Vilnius University
Institute of Mathematics and Informatics
L I T H U A N I A

INFORMATICS ENGINEERING (07 T)

MACHINE LEARNING BASED OPEN SOURCE INTELLIGENCE INFORMATION EXTRACTION AND ANALYSIS METHODS

Paulius Vaitkevičius

October 2019

Technical Report DMSTI-DS-T007-19-08

Vilnius University Institute of Data Science and Digital Technologies,
Akademijos str. 4, Vilnius LT-08663, Lithuania
www.mii.lt

Contents

1 Foreword

Research results presented in this technical report *are directly related* to the research aim and object of the doctoral studies and future dissertation.

Research results presented in this technical report covers objective No 1 of the doctoral studies and future dissertation: the analytical literature review of the related works in Machine Learning and Deep Machine Learning areas, comparing the best algorithms for phishing websites detection, is presented.

Research aim and object of the doctoral studies and future dissertation are introduced further in this section.

RESEARCH OBJECT, AIM, AND OBJECTIVES OF THE DOCTORAL STUDIES

Research object:

1. Machine Learning and Deep Machine Learning algorithms for phishing websites detection.
2. Adversarial Machine Learning algorithms.

Research aim: The research aim is to develop a new method for effective and reliable phishing websites detection, based on Deep Neural Networks and Adversarial Machine Learning algorithms.

Research objectives:

1. Performing literature review, analyzing *state-of-the-art* algorithms for phishing website detection.
2. Replicating the results of *state-of-the-art* algorithms.
3. Proposing new and more effective method for phishing website detection.
4. Creating datasets for new experiments.
5. Conducting experimental research comparing the proposed method with state-of-the-art algorithms.

2 Introduction

Phishing is a form of a cybercrime employing both social engineering and technical trickery to steal sensitive information, such as digital identity data, credit card data, login credentials and other personal data etc. from unsuspecting users by masking as a trustworthy entity. For example, the victim receives an e-mail from an adversary with a threatening message such as possible bank or social media account termination or fake alert on illegal transaction [10], directing him to a fraudulent website that mimics a legitimate one. The adversary can use any information that the victim enters in the phishing website to steal identity or money [26].

Though there are many existing anti-phishing solutions, phishers continue to lure more and more victims. In 2018, the Anti-Phishing Working Group (APWG) reported as many as 785,920 unique phishing websites detected, with a 69.5% increase during the last five years of monitoring, from 463,750 unique phishing websites detected in 2014 [2]. Global losses from phishing activities exceeded 2.7 billion USD in 2018, according to the FBI's Internet Crime Complaint Center [6].

Deceptive phishing attacks are still so successful nowadays because in essence they are "human-to-human" assaults performed by professional adversaries who (i) have financial motivation for their actions, (ii) exploit lack of awareness and computer illiteracy of common Internet users [1], and (iii) manage to learn from their previous experience and improve their future attacks to more successfully lure new victims into visiting new fraudulent websites. For this reason, common Internet users cannot keep up with new trends of phishing attacks and learn to differentiate a legitimate website's URL from a malicious one, relying solely on their own efforts.

In order to protect Internet users from criminal assaults, automated detection techniques for phishing websites recognition were started to develop. The oldest approach included manual blacklisting of known phishing websites' URLs in centralized databases, later used by Internet browsers to alert users about possible threats. The negative aspect of the blacklisting method is that these databases do not cover newly launched phishing websites and therefore do not protect Internet users from "the zero hour" attacks, as the most of phishing URLs are inserted in centralized databases only 12 hours after the first phishing attack [7]. More recent studies have attempted to solve phishing websites detection as a supervised machine learning problem. Many authors have conducted experiments using various classification methods and different phishing datasets with predefined features [4, 14, 16].

Although some scientific papers have described promising results, they are not comparable with each other due to the fact that authors used differently designed datasets and different scientific methods. To the best of our knowledge, no studies comparing classic classification algorithms' performance on all publicly available phishing datasets were conducted.

3 Related works

The scientific community has spent a lot of efforts to tackle the problem of phishing websites detection. In general, approaches to solving this problem can be grouped into three different categories: (i) blacklisting and heuristic based approaches (more in Section 3.1), (ii) supervised machine learning approaches (more in Section 3.2), and deep learning approaches (more in Section 3.3) [16].

3.1 Review of blacklisting and heuristics based research

Although there are initiatives to use a centralized phishing websites' URLs blacklisting solutions (e.g. PhishTank[2], Google Safe Browsing API [3], etc.), this method was proven unsuccessful as it takes time to detect and report a malicious URL, because phishing websites have a very short lifespan (from a few hours to a few days) [24] therefore new phishing websites' URL detection methods were started to implement by the science community.

Heuristic approaches are an improvement on blacklisting techniques where the signatures of common attacks are identified and blacklisted for the future use of Intrusion Detection Systems [18]. Heuristic methods supersede common blacklisting methods as they have better generalization capabilities and have the ability to detect threats in new URLs but they cannot generalize to all types of new threats [24].

3.2 Review of supervised machine learning based research

During the last decade, most of machine learning approaches to solve phishing websites detection problem were based on the supervised machine learning methods on phishing datasets with predefined features. In Table 1 we present a detailed summary of other authors' results of this problem solving during the last 10 years of study. Our review consists of the publication year, authors, used classifier, dataset composition (numbers of phishing and legitimate websites), and achieved classification accuracy. Results are sorted by accuracy from highest to lowest.

From this review, we can make the following observations:

- Two best approaches scored as high as a 99.9% accuracy.
- 15 best approaches scored above 99.0% accuracy.
- The most popular algorithms among researchers are: Random Forest (8 papers), Naïve-Bayes (7 papers), SVM (7 papers), C4.5 (7 papers[4]), Logistic Regression (6 papers).
- Best 5 approaches scored above 99.49% and were implemented using different types of classifiers: neural networks, regression, decision trees, ensembles, and Bayesian. We see no prevailing classification method or type of method among top results.
- Best 5 approaches use highly unbalanced datasets, therefore, evaluating classifier performance by accuracy is inadequate and does not tell how this classifier would perform on more balanced datasets.

[2] https://www.phishtank.com/

[3] https://developers.google.com/safe-browsing/

[4] Including J48, which is WEKA's class for generating pruned or unpruned C4.5 decision tree (http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html)

Table 1: Classification approaches to the solution of the phishing websites detection problem

Year	Authors	Classifier	Dataset # phish.	# legit.	Accuracy
2017	Marchal et al. [12]	Gradient Boosting	100,000	1000	99.90%
2010	Whittaker et al. [26]	Logistic Regression	16,967	1,499,109	99.90%
2011	Xiang et al. [27]	Bayesian Network	8,118	4,780	99.60%
2018	Cui et al. [5]	C4.5	24,520	138,925	99.78%
2013	Zhao et al. [30]	Classic Perceptron	990,000	10,000	99.49%
2018	Patil et al. [15]	Random Forest	26,041	26,041	99.44%
2013	Zhao et al. [30]	Label Efficient Perceptron	990,000	10,000	99.41%
2014	Chen et al. [3]	Logistic Regression	1,945	404	99.40%
2018	Cui et al. [5]	SVM	24,520	138,925	99.39%
2018	Patil et al. [15]	Fast Decision Tree Learner (REPTree)	26,041	26,041	99.19%
2013	Zhao et al. [30]	Cost-sensitive Perceptron	990,000	10,000	99.18%
2018	Patil et al. [15]	CART	26,041	26,041	99.15%
2018	Jain et al. [8]	Random Forest	2,141	1,918	99.09%
2018	Patil et al. [15]	J48	26,041	26,041	99.03%
2015	Verma et al. [25]	J48	11,271	13,274	99.01%
2015	Verma et al. [25]	PART	11,271	13,274	98.98%
2015	Verma et al. [25]	Random Forest	11,271	13,274	98.88%
2018	Shirazi et al. [20]	Gradient Boosting	1,000	1,000	98,78%
2018	Cui et al. [5]	Naïve-Bayes	24,520	138,925	98,72%
2018	Cui et al. [5]	C4.5	356,215	2,953,700	98.70%
2018	Patil et al. [15]	Alternating Decision Tree	26,041	26,041	98.48%
2018	Shirazi et al. [20]	SVM (Linear)	1,000	1,000	98,46%
2018	Shirazi et al. [20]	CART	1,000	1,000	98,42%

Continued on next page

Table 1 – continued from previous page

Year	Authors	Classifier	Dataset # phish.	# legit.	Accuracy
2019	Adebowale et al. [1]	Adaptive Neuro-Fuzzy Inference System	6,843	6,157	98.30%
2016	Vanhoenshoven al. [22]	Random Forest	1,541,000	759,000	98.26%
2018	Jain et al. [8]	Logistic Regression	2,141	1,918	98.25%
2018	Patil et al. [15]	Random Tree	26,041	26,041	98.18%
2018	Shirazi et al. [20]	k-Nearest Neighbors	1,000	1,000	98,05%
2016	Vanhoenshoven et al. [22]	Multi Layer Perceptron	1,541,000	759,000	97.97%
2015	Verma et al. [25]	Logistic Regression	11,271	13,274	97.70%
2018	Jain et al. [8]	Naïve-Bayes	2,141	1,918	97.59%
2016	Vanhoenshoven et al. [22]	k-Nearest Neighbors	1,541,000	759,000	97.54%
2018	Shirazi et al. [20]	SVM (Gaussian)	1,000	1,000	97,42%
2016	Vanhoenshoven et al. [22]	C5.0	1,541,000	759,000	97.40%
2018	Karabatak et al. [9]	Random Forest	6,157	4,898	97.34%
2016	Vanhoenshoven et al. [22]	C4.5	1,541,000	759,000	97.33%
2016	Vanhoenshoven et al. [22]	SVM	1,541,000	759,000	97.11%
2018	Karabatak et al. [9]	Multilayer Perceptron	6,157	4,898	96.90%
2018	Karabatak et al. [9]	Logistic Model Tree (LMT)	6,157	4,898	96.87%
2018	Karabatak et al. [9]	PART	6,157	4,898	96.76%
2018	Karabatak et al. [9]	ID3	6,157	4,898	96.49%
2019	Zhao et al. [29]	Random Forest	40,000	150,000	96.40%
2018	Karabatak et al. [9]	Random Tree	6,157	4,898	96.37%
2019	Chiew et al. [4]	Random Forest	5,000	5,000	96.17%
2018	Jain et al. [8]	SVM	2,141	1,918	96.16%
2016	Vanhoenshoven et al. [22]	Naïve-Bayes	1,541,000	759,000	95.98%

Continued on next page

Table 1 – continued from previous page

Year	Authors	Classifier	Dataset		Accuracy
			# phish.	# legit.	
2018	Shirazi et al. [20]	Naïve-Bayes	1,000	1,000	95,97%
2018	Karabatak et al. [9]	J48	6,157	4,898	95.87%
2009	Ma et al. [11]	Logistic Regression	20,500	15,000	95.50%
2018	Karabatak et al. [9]	JRip	6,157	4,898	95.01%
2014	Marchal et al. [13]	Random Forest	48,009	48,009	94.91%
2015	Verma et al. [25]	SVM	11,271	13,274	94.79%
2019	Chiew et al. [4]	C4.5	5,000	5,000	94.37%
2018	Karabatak et al. [9]	Randomizable Filtered Classifier	6,157	4,898	94.21%
2019	Chiew et al. [4]	JRip	5,000	5,000	94.17%
2019	Chiew et al. [4]	PART	5,000	5,000	94.13%
2017	Zhang et al. [28]	Extreme Learning Machines (ELM)	2,784	3,121	94.04%
2018	Karabatak et al. [9]	Stochastic Gradient Descent	6,157	4,898	93.95%
2018	Karabatak et al. [9]	Naïve-Bayes	6,157	4,898	93.39%
2018	Karabatak et al. [9]	Bayesian Network	6,157	4,898	92.98%
2019	Chiew et al. [4]	SVM	5,000	5,000	92.20%
2011	Thomas et al. [21]	Logistic Regression	500,000	500,000	90.78%
2019	Chiew et al. [4]	Naïve-Bayes	5,000	5,000	84.10%
2015	Verma et al. [25]	Naïve-Bayes	11,271	13,274	83.88%

3.3 Review of deep learning based research

During past few years, novel approaches to solve phishing websites detection problem using deep learning techniques were introduced by scientific community. Zhao et al. have demonstrated that Gated Recurrent Neural Network (GRU) without the need of manual feature creation is capable of classifying malicious URLs with 98.5% accuracy on 240,000 phishing and 150,000 legitimate websites URL samples [29]. Saxe and Berlin have performed an experiment with Convolutional Neural Network (CNN), automating the process of feature design and extraction from generic raw character strings (malicious URLs, file paths, etc.) and gaining 99.30% accuracy on 19,067,879 randomly sampled websites URLs [17]. Vazhayil et al. have performed a comparative study, demonstrating the

98.7% accuracy of CNN and 98.9% accuracy of CNN Long Short-Term Memory (CNN-LSTM) deep learning networks on 116,101 URL samples [23]. Selvaganapathy et al. have implemented a method where feature selection is done using Greedy Multilayer Deep Belief Network (DBN) and binary classification is done using Deep Neural Networks (DNN), capable of classifying malicious URLs with 75.0% accuracy on 17.700 phishing and 10,000 legitimate websites URL samples [19].

References

[1] Adebowale, M., Lwin, K., Sánchez, E., Hossain, M.: Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. Expert Systems with Applications **115**, 300–313 (jan 2019). https://doi.org/10.1016/J.ESWA.2018.07.067, https://www.sciencedirect.com/science/article/pii/S0957417418304925?via{%}3Dihub

[2] Anti-Phishing Working Group, I.: Phishing Activity Trends Reports (2018), https://apwg.org/resources/apwg-reports/

[3] Chen, T.C., Stepan, T., Dick, S., Miller, J.: An Anti-Phishing System Employing Diffused Information. ACM Transactions on Information and System Security **16**(4), 1–31 (apr 2014). https://doi.org/10.1145/2584680, http://dl.acm.org/citation.cfm?doid=2617317.2584680

[4] Chiew, K.L., Tan, C.L., Wong, K., Yong, K.S., Tiong, W.K.: A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences **484**, 153–166 (may 2019). https://doi.org/10.1016/j.ins.2019.01.064, https://www.sciencedirect.com/science/article/pii/S0020025519300763?via{%}3Dihubhttps://linkinghub.elsevier.com/retrieve/pii/S0020025519300763

[5] Cui, B., He, S., Yao, X., Shi, P., Yao, X., He, S., Cui, B.: Malicious URL detection with feature extraction based on machine learning. International Journal of High Performance Computing and Networking **12**(2), 166 (2018). https://doi.org/10.1504/ijhpcn.2018.10015545, http://www.inderscience.com/link.php?id=94367

[6] Internet Crime Complaint Center: 2018 Internet Crime Report. Tech. rep., Internet Crime Complaint Center at the Federal Bureau of Investigation of United States of America (2019), https://www.ic3.gov/media/annualreport/2018{_}IC3Report.pdf

[7] Jain, A.K., Gupta, B.B.: A machine learning based approach for phishing detection using hyperlinks information. Journal of Ambient Intelligence and Humanized Computing pp. 1–14 (apr 2018). https://doi.org/10.1007/s12652-018-0798-z, http://link.springer.com/10.1007/s12652-018-0798-z

[8] Jain, A.K., Gupta, B.B.: Towards detection of phishing websites on client-side using machine learning based approach. Telecommunication Systems **68**(4), 687–700 (aug 2018). https://doi.org/10.1007/s11235-017-0414-0, http://link.springer.com/10.1007/s11235-017-0414-0

[9] Karabatak, M., Mustafa, T.: Performance comparison of classifiers on reduced phishing website dataset. In: 2018 6th International Symposium on Digital Forensic and Security (ISDFS). pp. 1–5. IEEE (mar 2018). https://doi.org/10.1109/ISDFS.2018.8355357, https://ieeexplore.ieee.org/document/8355357/

[10] Lin Tan, C., Leng Chiew, K., Wong, K.S., Nah Sze, S., Tan, C.L., Chiew, K.L., Wong, K.S., Sze, S.N.: PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder. Decision Support Systems **88**, 18–27 (2016). https://doi.org/10.1016/j.dss.2016.05.005, http://dx.doi.org/10.1016/j.dss.2016.05.005

[11] Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Beyond blacklists. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09. p. 1245. ACM Press, New York, New York, USA (2009). https://doi.org/10.1145/1557019.1557153, http://portal.acm.org/citation.cfm?doid=1557019.1557153

[12] Marchal, S., Armano, G., Grondahl, T., Saari, K., Singh, N., Asokan, N.: Off-the-hook: An efficient and usable client-side phishing prevention application. IEEE Transactions on Computers **66**(10), 1717–1733 (2017). https://doi.org/10.1109/TC.2017.2703808

[13] Marchal, S., Francois, J., State, R., Engel, T.: Phish storm: Detecting phishing with streaming analytics. IEEE Transactions on Network and Service Management **11**(4), 458–471 (2014). https://doi.org/10.1109/TNSM.2014.2377295

[14] Marchal, S., Saari, K., Singh, N., Asokan, N.: Know Your Phish: Novel Techniques for Detecting Phishing Sites and Their Targets. In: 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS). pp. 323–333. IEEE (jun 2016). https://doi.org/10.1109/ICDCS.2016.10, http://ieeexplore.ieee.org/document/7536531/

[15] Patil, D.R., Patil, J.B.: Malicious URLs Detection Using Decision Tree Classifiers and Majority Voting Technique. Cybernetics and Information Technologies **18**(1), 11–29 (mar 2018). https://doi.org/10.2478/cait-2018-0002, http://content.sciendo.com/view/journals/cait/18/1/article-p11.xml

[16] Sahoo, D., Liu, C., Hoi, S.C.H.: Malicious URL Detection using Machine Learning: A Survey (jan 2017), http://arxiv.org/abs/1701.07179

[17] Saxe, J., Berlin, K.: eXpose: A character-level convolutional neural network with embeddings for detecting malicious URLs, file paths and registry keys. arXiv preprint arXiv:1702.08568 (feb 2017), http://arxiv.org/abs/1702.08568

[18] Seifert, C., Welch, I., Komisarczuk, P.: Identification of Malicious Web Pages with Static Heuristics. In: 2008 Australasian Telecommunication Networks and Applications Conference. pp. 91–96. IEEE (dec 2008). https://doi.org/10.1109/ATNAC.2008.4783302, http://ieeexplore.ieee.org/document/4783302/

[19] Selvaganapathy, S., Nivaashini, M., Natarajan, H.: Deep belief network based detection and categorization of malicious URLs. Information Security Journal: A Global Perspective **27**(3), 145–161 (may 2018). https://doi.org/10.1080/19393555.2018.1456577, https://www.tandfonline.com/doi/full/10.1080/19393555.2018.1456577

[20] Shirazi, H., Bezawada, B., Ray, I.: "Kn0w Thy Doma1n Name". In: Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies - SACMAT '18. vol. 18, pp. 69–75. ACM Press, New York, New York, USA (2018). https://doi.org/10.1145/3205977.3205992, http://dl.acm.org/citation.cfm?doid=3205977.3205992

[21] Thomas, K., Grier, C., Ma, J., Paxson, V., Song, D.: Design and Evaluation of a Real-Time URL Spam Filtering Service. In: 2011 IEEE Symposium on Security and Privacy. pp. 447–462. IEEE (may 2011). https://doi.org/10.1109/SP.2011.25, http://ieeexplore.ieee.org/document/5958045/

[22] Vanhoenshoven, F., Napoles, G., Falcon, R., Vanhoof, K., Koppen, M.: Detecting malicious URLs using machine learning techniques. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1–8. IEEE (dec 2016). https://doi.org/10.1109/SSCI.2016.7850079, http://ieeexplore.ieee.org/document/7850079/

[23] Vazhayil, A., Vinayakumar, R., Soman, K.: Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks. In: 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). pp. 1–6. IEEE (jul 2018). https://doi.org/10.1109/ICCCNT.2018.8494159, https://ieeexplore.ieee.org/document/8494159/

[24] Verma, R., Das, A.: What's in a URL. In: Proceedings of the 3rd ACM on International Workshop on Security And PrivacyAnalytics - IWSPA '17. pp. 55–63. ACM Press, New York, New York, USA (2017). https://doi.org/10.1145/3041008.3041016, http://dl.acm.org/citation.cfm?doid=3041008.3041016

[25] Verma, R., Dyer, K.: On the Character of Phishing URLs. In: Proceedings of the 5th ACM Conference on Data and Application Security and Pri-

vacy - CODASPY '15. pp. 111–122. ACM Press, New York, New York, USA (2015). https://doi.org/10.1145/2699026.2699115, http://dl.acm.org/citation.cfm?doid=2699026.2699115

[26] Whittaker, C., Ryner, B., Nazif, M.: Large-Scale Automatic Classification of Phishing Pages. The 17th Annual Network and Distributed System Security Symposium (NDSS '10) (2010). https://doi.org/10.1109/TDSC.2013.3, http://www.isoc.org/isoc/conferences/ndss/10/pdf/08.pdf{%}5Cnhttp://research.google.com/pubs/pub35580.html

[27] 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 **14**(2), 1–28 (sep 2011). https://doi.org/10.1145/2019599.2019606, http://dl.acm.org/citation.cfm?doid=2019599.2019606https://www.ml.cmu.edu/research/dap-papers/dap-guang-xiang.pdf

[28] Zhang, W., Jiang, Q., Chen, L., Li, C.: Two-stage ELM for phishing Web pages detection using hybrid features. World Wide Web **20**(4), 797–813 (jul 2017). https://doi.org/10.1007/s11280-016-0418-9, http://link.springer.com/10.1007/s11280-016-0418-9

[29] Zhao, J., Wang, N., Ma, Q., Cheng, Z.: Classifying Malicious URLs Using Gated Recurrent Neural Networks. In: International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. pp. 385–394. Springer (2018)

[30] Zhao, P., Hoi, S.C.: Cost-sensitive online active learning with application to malicious URL detection. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13. p. 919. ACM Press, New York, New York, USA (2013). https://doi.org/10.1145/2487575.2487647, http://dl.acm.org/citation.cfm?doid=2487575.2487647