Research on phishing email detection based on URL parameters using machine learning algorithms

Milda Tubyte and Agne Paulauskaite-Taraseviciene [a]

[a] *Department of Applied Informatics, Kaunas University of Technology, Studentu 50, Kaunas, Lithuania*
milda.tubyte@ktu.edu, agne.paulauskaite-taraseviciene@ktu.lt

Abstract

Abstract – phishing is the most frequent data breach problem in cybersecurity. Cyber scammers use the phishing approach to outwit or obtain sensitive information without a consumer agreement. The victims might receive an email that promotes clicking on the following malicious links that lead to sensitive data leaks. This problem is especially relevant to large companies. Attackers tend to prepare emails that contain work-related information and include familiar keywords in phishing URL[1]s. This paper addresses the URL Boolean classification problem using various machine learning methods such as Support Vector Machine, Random Forest, Decision Tree, Linear Discriminant Analysis, and Logistic Regression. This paper provides a comparative study on these algorithms applied for two different URL classification datasets.

Keywords

Phishing, URL, machine learning, cybersecurity, classification

1. Introduction

The phishing technique still scores as one of the top cybersecurity threats even though it was found in the late 90s. Cybercriminals use this approach because of its simplicity. By creating trustworthy email and encouraging to press provided URL, cybercriminals trick their victim and achieve sensitive data by manipulation. The website might appear as a well-known login form that asks for credentials. Additionally, the website could automatically install malware or inject drive-by exploits in the user's browser. The most targeted companies are Google, Microsoft, Dropbox, PayPal, Apple [1]. By targeting the most popular companies, cybercriminals attempt to gain credentials that would lead them to examine the organization's infrastructure and collect data. Moreover, some consumers choose to use the same credentials to other websites or applications that might open opportunities for attackers to expand criminal activities. Phishing emails and business email compromise cause 67% or more data breaches [2], as stated in the Verizon 2020 data breach investigation report. These numbers provide insight that as organizations moved to SaaS[2] applications, phishing will continue to grow. F5 labs provide vital statistics on phishing patterns. As reported by 2019 research, 85% of tested phishing sites used certificates signed by trusted Certificate Authorities, 71% of phishing sites used HTTPS[3] [3]. It is one of the cybercriminal strategies to outwit consumers into thinking that the website is legitimate. Although, 36% of phishing sites had certificates lasting only 90 days. It might be worth checking the expiration date since criminals do not use SSL[4] certificates for long periods.

Three main phishing target groups can be distinguished: indiscriminate, semi-targeted, and spear phisher [4]. The indiscriminate group usually gets the same content phishing email that often pretends to be a tech brand like Microsoft or Google. Semi-target groups most often are from the same working space. And the last one usually targets C-levels or system admins. Diverse phishing techniques apply to each group which is why phishing email detection is so troublesome. Attackers improve their methods

[1] Uniform Resource Locator
[2] Software as a service
[3] Hypertext Transfer Protocol
[4] Secure Sockets Layer

each month, besides various automation tools increase the processing speed. Machine learning is one of the components that could help to prevent phishing attacks. Combining it with authorization tools and phishing detection courses could minimize the possibility of a successful attack [5]. Since the most common phishing type is receiving a letter with a malicious URL, this paperwork investigates machine learning predictions based on URL features. By extracting URL features from a letter, machine learning algorithms can find the patterns that lead to identifying the URL type. In this research, two datasets were used with five different machine learning techniques. By creating the future extraction code, deployed machine learning models provide a prediction on the selected URL. Experiments include Logistic Regression, Linear Discriminant Analysis, Decision Tree, Support Vector Machine, and Random Forest machine learning methods.

The rest of the paper is organized as follows. Section 2 reviews related work and compares methods that are used in this paper. Section 3 describes used machine learning techniques. Section 4 reveals the results, selected each model parameter, and dataset comparison. Section 5 concludes the paperwork results.

2. Related work

Machine learning algorithms became a prevalent research study for the past years. Recent machine learning studies gave high accuracy results on classifying URLs as phishing and legitimate type [6], [7], [8]. For instance, a study declared 99.7% accuracy with a negligible false positive rate of about 0.06% using a random forest algorithm [9]. Another study based on phishing website detection has implemented the SVM method and reached 95% accuracy using six features only [10]. The study dataset has been created using legitimate URLs from browsing history and phishing URLs from the PhishTank database. However, the study estimates, if the URL does not include all features, the prediction gives wrong results. A. Akinyelu and A. O. Adewumi developed fifteen URL feature extraction code using C# language and performed the Random Forest algorithm. The model overall accuracy reached an impressive 99.7% result with a negligible false positive rate of about 0.06%. [11]

The study Unbiased Phishing Detection Using Domain Name-Based Features by Hossein Shirazi used Rami M. Mohammad, Fadi Thabtah, and Lee McCluskey created rule-based dataset that this paperwork investigates in section [12]. A rule-based threshold becomes a critical point that sets the features as phishing or legitimate type for each URL attribute [13]. Because of strict rules, the model might lose its flexibility. That is the reason why the researcher Hossein Shirazi removed rule-based thresholds and experimented with actual values. The SVM method with the Gaussian kernel parameter provided a 93.7% overall accuracy result. For comparison, the experiments were done on a threshold-based dataset in the experiments section.

3. Research methods

Five machine learning approaches used in this paper are related to the URL classification problem. The target result is to detect if the URL is a legitimate or a phishing type.

Logistic regression is an efficient and powerful algorithm for Boolean classification problems, therefore can be used to classify the URL as phished or legitimate [14]. Different studies have shown promising results while using logistic regression for predicting phishing [15], [16], [17]. The algorithm uses the Sigmoid function that maps result values between 0 and 1. The model requires defining a threshold value that sets a boundary between two classes. Predictions are made based on that boundary.

The Linear Discriminant Analysis (LDA) data classification technique is based on dimension reduction. The method maximizes class separability by reaching the maximum ratio of between-class variance [18]. However, in some cases, alternative and discriminant analysis-based methods can provide higher accuracy results than standard LDA. For example, Biased Discriminant Analysis (BDA) [19], cluster space linear discriminant analysis (CSLDA) [20].

Categorical variable decision tree targets to predict class type values. The model constructs a tree structure that contains a root node that represents entire dataset observations, leaves which are the last

node that does not split, decision nodes which are sub-node that splits into further sub-nodes. Some studies display quite great prediction results using the DT algorithm [21], [8], [22].

Support Vector Machine (SVM) approach for classification problem requires that classes could be separable by a linear boundary. The algorithm is constructed to predict Boolean type targets by determining the separating hyperplane between two classes. The method uses kernel functions for classification [23], [24].

The random forest (RF) algorithm randomizes classification or regression trees, averages calculated predictions, and merges them to achieve more accurate results. Instead of looking for the most important feature in the node, the algorithm searches for the best feature among randomized subsets. The algorithm is widely used in phishing URL classification problems because of its great performance [25], [26], [7].

4. Experiments

The experiments below represent two separate dataset accuracy results applying supervised machine learning algorithms. The first dataset focuses on URL symbol appearance and the second one on the created threshold rules. The experimental results provide a comparison of datasets and insight into each dataset's advantages and drawbacks.

The University of Maribor provided two datasets that contain 111 features of URL specifics [27]. The last dataset feature represents the target *phishing* attribute. The phishing attribute is the Boolean value (0 is legitimate, 1 is phishing URL). The class balance in provided datasets variates. One of the datasets has approximately the same amount of both classes. The second variant distribution is about 30% of phishing and 70 % of legitimate observations. The purpose of different variations is to refer to life conditions were about the same present URL distribution. The datasets examine Domain, Directory, File, Parameter, URL, and external services features (Figure 1). It is not enough to evaluate the symbols that appear in the URL. At first sight, the URL can seem trustworthy, so additional information from external services about the domain and URL is significant. The datasets with unbalanced variation were chosen for the first accuracy experiment.

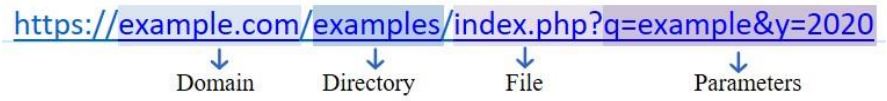

Figure 1: URL feature separation

Researchers Rami M. Mohammad, Fadi Thabtah, and Lee McCluskey created an intelligent rulebased phishing website classification method [13]. In 2012, seventeen features were described with rule-based threshold values that determine values as 1 (legitimate), 0 (suspicious), or -1 (phishing). The final feature represents the URL as a Boolean result (-1 - phishing, 1 - legitimate). The observations were collected from the Phishtank database and yahoo directory. In 2015 researchers added more rules and published a dataset that is still commonly used in scientific experiments. The dataset investigates five groups' attributes based on URL, DNS[5], External statistics, HTML[6], and JavaScript. The final dataset contains 31 features, including targeting URL type. Most features require additional information from external services like the WHOIS or the Alexa databases.

4.1. Accuracy evaluation

In this research, confusion matrix metrics were used to evaluate the performance of included machine learning algorithms. Since phishing URL classification is a two-class Boolean problem, confusion matrix dimensions are M_{2x2}. The confusion matrix's purpose is to display model predicted and actual class values [28]. The correct model provided predictions are True Negative (TN) and True Positive

[5] Domain Name System
[6] Hyper Text Markup Language

(TP) values. Incorrect predictions accordingly are False Negative (FN) and False Positive (FP) values. The following formulas need to be applied to obtain the accuracy (1) and the error rate (2) of the final model.

Table 1
Confusion matrix metrics

	Predicted negative	**Predicted Positive**
Actual negative	TN	FP
Actual positive	FN	TP

$$Accuracy = \frac{TN+TP}{TN+FN+FP+TP}, \quad (1)$$

$$Error = \frac{FP+FN}{TN+FN+FP+TP} \quad (2)$$

4.2. Experimental results

The experiments below represent the results of *.csv* type data. The second dataset format was converted from *.arff* to *.csv*.

The first experiments were done at the University of Maribor provided an unbalanced dataset that contained 88648 observations. A correlation matrix of a dataset was created with a threshold value of 0.7 based on the target column. Twenty-four features were selected with a higher than 70% positive correlation result based on correlation with phishing feature (Figure 2). Features also correlate with each other with a high positive result.

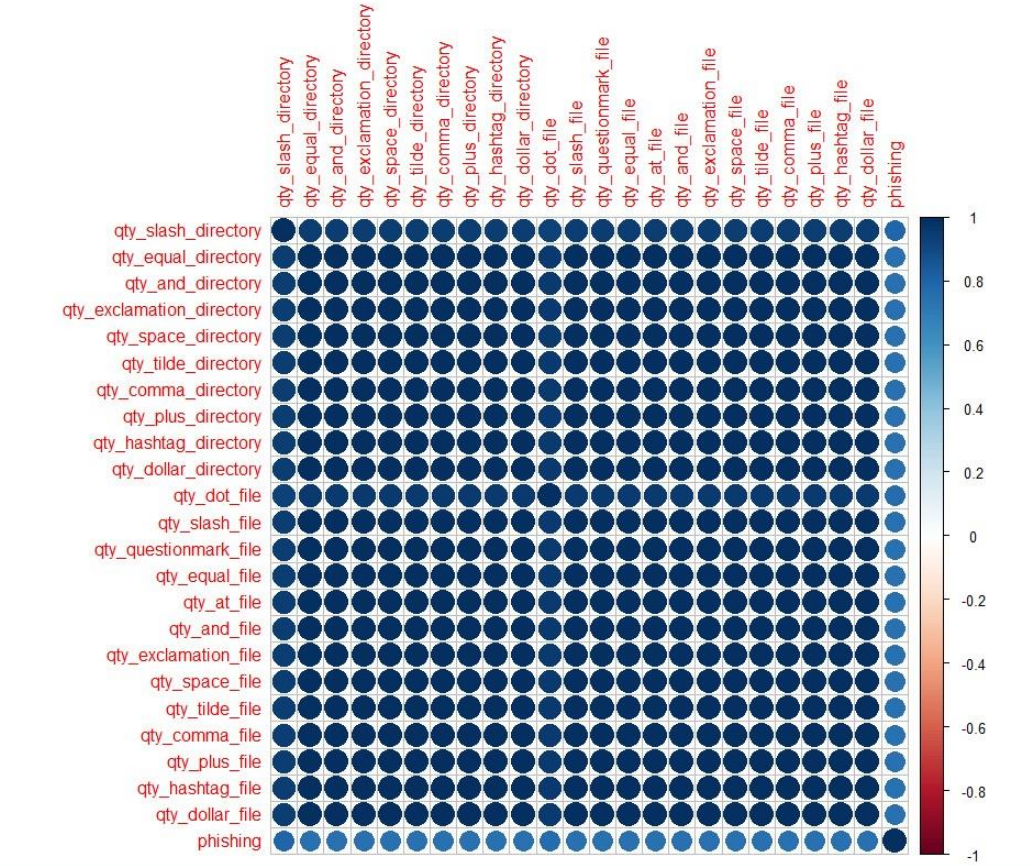

Figure 2: Correlation matrix with 0.7 threshold based on phishing feature.

Below presented tables (2, 3, 4, 5, 6) provide the confusion matrix of every classification algorithm. Table 7 pictures each class and overall model accuracy results. The LR model was implemented as a binomial family model with a threshold value of 0.5. The DT model was implemented as a categorical CART model. The most impactful parameter was detected as directory length. The SVM model provided the best performance with linear kernels (cost 10). The RF algorithm involves 80 iterations to grow trees. The most impactful parameter was indicated as time-domain activation according to the Mean Decrease Accuracy metrics and directory length according to the Mean Decrease Gini metrics. Directory length feature represents how many symbols the URL contains, and time-domain activation investigates how long the domain is active in days. These features are also investigated by the second dataset.

Table 2
Confusion matrix of RF algorithm

RF	Predicted class		
Actual Class	Legitimate	Phishing	Total
Legitimate	16972	428	17400
Phishing	529	8665	9194
Total	17501	9093	26594

Table 3
Confusion matrix of SVM algorithm

SVM	Predicted class		
Actual Class	Legitimate	Phishing	Total
Legitimate	16816	584	17400
Phishing	642	8552	9194
Total	17458	9136	26594

Table 4
Confusion matrix of Decision Tree (DT) algorithm

DT	Predicted class		
Actual Class	Legitimate	Phishing	Total
Legitimate	15774	1626	17400
Phishing	563	8631	9194
Total	16337	10257	26594

Table 5
Confusion matrix of LDA algorithm

LDA	Predicted class		
Actual Class	Legitimate	Phishing	Total
Legitimate	15514	440	17400
Phishing	886	8754	9194
Total	17400	9194	26594

Table 6
Confusion matrix of Logistic Regression (LR) algorithm

LR	Predicted class		
Actual Class	Legitimate	Phishing	Total
Legitimate	16379	1021	17400
Phishing	812	8382	9194
Total	17191	9403	26594

The second dataset contains 11056 observations. The correlation matrix below represented that the most correlated feature with the target value was the SSL final state and URL of Anchor (Figure 3). The URL of the Anchor feature examines if HTML tags and the website have different domain names. The SSL final state feature investigates if the URL uses HTTPS protocol and checks certificate age. The previous dataset also includes the SLL attribute that is called the tls ssl certificate.

Below presented tables (7, 8, 9, 10, 11) provide statistics of each machine learning accuracy result. Random Forest and Support Vector Machine provided the highest overall accuracy results that reach 96%. However, the SVM algorithm predicted each class with almost the same percentage accuracy, and RF predicted the Legitimate class with a higher result than the Phishing class. The decision tree provided 94% overall accuracy result, Linear Discriminant Analysis and Logistic Regression scored with lowest accuracy results 91 % and 92 % accordingly. Comparing the SVM algorithm with Hossein's research results, using the rule-based threshold dataset, the model provided a 3% higher overall accuracy result. However, the threshold method deprives model flexibility as rule boundaries decide feature class. Since phishing trends change over the years, threshold rules might variate also.

The SVM algorithm scored best with kernel option as polynomial, cost value as 50. RF algorithm used 150 tree iterations. The most impactful future was indicated as SSL final state according to Mean Decrease Accuracy and Gini metrics. Decision Tree also detected SLL final state as the most impactful feature. The pruned tree had 51 leaves, and the CP value was selected as 4.37-04. LR performed best with a threshold value of 0.6.

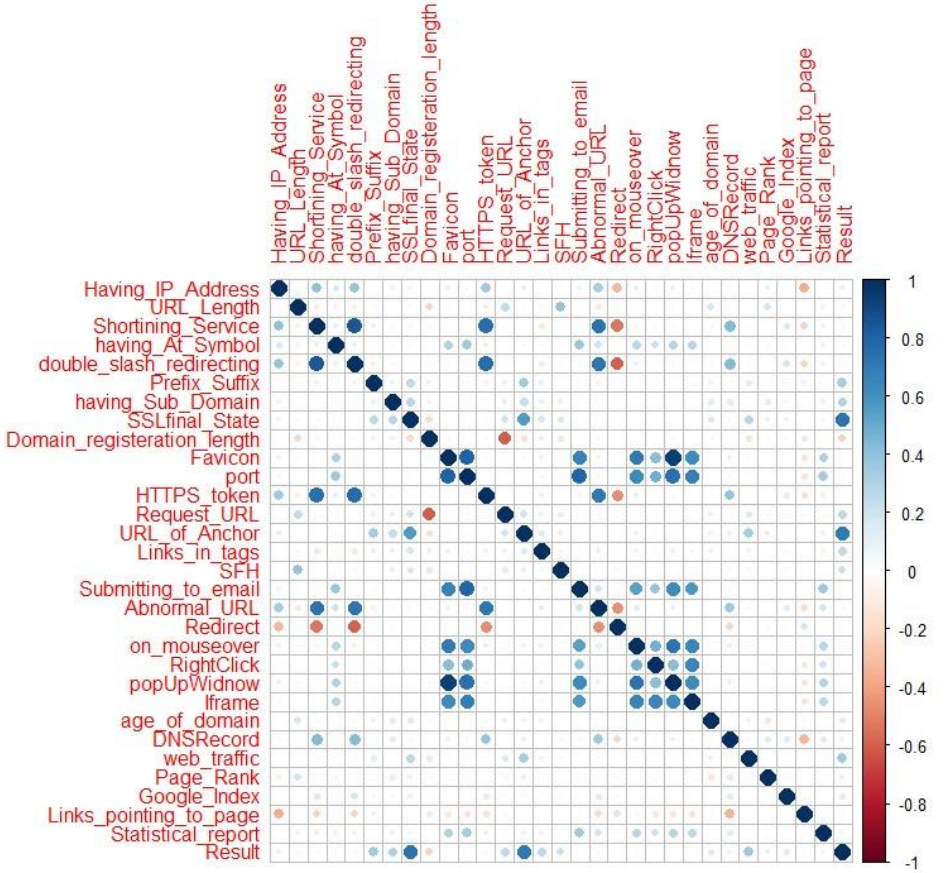

Figure 3: Correlation matrix with 0.7

Table 7
Confusion matrix of RF algorithm

RF	Predicted class		
Actual Class	Legitimate	Phishing	Total
Legitimate	1809	38	1847
Phishing	67	1402	1469
Total	1876	1440	3316

Table 8
Confusion matrix of SVM algorithm

SVM	Predicted class		
Actual Class	Legitimate	Phishing	Total
Legitimate	1788	59	1847
Phishing	59	1410	1469
Total	1847	1469	3316

Table 9
Confusion matrix of Decision Tree (DT) algorithm

DT	Predicted class		
Actual Class	Legitimate	Phishing	Total
Legitimate	1756	91	1847
Phishing	90	1379	1469
Total	1846	1470	3316

Table 10
Confusion matrix of LDA algorithm

LDA	Predicted class		
Actual Class	Legitimate	Phishing	Total
Legitimate	1708	139	1847
Phishing	139	1330	1469
Total	1847	1469	3316

Table 11
Confusion matrix of Logistic Regression (LR) algorithm

LR	Predicted class		
Actual Class	Legitimate	Phishing	Total
Legitimate	1703	144	1847
Phishing	114	1355	1469
Total	1817	1499	3316

Comparing the two dataset results, the overall prediction accuracy is approximately the same. The first dataset contains 3.6 times more features than the second dataset. It would take more time to develop code that would extract all features represented in a first dataset however, only 15 features require additional information from external services. Moreover, the first dataset uses actual values, and it might avoid URL trend change more over the years. The second dataset is based on strict rules and depends on many external recourses that could fail to provide data. Machine learning models compute faster since there are fewer observations and features. It is possible to remove threshold rules and use actual values, but model accuracy might slightly decrease.

Experiments were coded in the R Studio environment using the x86_64-w64-mingw32 platform, 4.0.2 version R. The training was executed on the PC with Intel(R) Core (TM) i5-9600KF CPU @ 3.70GHz 3.70 GHz processor, NVIDIA GeForce GTX 1660 SUPER graphic card 6GB.

The first dataset results reveal that the RF (95%) out-performed DT (91%), LDA (91%), LR (93%), SVM (94%) algorithms with the highest overall accuracy (Figure 4). The DT and LDA algorithm scored with the same 91% overall accuracy result. However, LDA classified the phishing feature with the best score of all provided models with 95% accuracy. The lowest precisely classified legitimate type URL was using LDA (89%) and DT (90%).

The second dataset result shows that RF and SVM scored with the best 96% overall accuracy results (Figure 5). DT (94%) outperformed LDA (91%) and LR (92%). The second dataset scored with 1% higher accuracy using the Random Forest algorithm. Altogether, each model performed with high accuracy results.

Table 12
Percentage accuracy results of included algorithms

ML algorithm	Average accuracy, %	Accuracy results each class	
		Legitimate, %	*Phishing. %*
The first dataset results			
RF	95	97.540	94.246
SVM	94	96.664	93.017
DT	91	90.655	93.876
LDA	91	89.160	95.214
LR	93	94.132	91.168
The second dataset results			
RF	96.8	97.942	95.543
SVM	96.6	96.661	96.304
DT	94	93.873	95.073
LDA	91	91.848	91.442
LR	92	92.239	92.220

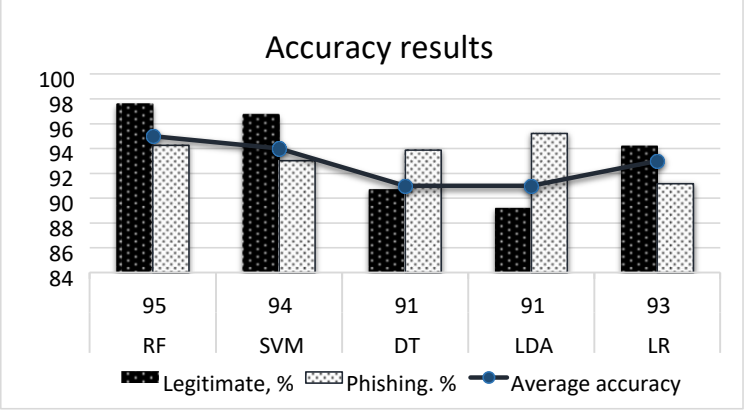

Figure 4: The first dataset accuracy results of five supervised learning algorithms

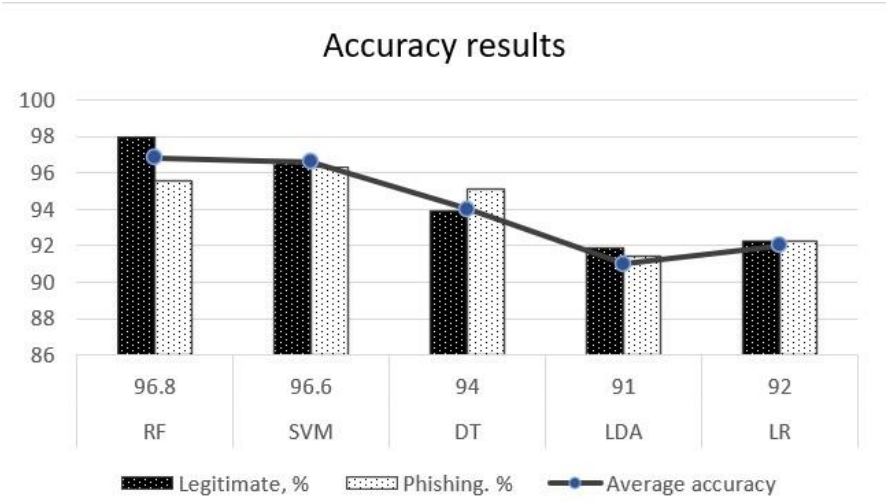

Figure 5: The second dataset accuracy results of five supervised learning algorithms

5. Conclusion

This paperwork represented supervised machine learning algorithm classification on phishing and legitimate type URLs. In this study, phishing URL classification is defined as a two-class problem. Five different algorithms were employed using two datasets: Random Forest, Support Vector Machine, Logistic Regression, Linear Discriminant Analysis, Decision Trees. The accuracy of each model was higher than 91% despite the dataset. The RF algorithm performed the highest overall accuracy result on both datasets. However, LDA classified the Phishing URL class with the highest 95% accuracy rate for the first dataset. The SVM algorithm provided the highest accuracy on classifying Phishing URLs for the second dataset. Each dataset accuracy test was performed on a subset that contained 30% of original dataset observations. The first dataset was published in 2020 and used actual values that mostly involve symbol search in the URL. The second dataset was created in 2015 and has strict rules that determine the threshold value of each feature. The better results were reached using the second dataset however, most features from the dataset require additional information from external services that could fail to provide accurate information. In future research, more effective features (new, derived features from classical URML features) can be included for determining the most relevant signs of malicious URLs.

6. Acknowledgements

I cannot express enough thanks to my lecture Agne Paulauskaite-Taraseviciene that consulted me in every step and helped me to overcome obstacles. Also, my project could not be completed without the Littelfuse Digital Innovation manager and IT security team that introduced me to this relevant problem and told me more about it. My heartful thanks.

7. References

[1] Webroot, "2019 Webroot Threat Report," p. 24, 2019.

[2] P. Langlois, "2020 Data Breach Investigations Report," p. 37, 2020.

[3] D. Warburton, R. Pompon, "2019 Phishing and Fraud Report," p. 10, 2019.

[4] P. N. Mangut, K. A. Datukun, "The Current Phishing Techniques – Perspective of the Nigerian Environment," World Journal of Innovative Research (WJIR), vol. 10, no. 1, pp. 34-44, 2021 .

[5] H. Wechsler. V. Ramanathan, "Phishing detection and impersonated entity discovery using Conditional Random Field and Latent Dirichlet Allocation," 2013.

[6] V. Marcinkevicius, P. Vaitkevicius, "Comparison of Classification Algorithms for Detection of Phishing Websites," Informatica, vol. 31, pp. 143-160, 2020.

[7] W. Ali, "Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection," International Journal of Advanced Computer Science and Applications,, vol. 8, no. 9, 2017.

[8] A. A. Orunsol, "PERFORMANCE COMPARISON OF PREDICTIVE MODELS BASED ON REDUCED PHISHING FEATURE CORPUS," Anale. Seria Informatică, vol. 18, no. 2, 2020.

[9] A. O. Adewumi, A. A. Akinyelu, "Classification of Phishing Email Using Random Forest Machine Learning Technique," Journal of Applied Mathematics, p. 6, 2014.

[10] B. Outtaj, M. Zouina, "A novel lightweight URL phishing detection system using SVM and similarity index," 2017.

[11] A. O. Adewumi, A. A. Akinyelu, "Classification of Phishing Email Using Random Forest Machine Learning Technique," vol. 2014 , 2014 .

[12] H. Shirazi, "Initial Attempt: Fresh-Phish Framework," Fort Collins, Colorado, 2018.

[13] F. Thabtah, L. McCluskey, R. M. Mohammad, "Intelligent rule-based phishing websites classification," Springer-Verlag, 2014.

[14] L. H. Ungar, A. I. Schein, "Active learning for logistic regression: an evaluation," Mach Learn, p. 235–265, 2007.

[15] M. M. Darabi, M. I. Vahid Shahrivari, "Phishing Detection Using Machine Learning Techniques," vol. 10, 2020.

[16] V. Anandkumar, "Malicious-URL Detection using Logistic Regression Technique," International Journal of Engineering Business Management, vol. 9, pp. 108-113, 2019.

[17] M. N. Kumar, M. Sowjanya, G. Kumari, "FAKE WEBSITE DETECTION USING REGRESSION," International Journal of Advance Reasearch in Science and Engineering, vol. 6, no. 8, 2017.

[18] G. A. Montazer, M. Imani, "Phishing Website Detection Using Weighted Feature Line Embedding," The ISC Int'l Journal of, vol. 9, no. 2, pp. 49-61, 2017.

[19] M. F. Moens, J. C. Gomez, "Using Biased Discriminant Analysis for Email Filtering," 2010.

[20] M. Imani, G. A. Montazer, "Email Spam Detection Using Linear Discriminant Analysis Based on Clustering," 2017.

[21] M. Kula, J. Bohacik, "Webpage phishing detection with data mining," Journal of Information, Control and Management Systems, vol. 17, no. 2, 2019.

[22] S. He, X. Y. Shi, B. Cui, "Malicious URL detection with feature extraction based on machine learning," International Journal of High Performance Computing and Networking , vol. 12, 2018.

[23] Y. Zhou, H. Liu, N. Zhu, Shourong Hou, "Wavelet Support Vector Machine Algorithm in Power Analysis Attacks," Shanghai, China, 2017.

[24] H. Mojeed, A. Balogun, A. N. Oluwatobi, V. Adeyemo, "Ensemble-Based Logistic Model Trees for Website Phishing Detection," in Advances in Cyber Security, Springer, 2021, pp. 627-641.

[25] S. Eddie, W. Shou, "Critical Analysis of Current Research Aimed at Improving Detection of Phishing Attacks," Selected Computing Research Papers, vol. 9, 2020.

[26] C. X. S. Nazir, A. Hafeez, S. Wan, S. Khan, "Deep Learning-Based Efficient Model Development for Phishing Detection Using Random Forest and BLSTM Classifiers," 2017.

[27] I. Fister, J. V. Podgoreleca, GregaVrbančič, "Datasets for phishing websites detection," Journal Data in Brief, vol. 33, 2020.

[28] M. B. Chaudhari, Purvi Pujara, "Phishing Website Detection using Machine Learning : A Review," International Journal of Scientific Research in Computer Science, vol. 3, no. 7.