), Support Vector Classification (SVC), Multilayer Perceptrons (MLPs), Random Forest (RF) and Gradient Boosting (GB). We used accuracy, precision, recall and AUC as performance metrics for evaluation. For model evaluations, we applied stratified k-fold cross validation method with k=10.

Using feature importance method by RF classifier, prediction influence by weight for each feature was determined and ranked as shown in fig. 2. With all the features, RF performed better across all the metrics compared to other classifiers with an accuracy, precision, recall and AUC of 98.279%, 0.992, 0.987 and 0.997 respectively. After

feature selection, RF achieved the highest accuracy of 98.363% with 20 features (from F.31 to F.8 in fig 2), as summarized in table 2. By performing parameter tuning using Random Search followed by Grid Search methods, the RF achieved an optimum accuracy of 98.45% using the best performing parameters automatically determined by the methods. The classifier achieved false positive (FP), false negative (FN) and classification error of 4.47%, 0.73% and 1.7% respectively.

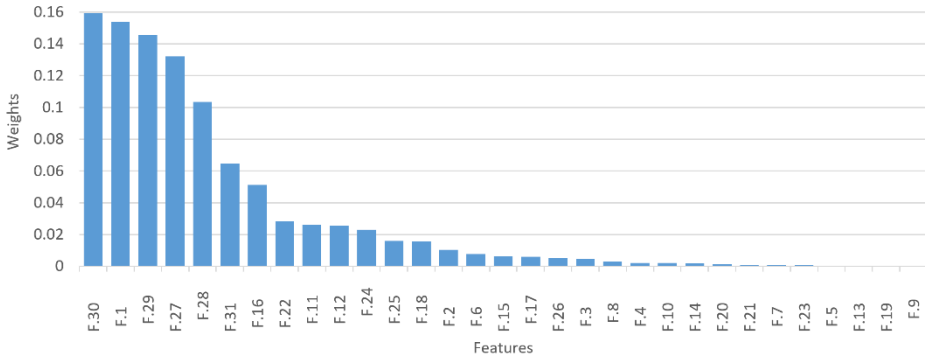

Fig. 2. Distribution of importance weights of the predictive features.

Table 2. Performance scores of all classifiers after applying feature selection.

Classifier	Accuracy (%)	Precision	Recall	AUC
LR	90.234	0.985	0.898	0.962
k-NN	91.239	0.980	0.914	0.947
DT	97.340	0.986	0.983	0.954
GNB	89.137	0.978	0.891	0.938
SVC	91.090	0.984	0.908	0.969
MLP	94.280	0.981	0.951	0.976
RF	98.363	0.994	0.987	0.997
GB	97.638	0.992	0.979	0.995

Each type of features was evaluated individually to determine its performance contributions. Their results in table 3 shows that domain registration and webpage reputation were the best performers across all metrics whereas SSL certificate was the least performer by far. Similarly, various combinations of types of features were evaluated to analyze their performances (see table 4). A combination of webpage structure, URL and domain registration features attained the highest performances while that of domain registration and webpage reputation had the lowest performances.

Table 3. RF performance scores of each type of features.

Type of features	Accur. (%)	Prec.	Recall	AUC
Webpage structure and contents	80.265	0.968	0.818	0.748
URL structure	90.485	0.459	0.660	0.888
Domain registration information	99.972	1.000	1.000	1.000
SSL certificate information	34.998	0.211	0.932	0.596

Webpage reputation	97.089	0.898	0.992	0.996

Performance contributions of our new features were also evaluated relative to the adopted features, as summarized in table 5. Though the new features achieved lesser accuracy than the adopted features, they performed far better in the other metrics and thus contributed significantly to the overall performances of the metrics.

Table 4. RF performance scores of various combinations of types of features.

Combination of Types of Features	Accur. (%)	Prec.	Recall	AUC
Webpage + URL	95.247	0.743	0.680	0.938
Webpage + URL + Domain	99.851	0.998	1.000	0.998
Webpage + URL + Domain + Reputation	91.137	0.849	0.767	0.943
URL + Domain	99.795	0.998	0.999	0.998
URL + Domain + Reputation	89.806	0.819	0.741	0.931
Domain + Reputation	82.943	0.635	0.682	0.861
Domain + Certificate + Reputation	82.980	0.639	0.672	0.861

Table 5. Performance contributions of new and adopted features on the overall performance.

Subset of features	Accur. (%)	Prec.	Recall	AUC
New + modified	90.867	0.839	0.768	0.938
Adopted	99.591	0.374	0.583	0.876

The runtime of each module was measured to evaluate the framework's prediction time per webpage, thus determining the overall overheads. The prediction runtime, as shown in table 6, was 7.63s while training the RF classifier took 16.93s. We also computed an average downloading time for 1,696 legitimate login webpages used in this study and was found to be 0.843s.

Table 6. Runtime for each framework's module.

Module	Runtime (s)
PII webpage filtering	0.02030
Phishing blacklist URL	0.27560
URL redirections check	1.69590
Feature data extraction	5.64000
Prediction analysis	0.00012
Total prediction runtime per webpage	**7.63190**
Training a classifier	16.9300

5 Discussions

FN is the most crucial type of an error for this problem, therefore having a relatively small rate is significant in reducing the risk of misdirecting users to phishing webpages. Compared to some of the related works with very high performances (summarized in table 7), our work compares favorably (in terms of accuracy and FN) against most of them. Other works by [1], [11] and [15], with higher accuracy than ours, did not report

FNs to compare with. However, our work has used different and more diversified features compared to all works, therefore it is more resilient to detection evasion, in addition to a high performance.

Table 7. Comparison of some of the related works with our work (rates in %).

Study	Acc.	FN	FP	Feature Types	Study	Acc.	FN	FP	Feature Types
[1]	98.50	-	1.5	4	[13]	99.65	0.34	0.42	4
[8]	97.98	-	-	1	[14]	97.71	-	1.7	4
[9]	-	0.02	1.17	2	[15]	99.55	-	-	4
[10]	97.30	1.61	4.46	2	[16]	92.54	-	0.41	4
[11]	99.09	-	1.25	2	Ours	98.45	0.73	4.47	5
[12]	92.18	-	-	3					

A slight difference in performances before and after feature selection suggests that all features are collectively effective in the prediction. Although some of the small subsets of features have achieved very high performances, they are still limited with few number of features and diversity of types of features, thus are likely to be vulnerable against detection evasions. A right balance between high prediction performances and resilience to detection evasions should be of high consideration in this problem.

Good performances of the new features suggest that there are many other undiscovered features that are as effective as the previously developed features. This study shows that by combining new features with some of the robust features previously used, new detection models are more likely to yield better performances compared to previous works.

A combined average downloading time for a login webpage and a prediction runtime per webpage, as computed in section 4, is 8.48s. This time is less than the current average downloading time for all types of webpages, which is 8.66s [17]. The average downloading time for login webpages is multiple times lesser than that of other types of webpages due to their light weight design, mostly containing few texts only. We therefore argue that our proposed framework brings an insignificant overhead over the current accepted web browsing speed, thus it is potential for real time deployments.

6 Conclusion

We have proposed a framework of new hybrid features to predict zero hour phishing websites using machine learning. A total of 31 features, 26 of them are novel, from five different types of webpage and third party related features were developed to learn the prediction model. Three webpage pre-processing modules were proposed for the framework to improve prediction performance and efficiency. Features' data were extracted and evaluated using eight machine learning classification algorithms, in which Random Forest achieved an optimal accuracy of 98.45% and false negatives of 0.73%. The framework took 7.63s to predict a new webpage, suggesting that it is promising for real time applications. Further research on new potential features and the use of recent ma-

chine learning methodologies such as deep learning and online learning should be pursued to improve prediction performances and efficiency beyond those of the existing works.

References

1. Lakshmi, V.S., Vijaya, M.: Efficient prediction of phishing websites using supervised learning algorithms. Procedia Engineering 30, 798-805 (2012)
2. PhishLabs, https://info.phishlabs.com/2017-phishing-trends-and-intelligence-report-pti, last accessed January 2017
3. Holz, T., Gorecki, C., Rieck, K., Freiling, F.: Measuring and Detecting Fast-Flux Service Networks. In: Proc. 16th Annual Network & Distributed System Security Symposium (NDSS). NDSS, San Diego, CA (2008)
4. Webroot, https://s3-us-west-1.amazonaws.com/webroot-cms-cdn/8415/0585/3084/Webroot_Quarterly_Threat_Trends_September_2017.pdf, last accessed November 2017
5. Sophos, https://secure2.sophos.com/en-us/medialibrary/Gated-Assets/white-papers/Dont-Take-The-Bait.pdf?la=en, last accessed August 2017
6. Sheng, S., Wardman, B., Warner, G., Cranor, L.F., Hong, J., Zhang, C.: An empirical analysis of phishing blacklists. In: Proc. 6th Conference on Email and Anti-Spam. Mountain View, CA (2009)
7. Gupta B.B, T.A., Jain A., Dharma A.: Fighting against phishing attacks: state of the art and future challenges. Neural Computing and Applications 28, 3629–3654 (2017)
8. Sahingoz, O.K., Buber, E., Demir, O., Diri, B.: Machine learning based phishing detection from URLs. Expert Systems with Applications 117, 345-357 (2019)
9. Zuhair, H., Selamat, A., Salleh, M.: New Hybrid Features for Phish Website Prediction. International Journal of Advances in Soft Computing & Its Applications 8, (2016)
10. Li, Y., Yang, Z., Chen, X., Yuan, H., Liu, W.: A stacking model using URL and HTML features for phishing webpage detection. Future Generation Computer Systems 94, 27-39 (2019)
11. Jain, A.K., Gupta, B.B.: Towards detection of phishing websites on client-side using machine learning based approach. Telecommunication Systems 68, 687-700 (2018)
12. Mohammad, R.M., Thabtah, F., McCluskey, L.: Predicting phishing websites based on self-structuring neural network. Neural Computing and Applications 25, 443-458 (2014)
13. Gowtham, R., Krishnamurthi, I.: A comprehensive and efficacious architecture for detecting phishing webpages. Computers & Security 40, 23-37 (2014)
14. Feng, F., Zhou, Q., Shen, Z., Yang, X., Han, L., Wang, J.: The application of a novel neural network in the detection of phishing websites. Journal of Ambient Intelligence and Humanized Computing 1-15 (2018)
15. Rao, R.S., Pais, A.R.: Detection of phishing websites using an efficient feature-based machine learning framework. Neural Computing and Applications (2018)
16. Xiang, G., Hong, J., Rose, C.P., Cranor, L.: Cantina+: A feature-rich machine learning framework for detecting phishing web sites. ACM Transactions on Information and System Security (TISSEC) 14, 21 (2011)
17. MachMetrics, https://www.machmetrics.com/speed-blog/average-page-load-times-websites-2018/, last accessed February, 2018