Feature Selections for Phishing URLs Detection Using Combination of Multiple Feature Selection Methods

Abulfaz Hajizada
Dove Science Academy - Tulsa HS, Tulsa, OK, USA
abulfaz.hajizada@gmail.com

Sharmin Jahan
Oklahoma State University, Stillwater, OK, USA
sharmin.jahan@okstate.edu

ABSTRACT

In this internet era, we are very prone to fall under phishing attacks where attackers apply social engineering to persuade and manipulate the user. The core attack target is to steal users' sensitive information or install malicious software to get control over users' devices. Attackers use different approaches to persuade the user. However, one of the common approaches is sending a phishing URL to the user that looks legitimate and difficult to distinguish. Machine learning is a prominent approach used for phishing URLs detection. There are already some established machine learning models available for this purpose. However, the model's performance depends on the appropriate selection of features during model building. In this paper, we combine multiple filter methods for feature selections in a procedural way that allows us to reduce a large number of feature list into a reduced number of the feature list. Then we finally apply the wrapper method to select the features for building our phishing detection model. The result shows that combining multiple feature selection methods improves the model's detection accuracy. Moreover, since we apply the backward feature selection method as our wrapper method on the data set with a reduced number of features, the computational time for backward feature selection gets faster.

KEYWORDS

Phishing, Feature Selection, Correlation, Machine Learning Model

ACM Reference Format:
Abulfaz Hajizada and Sharmin Jahan. 2023. Feature Selections for Phishing URLs Detection Using Combination of Multiple Feature Selection Methods. In *2023 15th International Conference on Machine Learning and Computing (ICMLC 2023), February 17–20, 2023, Zhuhai, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3587716.3587790

1 INTRODUCTION

In this modern internet era, phishing attacks are common internet-based attacks [1]. The phishing attack is a social engineering approach where attackers manipulate human emotions such as fear, greed, or compassion to persuade and gain access to a person's personal, organizational or financial information [2]. The common phishing approach is setting up a URL, which is very difficult to distinguish from a legitimate one. Then persuade people to click the URL that redirects them to malicious websites to steal their personal or sensitive information and/or gain access to their system.

The first reported phishing attack was reported in 1990, and since then, the volume of the attacks has kept rising [2]. In A report by Microsoft security intelligence, phishing attacks were on top of the list among discovered web attacks in 2018, and they predicted it to rise even more [3]. These numbers further prove that attackers apply advanced tactics in phishing attacks, and thus phishing detection is getting the attention of cyber researchers and developers. Determining the tactics used to identify phishing attempts is a major challenge. Attackers continuously improve their tactics, and they can make websites that can shield them from various types of detection. The machine learning approach can effectively determine phishing websites from their URLs' features [1], [4]. However, the challenge is to find out the right set of features that follows a phishing pattern [1], [4], [5]. The effectiveness of feature selection methods comes in handy in identifying the suitable feature set [5], [6]. The work has been done in this area where most researchers selected one or two feature selection methods to extract phishing patterns. However, there is no established agreement on what feature selection method to choose.

This research aims to present an effective feature selection method that combines three filter methods and a wrapper method to find the optimal features for phishing URLs. The reason to apply filter methods is to figure out the correlation of features with the target variable by statistical techniques, which is computationally faster and does not consider any specific machine learning model [1]. However, the problem with the filter method is that it is not clear how to determine the threshold point for rankings. We consider the top 12 features for each filter feature selection method and take the union of three feature sets. The reduced dataset by considering only the features in the union set means we already have chosen the highly correlated features using different statistical correlations. The reduced dataset is prepared to apply the backward feature selection (BFS) method for some specific machine learning model, which is a wrapper method, and apply a heuristic approach to select the optimal set of features to detect phishing URLs. Our proposed approach is computationally faster in selecting features using the BFS method than applying the BFS method to the all-feature dataset. We play with those machine learning models used in BFS to predict the phishing URL and evaluate their accuracy. The result shows that the model's accuracy improved when we chose the union features rather than considering features only from a particular filter method. The result also improves with the features selected by BFS than the features from the union set. We evaluate

ICMLC 2023, February 17–20, 2023, Zhuhai, China

ACM ISBN 978-1-4503-9841-1/23/02…$15.00
https://doi.org/10.1145/3587716.3587790

our feature selection approach using another dataset to detect malware, and the model accuracy result is mostly consistent, like the phishing dataset.

2 BACKGROUND

Phishing attacks generally include various combinations of social engineering and spoofing techniques to persuade users to share their sensitive personal information [7]. As the technologies evolve, attackers use different phishing techniques. Spear phishing attack is an attack that targets a specific group of people with a common interest [8]. Clone phishing is a technique where attackers mimic the webpage of a popular website and collect that user's information [7]. DNS-Based Phishing is an attack that manipulates DNS records to redirect traffic from a legitimate website to malicious site [7]. Email spoofing is an attack when email recipients click on the malicious link in the email and fall into the trap, which leads to installing malware in their system or attacker gaining access to the user's system.

Machine Learning (ML) became a popular tool to detect phishing websites and combat attacks [3]. Basnet et al. [4] apply multiple ML models to detect phishing URLs and conclude that the Random Forest model has the highest accuracy in detecting phishing URLs. In [9], the authors combine K-nearest neighbor (KNN) and support-vector machine (SVM) models to analyze the phishing data, allowing them to use both models' advantages. KNN provides cleanness to the data, while SVM brings effectiveness, and finally, the combined model achieves a 90.04% accuracy. Similarly, Gu et al. [10] combined Naïve Bayes and SVM to detect phishing URLs. The authors claimed that it was a faster approach with a high accuracy rate. It was further proven in the article that deploying this method over 600 phishing URLs yielded high accuracy score in a short amount of time. In [11], a text-based phishing detection model, CANTINA, is developed, which extracts keyword using a frequency-inverse document frequency algorithm and search in google to determine whether the URL is legitimate or not. However, the model produces a lot of false positives. Shahingoz et al. [12] apply Natural Language Processing (NPL) and propose a content-based phishing detection algorithm. The accuracy depends on the efficiency of generating the word vectors mechanism, which converts words into vectors for reaching some crucial features.

ML model's accuracy depends on appropriate feature selection. Feature selection is being used to shrink a high-dimensional dataset into a reduced-dimensional dataset for high-accuracy results with low computational times. In [13], authors use features extracted from URL attributes such as length, number of special characters, directory, domain name, and file name to identify phishing websites. Cai et al. [5] consider multiple feature selection methods in accordance with supervised, unsupervised, and semi-supervised learning models. They apply a two-stage process to combining the filter and wrapper methods to select the optimal features from the high-dimensional dataset. Saeys et al. [14] show the robustness of different feature selection methods and suggest combining filter, wrapper, and embedded methods finds a more stable set of features. In [1], the authors combine multiple feature selection methods to find optimal features to improve the model's accuracy.

3 APPROACH

This paper presents a procedural approach to select features for developing a machine learning model to detect phishing URLs. We combine filter and wrapper methods for the feature selection so that statistical and heuristic techniques are applied to evaluate the relationship between feature variables and the target classification variable. The purpose of combining the filter and wrapper method is that from the filter method, we can compute features' correlation without considering any specific machine learning model, and it is faster to compute. We choose three filter methods for feature selection: 1) Heatmap Correlation, 2) Anova test, and 3) Chi-square test. These three filter methods are chosen because each method uses different statistical techniques to evaluate the target variable's dependency relationship with feature variables. A specific statistical technique may not be effective enough to determine the relationship. Considering multiple filter methods gives us the confidence that the selected features are the right set of features. However, we should find optimal features to improve the model's accuracy. This mindset leads us to apply the heuristic-based wrapper method. The problem with the wrapper method is that it is computationally slow. So, we reduce the dimension of the dataset by only considering the features selected by three filter methods. It allows one to find optimal features for a specific machine learning model but in a faster manner.

The data flow diagram of our approach is shown in Figure 1. We start by splitting our data set with 80-20 into training and test data sets. Then three different feature selection methods are applied to the training data set and ranked based on their correlation values. We choose the top 12 features from each filter method. We consider each top 12 features as a set and take the union of those three feature sets. Taking the union of three feature sets allows us to consider all the features ranked top 12 according to the corresponding statistical technique for those filter methods. As a final feature selection step, the BFS method is applied to the dataset considering only union features for three different machine learning models. Finally, the machine learning models' accuracy in detecting phishing URLs is estimated using the test dataset.

4 RESULT

Our proposed approach is applied to detect phishing websites. We use a dataset from Kaggle [15] that includes 48 features and 10000 rows. PhishTank, Open-Fish, Alexa, and Common Crawl are the sources of data accumulation. Further, the features are divided into three subcategories: i) address bar-based features that cover basic details of URLs such as length and port number, ii) HTML/JavaScript features, which are used for scripting the web page, and iii) Phishing activities that the webpage leads to, such as object download from external domains [3].

There is a column name "labels" in the dataset, which represents the classification; i) legitimate (1), or ii) phishing (0). The dataset is balanced as 5000 rows are for legitimate websites, and the other 5000 rows are for phishing websites. We split the dataset into training and testing datasets in a 4: 1 ratio. We apply three filter methods: i) Heatmap correlation ii) Anova test, and iii) Chi-Square test for feature selections. We choose the top 12 mostly correlated features with the classification column by each feature selection method. The top 12 features and their corresponding correlation value from

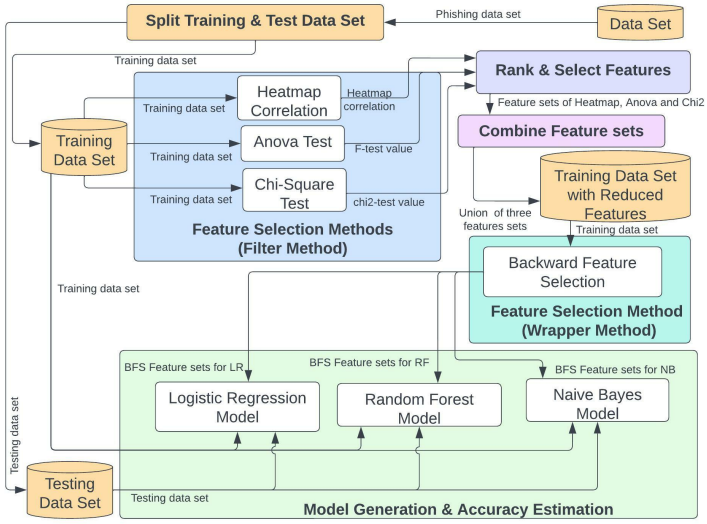


Figure 1: Dataflow diagram of Our Approach

Figure 2: Heatmap Correlation of Top 12 Mostly Correlated Features with Classification Column

the Heatmap correlation are shown in Figure 2. We defined a set of heatmap features, $F_{Heatmap}$ that contains the name of the features selected using Heatmap correlation as below:

$$F_{Heatmap} = \left\{ \begin{array}{l} \text{PctExtNullSelfRedirectHyperlinksRT,} \\ \text{FrequentDomainNameMismatch, NumDash,} \\ \text{SubmitInfoToEmail, PctNullSelfRedirectHyperlinks,} \\ \text{InsecureForms, NumDots,} \\ \text{NumSensitiveWords, PctExtHyperlinks,} \\ \text{IframeOrFrame, PathLevel, AbnormalExtFormActionR} \end{array} \right\}$$

We apply Anova test to determine the top 12 mostly correlated features based on their F-test value as shown in Figure 3. We defined a set of anova features, F_{Anova} that contains the name of the features selected using Anova F-test correlation as below:

$$F_{Anova} = \left\{ \begin{array}{l} \text{HttpsInHostname, PctExtNullSelfRedirectHyperlinksRT,} \\ \text{FrequentDomainNameMismatch, NumDash,} \\ \text{SubmitInfoToEmail, PctNullSelfRedirectHyperlinks,} \\ \text{InsecureForms,} \\ \text{NumDots, NumSensitiveWords, PctExtHyperlinks,} \\ \text{IframeOrFrame, PathLevel} \end{array} \right\}$$

We choose top 12 mostly correlated features from the Chi-square test result as shown in Figure 3. The set of Chi-2 features, F_{Chi2} for

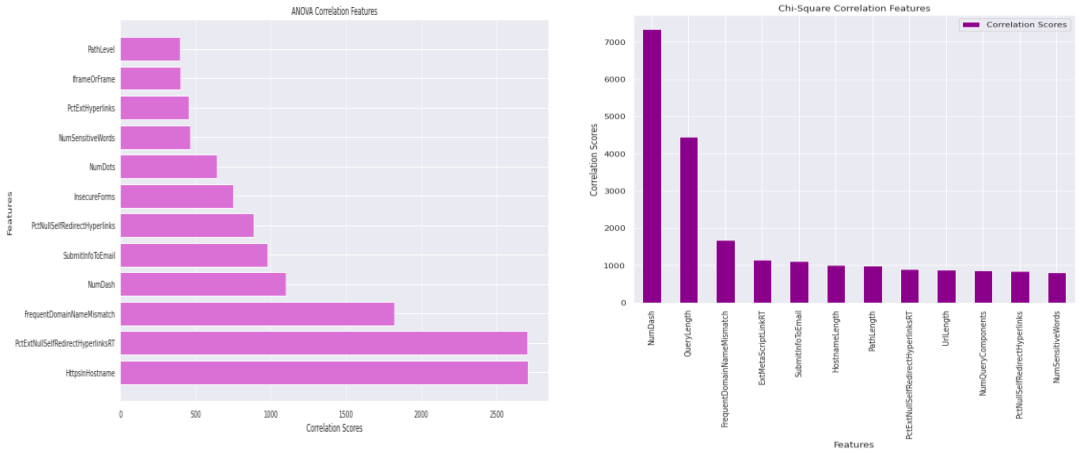


Figure 3: Anova Test (Left) and Chi-Square Test (Right) Correlation of Top 12 Mostly Correlated Features with Classification Column

Table 1: Selected Features by BFS for Three Different Regression Models

Logistic Regression	Random Forest	Naive Bayes
PctExtNullSelfRedirectHyperLinksRT	PctExtNullSelfRedirectHyperLinksRT	PctExtNullSelfRedirectHyperLinksRT
FrequentDomainNameMismatch	FrequentDomainNameMismatch	FrequentDomainNameMismatch
NumDash	NumDash	NumDash
SubmitInfoToEmail	SubmitInfoToEmail	SubmitInfoToEmail
InsecureForms	PctNullSelfRedirectHyperlinks	NumDots
NumSensitiveWords	InsecureForms	InsecureForms
PctExtHyperlinks	NumDots	PctExtHyperlinks
IframeOrFrame	NumSensitiveWords	IframeOrFrame
PathLevel	PctExtHyperlinks	PathLevel
AbnormalExtFormActionR	IframeOrFrame	AbnormalExtFormActionR
ExtMetaScriptLinkRT	PathLevel	ExtMetaScriptLinkRT
UrlLength	AbnormalExtFormActionR	HostnameLength
NumQueryComponents	QueryLength	NumQueryComponents
PctNullSelfRedirectHyperlinks	ExtMetaScriptLinkRT	
	HostnameLength	
	NumQueryComponents	

the features selected using Chi-square test as below:

$$F_{Chi2} = \begin{Bmatrix} \text{NumDash, QueryLength, FrequentDomainNameMismatch,} \\ \text{ExtMetaScriptLinkRT, SubmitInfoToEmail,} \\ \text{HostnameLength, PathLength,} \\ \text{PctExtNullSelfRedirectHyperlinksRT, UrlLength,} \\ \text{NumQueryComponents,} \\ \text{PctNullSelfRedirectHyperlinks, NumSensitiveWords} \end{Bmatrix}$$

Three feature sets contain some common features, and some are different from the other feature sets. So, we take the union of three feature sets that is defined as the union set as: $F_{Union} = F_{Heatmap} \cup F_{Anova} \cup F_{Chi2}$

The union feature set,

$$F_{Union} = \begin{Bmatrix} PctExtNullSelfRedirectHyperLinksRT, \\ FrequentDomainNameMismatch, NumDash, \\ SubmitInfoToEmail, InsecureForms, NumDots, \\ PctExtHyperlinks, IframeOrFrame, PathLevel, \\ AbnormalExtFormActionR, \\ PctNullSelfRedirectHyperlinks, \\ NumSensitiveWords, QueryLength, \\ ExtMetaScriptLinkRT, HostnameLength, \\ PathLength, UrlLength, NumQueryComponents \end{Bmatrix}$$

We create a dataset from the original dataset, only considering the 18 features from the union set. Thus, we reduce the number of features to apply BFS method for final feature selection for three different regression model for phishing detection, as shown in Table 1.

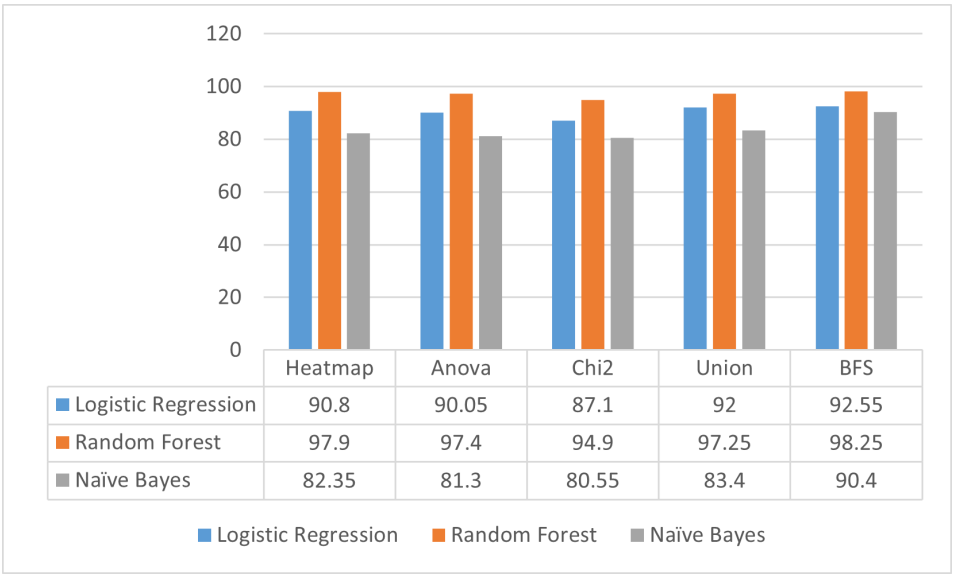


Figure 4: Comparison of detection accuracy of logistic regression, random forest, and naïve baye models for five sets of features

Our first regression model is Logistic regression. We build a logistic regression model for the top 12 features selected by Heatmap, Anova, and Chi-square feature selection methods. Those model's detection accuracy are 90.8%, 90.05% and 87.1% respectively. We then combine the features selected by all three filter methods by applying the union set. We build a logistic regression model for those 18 features from the union. The model's detection accuracy is 92%, more than the model's accuracy built using the features from the individual feature selection method. The BFS method reduces some features from the union set. We build another logistic regression model using the features selected by BFS. The model's accuracy is 92.55%, which is even better than the model of the union feature set. The comparison result of the five logistic regression models' detection accuracy is shown in Figure 4. Similarly, the union feature set have better accuracy in detecting phishing URLs than individual filter method for random forest and naïve bayes models, as shown in Figure 4. The model using the BFS features has more accuracy than individual filter method and union features.

5 EVALUATION

For evaluation purposes, we evaluate our approach in two ways. One is performance evaluation to determine the efficiency of the prediction models and how appropriate feature selection improves the efficiency. Another evaluation is on applying the similar approach on a different dataset showing the effectiveness of our feature selection approach to improve the prediction accuracy to any structured dataset.

5.1 Performance Evaluation

We use the receiver operating characteristic curve (ROC) to evaluate the models' prediction performance. The ROC plot has been frequently used in the literature for performance evaluation. ROC plots the True Positive Rate (TPR) against the False Positive Rate (FPR) to represent how much the model can distinguish between classes. The TPR is also known as recall, which has been defined as

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

The FPR is defined as

$$FPR = 1 - \frac{TN}{TN + FP} \tag{2}$$

Here, TP = The number of phishing classified URLs which are actually phishing URLs
TN = The number of non-phishing classified URLs which are actually non-phishing URLs
FP = The number of phishing classified URLs are actually non-phishing URLs
FN = The number of non-phishing classified URLs are actually phishing URLs

The ROC curve is plotted using TPR in the Y-axis and FPR in X-axis, and when the ROC area is close to 1, the model has a higher chance of classifying the actual phishing and non-phishing URLs. The ROC curves for three different machine-learning models are shown in Figure 5. We plot the ROC curve for five sets of features for each specific machine learning model. The plots show that the prediction efficiency improves when we select features by taking the union of the Top 12 Heatmap, Anova and Chi-square selected features. After that, the prediction efficiency improves by taking the BFS features picked from the union set. Figure 5 shows that ROC curves are consistent with our accuracy result, as mentioned in Section 4.

We also have plotted Precision-Recall Curve (PRC), which captures the tradeoff between precision and recall. Precision is the percentage of phishing classified URLs that are phishing URLs, defined as

$$precision = \frac{TP}{TP + FP} \tag{3}$$

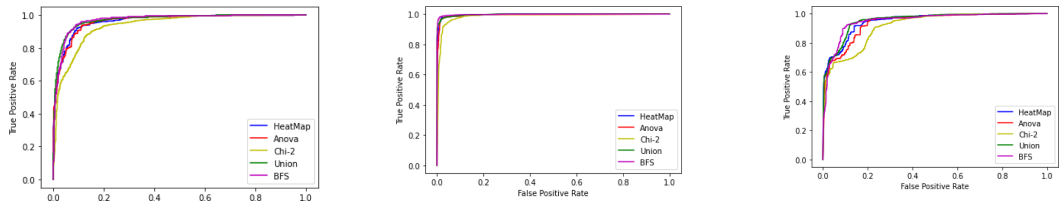


Figure 5: ROC for five sets of features for three machine learning models: Logistic Regression (left), Random Forest (middle) and Naïve Bayes (right)

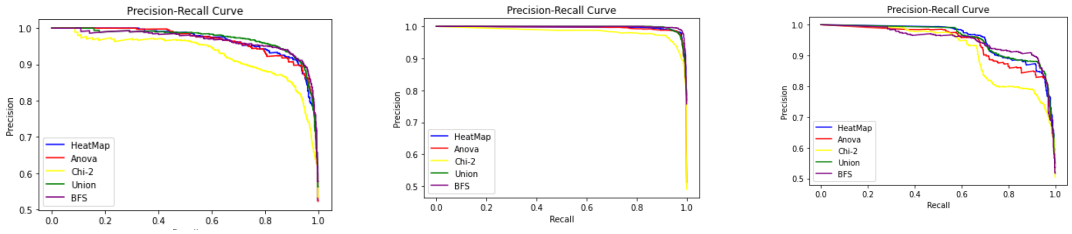
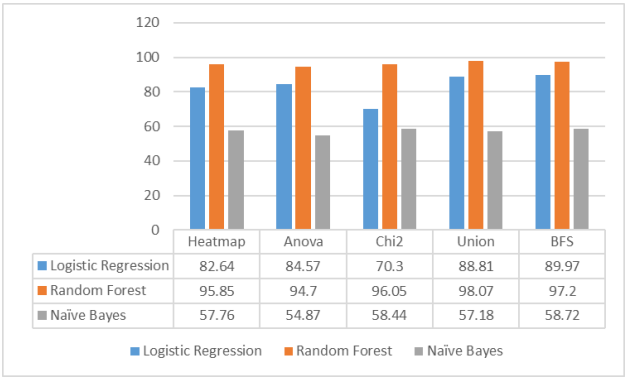


Figure 6: PRC for five sets of features for three machine learning models: Logistic Regression (left), Random Forest (middle) and Naïve Bayes (right)



	Heatmap	Anova	Chi2	Union	BFS
Logistic Regression	82.64	84.57	70.3	88.81	89.97
Random Forest	95.85	94.7	96.05	98.07	97.2
Naïve Bayes	57.76	54.87	58.44	57.18	58.72

Figure 7: Comparison of detection accuracy of logistic regression, random forest, and naïve baye models for five sets of features from malware dataset

Recall is TPR which is the percentage of phishing URLs that are predicted phishing URLs. If the PRC curve has a high area under it, the prediction model has both high recall and high precision. In other words, high precision means a low false positive rate, and high recall expresses a low false negative rate, which can be interpreted as the model providing high accuracy results. The PRC curves for five sets of selected features for each specific machine learning model are shown in Figure 6, consistent to our accuracy result as mentioned in Section 4.

5.2 Effectiveness evaluation

We choose the malware detection dataset [16] to evaluate the approach's effectiveness in improving the efficiency of another dataset. That provides the confidence that the approach applies to any structured dataset in prediction modeling. The dataset contains 55 features of malware, and the total sample size is 5184. We follow our approach for feature selection and play with the same three ML modeling algorithms. The result shows that almost consistent with the phishing dataset. For logistic regression and the naïve bayes model, the models' accuracy improves with the union feature set. Models using BFS features have improved accuracy than models with union features. However, the random forest model has some discrepancies from the expectation, as shown in Figure 7. We suggest that the model developer can decide on choosing the features from the five feature sets as per their preferences.

6 CONCLUSION

The paper presents an approach for selecting appropriate features for the ML model to detect phishing URLs. We combine multiple filter and wrapper methods in a procedural way to choose high correlated features but in a faster computational manner. We develop three ML models commonly used for classification choosing the five sets of features from the heatmap, Anova, and chi-square test, the union of these three, and BFS methods. The result shows that the model's accuracy improves when we choose the optimal features from the final BFS method. Our limitation is that we consider a balanced phishing dataset, which is prone to overfitting. Moreover, we prefer these three filter methods and a wrapper method without having any particular justification. We just follow the other researchers' work and pick these methods since they are widely used. In the future, we will investigate the reasoning behind our choice of feature selection methods.

ACKNOWLEDGMENTS

This research is sponsored by NSF RET Grant: Research Experiences in Big Data and Machine/Deep Learning for Oklahoma STEM, grant number 2055557.

REFERENCES

[1] Ram Basnet, Andrew H. Sung, and Qingzhong Liu. 2012. Feature selection for improved phishing detection. *Advanced Research in Applied Artificial Intelligence*, 252–261. https://doi.org/10.1007/978-3-642-31087-4_27~

[2] Zainab Alkhalil, Chaminda Hewage, Liqaa Nawaf, and Imtiaz Khan. 2021. Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. Cardiff Metropolitan University. Journal contribution. https://hdl.handle.net/10779/cardiffmet.16988479.v1

[3] Mohammad Almseidin, Almaha Abuzuraiq, Mouhammd Alkasassbeh, and Nidal Alnidami. 2019. Phishing detection based on machine learning and feature selection methods. *International Journal of Interactive Mobile Technologies (IJIM)*, *13*(12). https://doi.org/10.3991/ijim.v13i12.11411~

[4] Ram Basnet, and Tenzin Doleck. 2015. Towards developing a tool to detect phishing urls: a machine learning approach. In 2015 IEEE International Conference on Computational Intelligence & Communication Technology, 220–223. https://doi.org/10.1109/cict.2015.63

[5] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. 2018. Feature selection in Machine Learning: A new perspective. *Neurocomputing*, *300*, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077~

[6] Girish Chandrashekar, and Ferat Sahin. 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024~

[7] M. Nazreen Banu, and S. Munawara Banu. 2013. A Comprehensive Study of Phishing Attacks. *International Journal of Computer Science and Information Technologies*, 783–786.

[8] Bimal Parmar. 2012. Protecting against Spear-Phishing. *Computer Fraud & Security*. 8–11. https://doi.org/10.1016/s1361-3723(12)70007-6

[9] Altyeb Altaher. 2017. Phishing websites classification using hybrid SVM and KNN approach. International Journal of Advanced Computer Science and Applications, 8(6).

[10] Xiaoqing GU, Hongyuan WANG, and Tongguang NI. 2013. An efficient approach to detecting phishing web. *Journal of Computational Information Systems*, *9*(14), 5553–5560.

[11] Yue Zhang, Hong I. Jason, and Cranor F. Lorrie. 2007. Cantina: a content-based approach to detecting phishing web sites. Proceedings of the 16th international conference on World Wide Web.

[12] Ozgur K. Sahingoz, Ebubekir Buber, Onder Demir, and Banu Diri. 2019. Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345-357.

[13] Anh Le, Athina Markopoulou, and Michalis Faloutsos. 2011. PhishDef: URL names say it all. 2011 Proceedings IEEE INFOCOM, 2011, pp. 191-195, doi: 10.1109/INFCOM.2011.5934995.

[14] Yvan Saeys, Thomas Abeel, and Yves Van de Peer. 2008. Robust feature selection using ensemble feature selection techniques. *Machine Learning and Knowledge Discovery in Databases*, 313–325. https://doi.org/10.1007/978-3-540-87481-2_21~

[15] Pishing Detection Using Machine Learning. Available at: https://www.kaggle.com/code/fadilparves/pishing-detection-using-machine-learning/notebook

[16] Classification of Malwares (CLaMP). Available at: https://www.kaggle.com/datasets/saurabhshahane/classification-of-malwares