[1]*Dr Amit S. Savyanavar
[2]Pradnya Sankpal
[3]Dr Nikhil C. Mhala

Phishing Webpage Detection using Feature Selection Methods

Journal of Electrical Systems

Abstract: - Phishing attacks are rapidly taking place around the globe. This makes it vital to have efficient phishing detection methods in place. All the datasets that are available are voluminous generally with a vast number of features. Furthermore, many of the features present are redundant or irrelevant and don't substantially help in determining the final outcome. Therefore, it is necessary to identify those features and eliminate them to help reduce resources & time. This paper proposes two phishing detection techniques wherein one method incorporates ensemble feature reduction method and the other incorporates a feature reduction method based on average weight which help in eliminating irrelevant features and making a compact subset of the features to identify phishing attacks. These two methods are based on correlation, chi square, gain ratio, and information gain. The system uses Random Forest classifier which outperforms the rest of the classifiers. The comparison between both the methods is provided and the best method is determined taking factors like accuracy and computational time into consideration. The Phishing Webpage dataset is taken from Mendeley data.

Keywords: Phishing, Phishing Detection, Feature Selection, Ensemble Feature Selection.

I. INTRODUCTION

In an age where everything is digitalized, where the primary means of communication is the internet, cybercrimes are prevalent in the society and are rising rapidly. Among these, phishing is one of the most damaging and adverse attacks. Phishing mainly involves tricking users and snatching away their sensitive information deceitfully. There are different types of phishing attacks namely email phishing, smishing and vishing, whaling, etc. Phishing does happen because of technical vulnerabilities, but along with that, there is some sort of psychological manipulation involved as well. It is necessary to stay vigilant and careful but along with that, there is an alarming need to detect phishing attacks and prevent financial losses, data breaches, identity thefts, etc.

Machine learning along with feature selection techniques provide an efficient way to detect phishing attacks.

The major contributions of this work are:

•This paper proposes two feature reduction techniques which use filter techniques like gain ratio, correlation, chi square, and information gain.

•The reduced set of features obtained is then used to classify phishing attacks using Random Forest classifier.

•Additionally, the two feature reduction techniques are compared on the Mendeley dataset.

II. LITERATURE SURVEY

Deepak Kshirsagar, Deepak Kumar, [1], propose an ensemble feature reduction method for the detection of web attacks which is used to reduce features with the help of filter methods like Information Gain, Correlation, Gain Ratio, Chi Square, and ReliefF. They use J48 classifier in this study for improvised accuracy. The method proposed by them is implemented in this paper for phishing webpage detection.

Deepak Kshirsagar, Deepak Kumar, [2], propose a feature reduction technique which is based on average weight method. This method is used for the detection of DoS attacks. It uses Information Gain, Correlation, and ReliefF. The features are reduced with the help of this method and accuracy is improvised. The method proposed by them is implemented in this paper with a few changes for phishing webpage detection.

Khonji et al[3] extensively study the literature that is based on phishing attacks. This paper surveys various phishing mitigation techniques of machine learning. It also reviews various anti-phishing software techniques. The authors imply in their paper that user education or training is also important and this reduces the susceptibility to phishing attacks.

[1] *Corresponding author: Department of Computer Engineering and Technology, Dr Vishwanath Karad MIT World Peace University, Pune, India, *amitsavyanavar@gmail.com
[2,3] Department of Computer Engineering and Technology, Dr Vishwanath Karad MIT World Peace University, Pune, India

S. Eftimie, et al., [4] presents a study which investigates the impact of personality traits of users in the contect of spear phishing attacks. Before conducting Phishing campaigns, personality tests, cybersecurity courses, the results were aggravated. However, after all these effects, there was a reduction in people falling prey to such attacks.

R. Valecha, et al., [5], assess the effectiveness of using persuasion cues in detection of phishing emails. The focus is on gain and loss persuasion cues. Three machine learning models are created with loss persuasion cues, gain persuasion cues, and combined loss and gain persuasion cues. The drawbacks of this research are that that there could be information loss during analysis and coding.

Yijun Xia, et al..,[6], propose an attributed ego graph embedding framework on Ethereum which is used to differentiate phishing accounts. Furthermore, to make this model more suitable, a transaction attribute based strategy, number, and transaction directions were designed.

M. Chatterjee and A.-S. Namin [7] introduce an approach which is novel and based on deep reinforcement learning. The proposed model adapts to dynamic behaviour of phishing webites and learns the features which are associated with those phishing websites. The work is not optimized for real world implementation.

S. Mahdavifar and A. A. Ghorbani [8] propose DeNNeS: deep embedded neural network expert system, this system extracts rules from a trained deep neural network architecture. It is evaluated on two datasets. It is concluded that DeNNeS outerforms standalone DNN, JRip and other various machine learning algorithms. It also outperforms KNN, SVM, Random Forest. More development is needed to refine the extracted rules of the model.

III. METHODOLOGY

This paper presents two methods for phishing detection which make use of feature reduction techniques.

A. Method one

The first method uses ranking with filter techniques. It consists of Set of Feature Occurrence (SFO) and Reduced Feature Set (RFS). Figure 1 shows the representation of the first method.

The dataset is consistent as it is already preprocessed.

For feature selection, filter techniques like Chi square, Information Gain (IG), Gain Ratio (GR), Correlation (CR) are used with ranking and applied on the dataset. The filter techniques calculate the weight of each feature in a descending order, along with the ranking. From each filter technique and based on their ranks, five sets are created by taking 21 features from the total number of features, i. e. top 25% of the total number of features. Examination of the occurrence of features in each set is done. This obtains four Sets of Feature Occurrence (SFO) and here, ranking is not considered. If the occurrence of a particular feature is there in at least one set out of the four sets, it is included in SFO1. If the occurrence of a particular feature is there in at least two sets out of the four sets, it is included in SFO2. If the occurrence of a particular feature is there in at least three sets out of the four sets, it is included in SFO3. If the occurrence of a particular feature is there in at least four sets out of the four sets, it is included in SFO4. The four sets (Selection of Frequency sets) along with the sets of features obtained through Chi square, Information Gain (IG), Gain Ratio (GR), and Correlation (CR) are provided to Random Forest classifier to obtain results for phishing detection. Eventually, the set which produces the best results is selected.

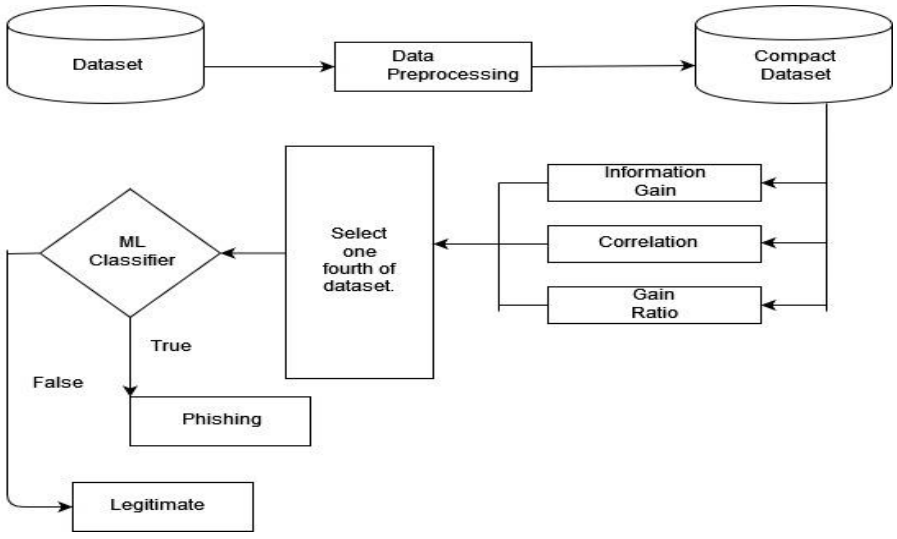

Fig 1: Ensemble Feature Reduction technique.

B. Method two

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations. The second method also uses Information Gain (IG), Correlation (CR), Gain Ratio (GR). These methods used calculate the weight of each feature and a score is assigned to each one of them based on statistical measures. Figure 2 shows the representation of the second method.

The average weight which is obtained is used as a benchmark and on that basis, the features which are equal to or greater than the benchmark average weight are selected. The new set of features pertaining to Correlation (CR), Information Gain (IG), Gain Ratio (GR) are named as CR-1, IG-1, GR-1 respectively. Furthermore, three more New Sets (NS-1, NS-2, NS-3) are obtained. These three new sets observe the feature occurrence in the above sets of CR-1, IG-1, GR-1. If a feature is present in at least one of the sets of CR-1, IG-1, GR-1, it is included in NS-1. Similarly, if a feature is present in at least two and three of the sets of CR-1, IG-1, GR-1, it is included in NS-2 and NS-3 respectively. All these 6 sets are tested on Random Forest classifier.

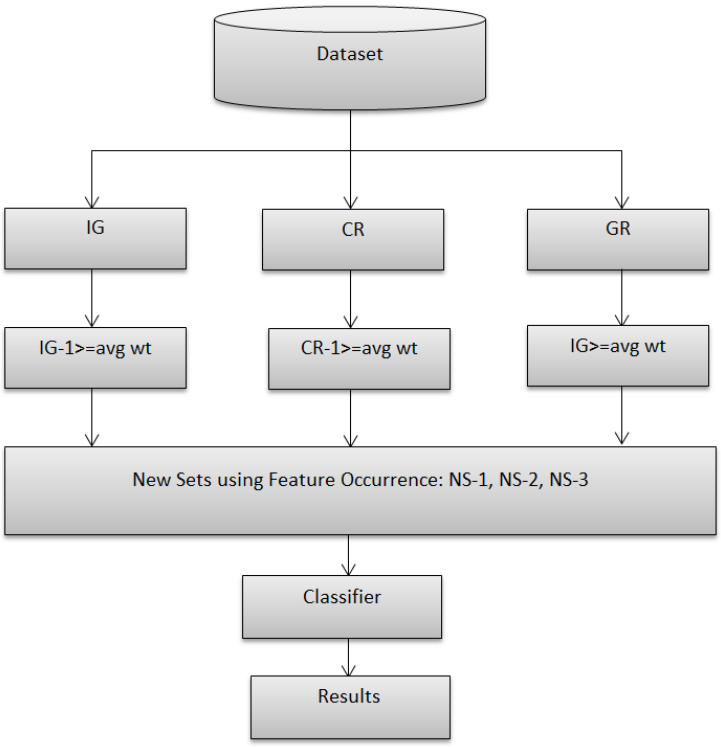

Fig 2: Feature Reduction Method based on Average Weight

IV. RESULTS

Weka, which is an open source tool, is widely used for conducting machine learning experiments. This experiment is performed on Intel(R) Core(TM) i3-8130U CPU @ 2.20GHz with 12 GB of RAM. The proposed method is tested on the 2020 Mendeley dataset. It is balanced, it contains exactly 50% legitimate and 50% Phishing URLs. It has 11430 instances with 87 attributes.

Method One: The filter based feature selection methods such as Information Gain, Correlation, Gain ratio, Chi Square are applied to the dataset for selection of features with ranking. Out of the total features, the top 25% were considered, that is 21 features out of 87 were considered. Therefore, for each individual filter technique, the top 21 features were selected.

Methods	Feature Numbers
Information Gain(IG)	86, 84, 57, 87, 83, 59, 58, 75, 82, 47, 21, 63, 68, 26, 51, 41, 45, 43, 50, 1, 49
Gain Ratio (GR)	86, 21, 6, 18, 3, 31, 84, 51, 7, 79, 8, 53, 87, 39, 27, 10, 75, 32, 85, 19, 54
Correlation (CR)	86, 87, 21, 26, 79, 57, 51, 83, 3, 7, 1, 58, 14, 2, 10, 27, 43, 34, 47, 31, 78
Chi Square	86, 84, 57, 87, 83, 59, 58, 75, 21, 63, 82, 47, 68, 26, 41, 43, 51, 45, 50, 65, 79

Table 1: Top 25% features corresponding to the filter techniques with ranking

Sets	Feature Numbers
SFO1	86, 84, 57, 87, 83, 59, 58, 75, 82, 47, 21, 63, 68, 26, 51, 41, 45, 43, 50, 1, 49, 6, 18, 3, 31, 7, 79, 8, 53, 39, 27, 10, 32, 85, 19, 54, 14, 2, 34, 65
SFO2	86, 84, 57, 87, 83, 59, 58, 75, 82, 47, 21, 63, 68, 26, 41, 51, 45, 43, 50, 1, 10, 3, 31, 7, 79
SFO3	86, 84, 57, 87, 83, 58, 75, 47, 21, 26, 51, 43, 79
SFO3	86, 87, 21, 51

Table 2: Sets of Feature Occurrence

For the formation of four subsets, without considering ranks of features, the strategy of Set of Feature Occurrence was applied. All the SFO sets were experimented and evaluated on classifiers like J48, Naïve Bayes, and Random Forest. Out of all these classifiers, Random Forest outperformed the rest in terms of accuracy. Random Forest was first applied on all features and it achieved an accuracy of 95.045% with 7.37 seconds for building the model. Further experimentation was performed with Random Forest classifier where 10-fold cross validation with the SFO sets of the Mendeley dataset. After applying Random Forest to all of all the methods and the SFO sets, results were obtained and are shown in the below table.

Methods	No. of Features	Accuracy (%)	Time (Seconds)
-	87	95.045	7.37
Information Gain(IG)	21	95.923	3.56
Gain Ratio (GR)	21	94.628	3.22
Correlation (CR)	21	96.019	3.31
Chi Square	21	95.844	3.42
SFO1	40	96.194	4.05
SFO2	25	96.281	3.07
SFO3	13	95.765	2.56
SFO4	4	92.021	1.35

Table 3: Results of Ensemble Feature Reduction Method using Random Forest

The proposed method, ensemble feature reduction with Set of Feature Occurrence 2 (SFO2) outperforms the rest as shown in the above table and achieves an accuracy of 96.281% with a model build time of 3.07 seconds.

b. Method Two: This method like discussed above incorporates a feature reduction technique which is based on average weight. Information Gain was first applied to the dataset and corresponding weights of each feature were calculated. Similarly, Correlation and Gain Ratio were also applied to the dataset and the corresponding weights were calculated. The average weight for Information Gain was calculated. Likewise, the average weights for Correlation and Gain Ratio were also calculated. The average weight for IG, CR, GR that we got was 0.065, 0.138, 0.057 respectively. The IG-1, CR-1, GR-1 sets were formed. With the help of these sets, NS-1, NS-2, NS-3 were created by examining the feature occurrence in at least one, two, and three sets respectively as discussed above. And on these sets, Random Forest classifier was applied to obtain results. Results are demonstrated in the below tables.

Methods	Feature Numbers
Information Gain-1 (IG-1)	48, 27, 42, 7, 10, 71, 3, 70, 79, 4, 65, 2, 49, 1, 50, 43, 45, 41, 51, 26, 68, 63, 21, 47, 82, 75, 58, 59, 83, 87, 57, 84, 86
Correlation- 1 (CR-1)	71, 6, 56, 67, 63, 22, 82, 48, 8, 80, 75, 68, 40, 70, 49, 50, 45, 4, 78, 31, 47, 34, 43, 27, 10, 2, 14, 58, 1, 7, 3, 83, 51, 57, 79, 26, 21, 87, 86
Gain Ratio- 1 (GR-1)	63, 43, 78, 16, 83, 65, 28, 55, 17, 58, 26, 23, 57, 56, 59, 15, 54, 19, 85, 32, 75, 10, 27, 39, 87, 53, 8, 79, 7, 51, 84, 31, 3, 18, 6, 21, 86

Table 4: Features based on Average Weight

Sets	Feature Numbers
NS-1	48, 27, 42, 7, 10, 71, 3, 70, 79, 4, 65, 2, 49, 1, 50, 43, 45, 41, 51, 26, 68, 63, 21, 47, 82, 75, 58, 59, 83, 87, 57, 84, 86, 6, 56, 67, 22, 8, 80, 40, 78, 31, 34, 14, 16
NS-2	48, 27, 7, 10, 71, 3, 70, 79, 4, 65, 2, 49, 1, 50, 4, 43, 51, 21, 26, 63, 68, 47, 82, 75, 59, 58, 83, 87, 86, 57, 56, 8, 78, 31
NS-3	827, 10, 7, 79, 51, 43, 26, 63, 21, 75, 87, 57, 86

Table 5: New Sets of Feature Occurrence

Methods	No. of Features	Accuracy (%)	Time (Seconds)
Information Gain- 1(IG-1)	33	95.923	3.94
Correlation- 1 (CR-1)	39	94.628	3.86
Gain Ratio (GR-1)	37	96.019	4.06
NS-1	58	96.570	4.33
NS-2	34	96.045	3.97
NS-3	13	95.188	2.29

Table 6: Results of Feature Reduction Method based on Average Weight using Random Forest.

It is observed that out of all the sets above, NS-1 outperforms the rest with an accuracy of 96.570 and a model build time of 4.33 seconds.

Now, in the ensemble feature reduction method, SFO2 outperforms the rest with an accuracy of 96.281% and model build time of 3.07 seconds with a total of 25 features(refer table 3). And, in the feature reduction method which is based on average weight, NS-1 outperforms the rest with an accuracy of 96.570% and model build time of 4.33 seconds with a total of 58 features(refer table 6). A comparative analysis in table format is provided below:

Methods	No. of Features	Accuracy(%)	Time(Seconds)
Ensemble Feature Reduction	25	96.281	3.07
Feature Reduction based on Average Weight	58	96.570	4.33

Table 7: Comparison of both the methods

V. CONCLUSION

This paper proposes two Feature Reduction methods for Phishing Webpage Detection. Both the methods produce higher accuracies, except there is a difference between the model build time and the features used. The ensemble feature reduction method provides a high accuracy of 96.281% with a reduced feature number of 25 out of 87 and a model build time of 3.07 seconds. The Feature Reduction method based on average weight produces a high accuracy with 96.570%, however the features used are 58, which are greater than the features used in the previous method, and a model build time of 4.33 seconds. When we compare the method used which doesn't incorporate feature reduction, the accuracy obtained is 95.045%, with a model build time of 7.37 seconds and with the entire 87 features. The implemented methods have produced promising results by increasing the accuracy,

reducing the model build time, and reducing the number of features. As a part of future work, ML [9] and DL [10-14] approaches could be explored for Phishing Webpage Detection.

REFERENCES

[1] Deepak Kshirsagar, Deepak Kumar, “An ensemble feature reduction method for web attack detection,” Journal of Discrete Mathematical Sciences and Cryptography, 23:1, 283-291, DOI: 10.1080/09720529.2020.1721861, 2020.

[2] Deepak Kshirsagar, Deepak Kumar, “An efficient feature reduction method for the detection of DOS attack,” ICT Express, Volume 7, Issue 3, 2021.

[3] Mahmoud Khonji, Youssef Iraqi, Andrew Jones, “Phishing Detection: A Literature Survey,” IEEE Communication Surveys and Tutorials, Vol. 15, No.4, Fourth Quarter, 2013

[4] S. Eftimie, R. Moinescu and C. Răcuciu, "Spear-Phishing Susceptibility Stemming From Personality Traits," IEEE Access, vol. 10, pp. 73548-73561, 2022, doi: 10.1109/ACCESS.2022.3190009.

[5] R. Valecha, P. Mandaokar and H. R. Rao, "Phishing Email Detection Using Persuasion Cues," IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 2, pp. 747-756, 1 March-April 2022, doi: 10.1109/TDSC.2021.3118931.

[6] Y. Xia, J. Liu and J. Wu, "Phishing Detection on Ethereum via Attributed Ego-Graph Embedding," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 69, no. 5, pp. 2538-2542, May 2022, doi: 10.1109/TCSII.2022.3159594.

[7] M. Chatterjee and A. S. namn, “Detecting phishing websites through deep reinforcement learning,” Proc. IEEE 43rd Annu. Comput. Softw. Appl. Conf. (COMPSAC), Jul. 2019

[8] S. Mahdavifar and A. A Ghorbani, “DeNNeS- Deep embedded neural network expert system for detecting cyber attacks,” Neural Comput. Appl. Vol. 32, no. 18, Sep 2020.

[9] Amit Savyanavar, Ghumare, T., Ghorpade, V. (2022), “Applying ML on COVID-19 Data to Understand Significant Patterns”, Lecture Notes on Data Engineering and Communications Technologies, vol 116. Springer, https://doi.org/10.1007/978-981-16-9605-3_35

[10] A. Jain, A. Malviya, D. Bajaj, R. Bhavsar and Amit Savyanavar, "Brain Tumor Detection using MLops and Hybrid Multi-Cloud," 2022 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), Pune, India, 2022, pp. 1-6, doi: 10.1109/ICBDS53701.2022.9936020.

[11] R. Bhandigani, N. Ujjwal, S. Pattekar and Amit Savyanavar, "Enterprise Optimization in Pharmaceutical Industry," 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Pune, India, 2021, pp. 1-7, doi: 10.1109/SMARTGENCON51891.2021.9645744.

[12] Amit S. Savyanavar and V. R. Ghorpade, “Efficient resource allocation scheme for on-the-fly computing based mobile grids”, International Journal of Information Technology, 14, 943–954, 2022, https://doi.org/10.1007/s41870-018-0269-y.

[13] Amit Sadanand Savyanavar, Vijay Ram Ghorpade, “Applicability of edge computing paradigm for Covid-19 mitigation”, In Intelligent Data-Centric Systems, Intelligent Edge Computing for Cyber Physical Applications, Academic Press, 2023, Pages 151-166, ISBN 9780323994125, https://doi.org/10.1016/B978-0-323-99412-5.00011-3.

[14] Amit Sadanand Savyanavar, Nikhil Mhala, Shiv H. Sutar, “Star-Galaxy classification using machine learning algorithms and deep learning”, International Journal on Information Technologies and Security, vol.15 , no.2, 2023, pp. 87-96. https://doi.org/10.59035/VVLR5284