ISSN 2347 - 3983
Volume 8, No. 1 January 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter13812020.pdf
https://doi.org/10.30534/ijeter/2020/13812020

ERCRFS: Ensemble of Random Committee and Random Forest using StackingC for Phishing Classification

Niranjan A[1],Venkata Krishnasai Sakhamuri [2], P Deepa Shenoy[3], Venugopal K R[4]

Department of CS & E, UVCE, Bangalore University, Bangalore, India. a.niranjansharma@gmail.com
Department of CS & E, KLEF, Vaddeswaram, India. venkatsakhamuri38@gmail.com
Department of CS & E, UVCE, Bangalore University, Bangalore, India. shenoypd1@gmail.com
Bangalore University, Bangalore, India. venugopalkr@gmail.com

ABSTRACT

The process of gathering sensitive information such as credit card/ debit card number, password, OTP etc., through a deceptive e-mail or SMS with links to Malicious site is called Phishing. An attacker during the Phishing attack masquerades as a trustworthy entity of some kind and lures the victim to provide his/her sensitive information. As per the latest data breach reports, the Phishing attacks make up almost a third of all data breaches. The attackers are using more sophisticated tools and are getting away with all the important data . Our work examines the application of Machine Learning Algorithms for the efficient classification of Phishing. The MMR, MSDMR and FSF Algorithms contribute towards the selection of significant Features from the data set such that there is no degradation in the performance of Classification. The hybrid framework ERCRFS, proposed by us ensures better Detection Accuracy values amongst all existing models.

Key words: Ensemble Learning, Hybrid Feature Selection, Phishing attack Analysis, Phishing Classification

1.INTRODUCTION

As per the 2019 Phishing Trends and Intelligence Report of the PhishLabs, phishing grew by a volume of 40.9% in the year 2018. The figures are scary for the people who get duped by the fraudsters through the SMSes,and Mails sent by them. The process of gathering sensitive information such as credit card/debit card number, password, OTP etc., through a deceptive e-mail or SMS with links to Malicious site is called Phishing[1]. An attacker during the Phishing attack masquerades as a trustworthy entity of some kind and lures the victim to provide his/her sensitive information. Based on
the mode used by the fraudster to launch the Phishing attack, there are 5 categories of Phishing namely Vishing, Smishing, Search Engine Phishing, Spear Phishing, and Whaling. When the fraudsters make voice calls to launch phishing, it is referred to as Vishing (Voice based Phishing) and when SMSes are used, it is called Smishing, Search Engine Search refers to the creation of Webpage focusing on a specific set of keywords and the pages wait for the victims who would search with those set of keywords and eventually land on the intended webpage. Spear phishing unlike other types of Phishing target a broad set of users of an organization for the attack. Hence it requires a lot of research to be done by the fraudster to get all the required profile information about these users. Whaling on the other hand targets only specific subset of users such as CEOs who would have more access to the information chain of an organization. In this context, this specific group of users could be considered as whales of the organization. Phishing attacks can succeed for a variety of reasons. Usually, they generate a feel of interest and urgency and prompt the users to click on the link and redirect them to a spoofed page that collects sensitive credentials about them. Effective Phishing Classification at an earlier stage only can facilitate the Phishing Detection process. That is, an accurate classification of a sample as phishing or benign would enable the Phishing Detection Mechanism to stamp out a phishing attack on its very onset. The application of Machine Learning techniques has been proven to be an effective tool for any classification problem. Use of Hybrid models involving Multi-layered classification schemes is not a new idea in research.

The current work, examines and investigates few such Hybrid models involving Multi-layered classification for effective prediction of Phishing samples. We explore two Ensembling Schemes namely StackingC and Voting for implementing the Hybrid model and based on the results, the most efficient model is proposed. Prior to classification, we explore various Rank Based Feature Selection Algorithms and determine the ranks of all the features present in the data set. Based on the rank or the weight values given by various Algorithms for a feature, Mean and Standard Deviation values are determined. The values are repeatedly computed for every feature present in the data set. Finally, a Mean of Mean of Ranks and a Mean of Standard Deviation of Ranks is computed for all features. All such features whose Mean and Standard Deviation values are lesser than the final Mean of Mean of Ranks and a Mean of Standard Deviation of Ranks are discarded and two

resulting Feature subsets involving only relevant and significant features are chosen for Classification. This process thus ensures that the time complexity of classification is reduced. We call these techniques as MMR and MSDR respectively. Furthermore, we perform a Union operation on the two feature subsets to choose the most significant features from both MMR and MSDR feature subsets. To further validate our approach, we also run various non-rank-based Feature Selection Algorithms with different Search techniques and determine those features that are picked up at least five times by various non-rank-based Feature Selection Algorithms. We call this technique as FSF (Feature Selection using Frequency). This step ensures that only most significant Feature subset is chosen for the further process and helps in reducing the dimensionality of the data set there by improving the speed. We further perform a union operation on the subsets generated by MMR and MSDR with FSF to select the final feature subset for the classification phase. Various built in classifiers are run on the feature subsets obtained after the union of MMR and MSDR, FSF and the feature subset obtained on the subsets after performing the union operation between these Algorithms. Top two performing classifiers in terms of different performance metrics are determined. Such classifiers are then subjected to ensemble models such as StackingC and Voting. The performance metrics are again recorded for the two ensemble models. The top performing model between the two, involving the top two performing Classifiers is finally chosen as our proposed model for the efficient classification of Phishing samples.

It was observed through our experiments that a hybrid model involving Random Committee and Random Forest when ensembled through StackingC offers better results in terms of the chosen metrics such as Prediction Accuracy. We call this approach as ERCRFS. For demonstrating the efficiency of the proposed ERCRFS approach, extensive experiments were conducted on three data sets from publicly available malware samples collections namely UCI Phishing Data set and two data sets by Mohammad et al.[12], consisting of 11055, 2456, and 2670 instances respectively. The feature subset is subjected to various Machine Learning Classifiers and different performance metrics such as Precision, Recall of benign and malware samples, and Weighted F-Measure are recorded in the final phase of the model. The top performing classifiers in terms of Prediction accuracy values on all the data sets is determined thereafter. The top classifiers are then combined using various ensemble approaches such as StackingC, Stacking, Grading and Voting. The experimental results obtained on the data sets indicate that ERCRFS outperforms the existing Models [2] in terms of Prediction Accuracy.

In Random committee, a number of Base classifiers are constructed by using unique random number seed values and the predictions of individual base classifiers are computed. The final prediction is made by averaging the classification prediction values of individual base classifiers.

A special type of Decision trees which is built on a random subset of chosen attributes is called a Random Tree. A decision tree is basically a set of nodes and their branches. A node illustrates a test on an attribute while the branch represents the outcome of the test. The external nodes (leaves) denote the final decision taken. The path covered from root to a leaf forms a classification rule. A classifier that is composed of several such independent Random trees is called a Random Forest. Each Random Tree of a Random Forest is built by the use of Bagging and Feature Randomness such that correlation between the trees is zero. As the Random Forest, makes the final prediction based on the maximum votes garnered for a specific prediction it would be more accurate than that of any individual Random Tree.

Some of the commonly used Ensemble techniques are Bagging, Boosting, Voting, Stacking, StackingC, and Grading. Voting involves the creation of a number of sub-models and involving each of them in the voting process of choosing on what should be the outcome of a prediction. Stacking involves the individual and independent training of heterogeneous learning algorithms on the data and considering the outcomes of each of them as additional inputs to the combiner algorithm for the final training. StackingC an improvised version of Stacking, makes use of Linear Regression as the Meta Classifier. Linear Regression is a process of merging a set of numeric values (x) into a predicted output value (y). Grading is one of the meta classification techniques that involves the process of identifying and correcting incorrect predictions if any. Unlike Stacking that uses the predictions of the base classifiers as metalevel attributes, Grading makes use of graded predictions (correct or incorrect) as meta-level classes.

The contributions of this article can be summarized as below:

- A novel general-purpose classifier framework involving Hybrid approach (ERCRFS) has been presented along with its evaluation on two data sets.

- We also propose three Feature Selection algorithms MMR, MSDR and FSF for the efficient selection of the Features without compromising on the Performance of Classification.

- The results of the extensive experiments that are conducted on individual classifiers and ensemble classifiers such as StackingC, Voting and Grading are presented to demonstrate the effectiveness of our proposed approach.

- We also present results of a performance comparison of ERCRFS with EKRV. The remainder of the article is organized as follows. Related work with respect to this field is discussed in Section II while Section III presents the proposed ERCRFS framework. Section IV elaborates the investigation methodology, while section V presents results, with analyses and discussions. Conclusion forms the final Section of this paper.

2. RELATED WORK

EKRV[2] involves the selection of Features through Consistency Subset Eval Feature Selection Algorithm using Greedy Stepwise search method. A Feature subset totaling 23 features with an additional class label is selected by this Algorithm. The Feature subset is then subjected to an ensemble of Random Committee with Random Tree as the base classifier and kNN with Cover Tree as the base classifier using voting for classifying the incoming sample as phishy or benign. The results indicate that the Prediction Accuracy is about 97.4%. The authors of [3], emphasize the importance of categorizing the features into three sets namely NLP based features, word vectors, and hybrid features before classifying the test samples. According to their findings, the NLP based features have better performance than word vectors with an average rate of 10.86%. Furthermore, they show that the combined use of NLP based features and word vectors would increase the performance of the phishing detection system. They make use of seven different machine learning algorithms namely Decision Tree, ADABOOST, K-star, kNN (n = 3), Random Forest, SMO and Naive Bayes for Classification. Their Detection Accuracy for different Classifiers ranges between 93.24%(ADABOOST) and 97.98% (Random Forest) for NLP based features.

The methodology presented in [4] illustrates the application of Clustering technique prior to Classification. Two Clustering algorithms to determine the subset based on the similarity that exists between the data instances that are used by them are the k-Means and the k-Medians. To determine the cluster size, they use Silhouette Criterion. Their results show that the Random Forest classifier when combined with k-Means(93.31% for k=5) and k-Medians (95.10% for k=3)offer better Detection rates.

The authors in their work [5], propose a new feature selection framework called Hybrid Ensemble Feature Selection. Their methodology involves generation of primary Feature subset using Cumulative Distribution Function Gradient Algorithm(CDF-g) and generation of secondary Feature subset using Data Perturbation Ensemble technique in Phase one and in phase two, Baseline Features are generated from secondary Features using Function Perturbation Ensemble method. Their approach offers better results when combined with Random Forest Classifier offering a Detection accuracy of 94.6% using only 20.8% of the Features from the overall Features.

The work proposed by [6], investigates the use of two types of Neural Networks namely, Ensemble Feedforward Neural Network (EFFNN) and Deep Learning Neural Network (DLNN) for Phishing Detection on the CSDMC2010 SPAM corpus data set. Their EFFNN employs a two-layer FFNN architecture with 18-10-1 neurons and Backpropagation being adopted as the learning algorithm. The Training phase makes use of sigmoid transfer function with a learning rate of 0.1 and 1000 iterations. The DLNN also has 18 input neurons with two hidden layers each layer having 10 hidden neurons. The output layer contains only one output neuron. The other settings are same as EFFNN experiment setting. The NN parameters are fine tuned by a series of experiments to achieve greater accuracy rates. They achieve a accuracy of 94.41% in EFFNN and 94.27% in DLNN.

The method suggested in [7] identifies the phishing sites based on the heuristic features extracted from URL, Website content and third-party services using machine learning algorithms. Some of the advantages of this model are detection of phishing sites that imitate legitimate sites by replacing the website content with an image, detection of zero-day phishing sites and offers high detection rate of 9.55% by the use of oblique Random Forest algorithm.

3. ERCRFS: A FRAMEWORK FOR EFFICIENT PHISHING CLASSIFICATION

Majority of the applications that involve Machine Learning techniques would revolve around Preprocessing and Classification phases. Selection of significant features from the data set becomes an important step of preprocessing, as all features in the data set are not relevant during the final prediction. We propose MMR, MSDR and FSF algorithms for the selection of Features by exploiting the numerous advantages of existing Weight Based Ranking Algorithms and Non-Weighted Feature Selection Algorithms. The classification phase involves, subjecting the resultant Feature subset to the proposed **ERCRFS** framework by means of ten-fold cross validation involving Random Committee, and Random Forest that are ensembled through StackingC and various performance metrics such as Prediction Accuracy, Precision, Recall of benign and malware samples, and Weighted F-Measure are recorded. The **ERCRFS** framework is depicted in Fig.1. The data sets that are used for the experimentation purpose are UCI

Phishing Data set and two data sets by Mohammad et al., consisting of 11055, 2456, and 2670 instances respectively.

The proposed MMR (Mean of Mean of Ranks) Feature Selection algorithm as listed in Algorithm 1 and MSDR (Mean of Standard Deviation of Ranks) employ several Rank Based Feature Selection Algorithms for determining the ranks of all the features that are present in the data set. Based on the rank values given by various Algorithms for a Feature, a Mean and a Standard Deviation values are determined for the Feature. This is repeatedly computed for every feature present in the data set. Finally, an overall Mean for the Mean and Standard Deviation values of all features are computed. All such features whose Mean values of both Mean and Standard Deviation that are lesser than the Overall Mean values are discarded and only Feature subsets that are relevant and significant for Classification are chosen. Correlation Attribute Eval, Gain Ratio Attribute Eval, Info Gain Attribute Eval , Relief Feature Attribute Eval, and Symmetrical Uncert Attribute Eval were employed to compute the Ranks R_i in the proposed MMR and MSDR. The MMR for UCI data set is 0.072957 and is 0.084369 and 0.084063 for Data set1 and Data set2 by Mohammed et.al., respectively. The MSDR values are 0.056997, 0.045986 and 0.0447for UCI, Data set1 and Data set2 respectively. Features having lesser Weights than the MMR and MSDR values were ignored. The proposed FSF returned a subset of 22 features in case of UCI Phishing data set and 13 features in case of Data set1 and 11 features from Data set2.

To determine the most common features of various Non-Weight Based Feature Selection algorithms and to further ensure that only significant Feature subset is selected for the classification phase, we propose another simple but useful algorithm called FSF (Feature Selection using Frequency). The FSF algorithm as listed in Algorithm 3 involves the determination of most significant features through the use of various Non-Weight Based Feature Selection algorithms with different Search Algorithms and to determine those features that appear in at least 5 different Non-Weight Based Feature Selection and various Search Algorithm combinations. The feature subset that is chosen in this manner by the FSF is compared with the feature subsets that are chosen by the MMR and MSDR to ensure no significant feature is left out from the final feature subset that is chosen for the classification phase, we hence perform an Union operation of the subsets. The various Non-Weight Based Feature Selection algorithms that were employed for experimentation are CFS Subset Eval, Consistency Subset Eval, and Filtered Subset Eval. We employed Genetic, Evolutionary, Linear Forward, Greedy stepwise, and BFS search techniques for all of the above Non-Weight Based Feature Selection algorithms. Features having lesser Weights than the MMR and MSDR values were ignored. The proposed FSF returned a subset of 22 features in case of UCI Phishing data set and 13 features in case of Data set1 and 11 features from Data set2. This amounts to a total reduction of features by 30% on UCI Phishing data set, 56.66% on Data set1 and 63.33% on Data set2 averaging at 63.33 % of Feature Reduction without any noticeable reduction in the performance.

Algorithm 1: MMR Feature Selection (Mean of Mean of Ranks)

Input: Data set *D* having *n* number of Features
Output: $F \subseteq D$ with Significant Features

1. for each feature $f_i \in D$ do
   Determine Rank R_i using chosen Weight Based Feature Selection Techniques
   next
2. for each feature $f_i \in D$ do
   Compute Sum and Mean of Ranks ($\sum R_i$) and (mR_i) $\sum R_i = R_1 + R_2 + \ldots + R_n$ and $mR_i = \sum R_i / n$
   next
3. for each feature $f_i \in D$ do
   Determine final Mean *M* for the mean of Ranks $M(mR_i)$
   next
4. Discard all $f_i \in D < M(mR_i)$
5. **return** *F*

Algorithm 2: MSDR Feature Selection (Mean of Standard Deviation of Ranks)

Input: Data set *D* having *n* number of Features
Output: $F \subseteq D$ with Significant Features

1. for each feature $f_i \in D$ do
   Determine Rank R_i using various Weight Based Feature Selection Techniques
   next
2. for each feature $f_i \in D$ do
   Determine Standard Deviation σ_r for the feature
   next
3. Determine Mean M of Standard deviation of Ranks $M(\sigma_r)$
4. Discard all $f_i \in D < M(\sigma_r)$
5. **return** *F*

Algorithm 3: FSF (Feature Selection using Frequency)

Input: *D* data set having *n* number of Features
Output: $F \subseteq D$ with Most Significant Features

1. Perform the intersection ($\cap$) of the features selected by the various Non-Weight Based Feature Selection Algorithms and their Search technique combinations to determine the Common Features.
2. Determine the Frequency of the Common Features.
3. Select only those Features whose Frequency is >=5

4. Perform the Union (∪) of the features obtained in Step 3 and Features obtained from MMR and MSDR.
5. **return** ***F***

Algorithm 4: Ensemble of Random Committee & Random Forest using StackingC **(ERCRFS)**
Input: $F \subseteq D$ obtained after applying **FSF**
Output: Performance Metrics

1. Provide input to the proposed Ensemble of Random Committee, and Random Forest using StackingC.
2. Apply ten-fold cross validation and the performance metrics are recorded.

The resulting feature subset from FSF is finally subjected to the proposed ERCRFS framework. The ERCRFS algorithm for efficient Phishing Classification is presented in Algorithm 4. Table 1 lists the selected feature subsets of the MMR on all the data sets. The proposed MMR, MSDR and FSF for Feature Selection are presented as Algorithm1, Algorithm2 and Algorithm3 respectively. Table 2 lists the selected feature subsets of the MSDR on all the data sets.

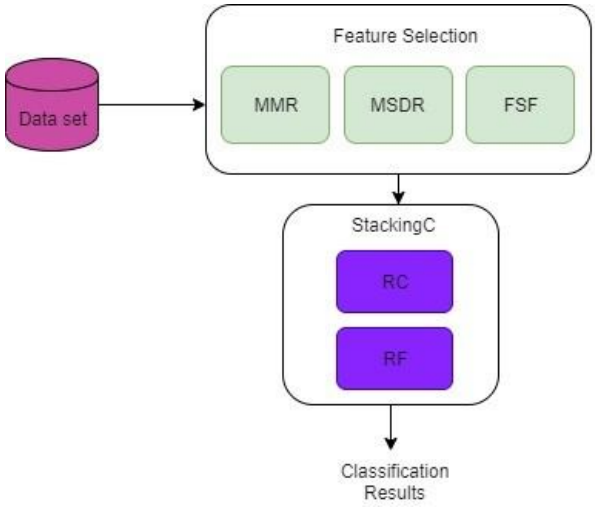

Figure 1: Proposed framework for Phishing Detection

The features listed in Table 3 are the selected features by the FSF algorithm from the three data sets. The Feature subset selected by the FSF is subjected to various Classifiers and the top two performing classifiers in terms of different performance metrics are determined. Fig. 2 depicts the performance comparison of various Classifiers. Such classifiers are then subjected to ensemble models such as StackingC, Stacking, Grading and Voting. The performance metrics are again recorded for the all ensemble models. Fig. 3 indicates the Comparative results of various Ensemble Models. The top performing model (StackingC) amongst them, involving the top two performing Classifiers (Random Committee and Random Forest) is finally chosen as our proposed model for the efficient classification of Phishing samples. Fig.4 depicts the Comparative Analysis of the proposed ERCRFS with other existing Models.

Table 1: Features Selected by MMR

UCIPhishingDataset	**Data set 1**	**Data set 2**
NumDots	Prefix_Suffix	Prefix_Suffix
PathLevel	having_Sub_Domain	having_Sub_ Domain
NumDash	SSLfinal_State	URL_of_Anchor
NumNumericChars	Domain_registeration_ length	Domain_registera tion_length
Domain_registeration_ length	URL_of_Anchor	SSLfinal_State
PctExtHyperlinks	age_of_domain	age_of_domain
PctExtResourceUrls	web_traffic	web_traffic
InsecureForms	Page_Rank	Page_Rank
PctNullSelfRedirect Hyperlinks		
FrequentDomainName Mismatch		
SubmitInfoToEmail		
IframeOrFrame		
PctExtNullSelfRedirect HyperlinksRT		

Table 2: Features Selected by MSDR

UCIPhishingDataset	**Data set-1**	**Data set 2**
NumDots	Prefix_Suffix	Prefix_Suffix
PathLevel	having_Sub_Domain	having_Sub_ Domain
NumDash	SSLfinal_State	URL_of_Anchor
NumNumericChars	Domain_registeration _length	Domain_regis teration_length
NumSensitiveWords	URL_of_Anchor	SSLfinal_State
PctExtHyperlinks	age_of_domain	age_of_domain
PctExtResourceUrls	web_traffic	web_traffic
InsecureForms	Page_Rank	Page_Rank
PctNullSelfRedirect Hyperlinks	DNSRecord	DNS
FrequentDomain NameMismatch	Google_Index	
SubmitInfoToEmail		
IframeOrFrame		
PctExtNullSelfRedirect HyperlinksRT		
NumDashInHostname		
IpAddress		
HostnameLength		
UrlLengthRT		
AbnormalExtForm ActionR		
ExtMetaScriptLink RT		

Table 3: Features Selected by FSF

UCIPhishingDataset	Data set-1	Data set-2
NumDots	Prefix_Suffix	Prefix_Suffix
PathLevel	having_Sub_Domain	having_Sub_ Domain
NumDash	SSLfinal_State	URL_of_Anchor
NumNumericChars	Domain_registeration_ length	Domain_registe ration_length
NumSensitiveWords	URL_of_Anchor	SSLfinal_State
PctExtHyperlinks	age_of_domain	age_of_domain
PctExtResourceUrls	web_traffic	web_traffic
InsecureForms	Page_Rank	Page_Rank
PctNullSelfRedirect Hyperlinks	DNSRecord	DNS
FrequentDomain NameMismatch	Google_Index	Request_URL
SubmitInfoToEmail	Request_URL	Links_in_tags
IframeOrFrame	Links_in_tags	SFH
PctExtNullSelfRedirect HyperlinksRT	SFH	
NumDashInHostname	Links_pointing_to_page	
IpAddress		
HostnameLength		
UrlLengthRT		
AbnormalExtForm ActionR		
ExtMetaScriptLinkRT		
RandomString		
AbnormalFormAction		
PopUpWindow		

4. INVESTIGATION METHODOLOGY

The data sets that are used for the experimentation purpose are UCI Phishing Data set and two data sets by Mohammad et al., consisting of 11055, 2456, and 2670 instances respectively. The proposed FSF returned a subset of 22 features in case of UCI Phishing data set and 13 features in case of Data set1 and 11 features from Data set2. This amounts to a total reduction of features by 30% on UCI Phishing data set, 56.66% on Data set1 and 63.33% on Data set2 averaging at 63.33 % of Feature Reduction without any noticeable reduction in the performance. A rigorous ten-fold cross validation was performed, and the performance metrics were recorded. A ten-fold cross validation typically requires dividing the data set into ten parts and the model would be trained with the nine parts of the data while the excluded part would be as the test set and the process would be repeated for ten rounds and each unused test set would be used during each round. Prediction Accuracy value is used as the yard stick for determining the efficiency of the classifiers. Different Ensemble approaches on the top performing classifiers are also tried to enhance the Performance of Classification.

A classifier with higher true positive rate and lower false positive rate is considered to be efficient. We define 8 Performance metrics of a classical classification methodology. N_{ben} is the number of normal or benign samples while N_{phi} is the number of phishing samples in the phishing data set. True Positive (TP) is the number of benign samples classified accurately as benign and is denoted as $N_{ben \rightarrow ben.}$ True Negative (TN) is the number of phishing samples classified accurately as phishing. It is denoted as $N_{phi \rightarrow phi.}$ False Positive (FP) is a measure of benign samples misclassified as phishing. It is denoted as $N_{ben \rightarrow phi}$ and False Negative (FN) is a measure of phishing instances misclassified as benign. It is represented as $N_{phi \rightarrow ben}$. The Detection Rate (DR) is the rate of phishing samples being classified accurately as phishing.

$$TP = \frac{N_{ben \rightarrow ben}}{(N_{ben \rightarrow ben} + N_{phi \rightarrow ben})} \times 100 \quad (1)$$

The rate of benign samples being classified inaccurately as phishing samples is referred to as False positive rate (FPR).

$$FPR = \frac{N_{ben \rightarrow phi}}{(N_{phi \rightarrow phi} + N_{ben \rightarrow ben})} \times 100 \quad (2)$$

The rate of phishing samples being classified inaccurately as normal samples is called False Negative Rate (FNR).

$$FNR = \frac{N_{phi \rightarrow ben}}{(N_{phi \rightarrow ben} + N_{phi \rightarrow phi})} \times 100 \quad (3)$$

The rate of benign samples being classified accurately as benign out of the total available benign samples is known as True Negative Rate (TNR).

$$TNR = \frac{N_{ben \rightarrow ben}}{(N_{ben \rightarrow ben} + N_{ben \rightarrow phi})} \times 100 \quad (4)$$

The total number of phishing and benign samples that are identified accurately with respect to the total number of all available instances is called Prediction Accuracy (PA).

$$PA = \frac{(N_{phi \rightarrow phi} + N_{ben \rightarrow ben})}{(N_{phi \rightarrow phi} + N_{ben \rightarrow ben} + N_{ben \rightarrow phi} + N_{phi \rightarrow ben})} \times 100 \quad (5)$$

Precision is the number of true positives divided by the total number of instances labeled as belonging to the positive class.

$$Precision = \frac{N_{phi \rightarrow phi}}{(N_{phi \rightarrow phi} + N_{ben \rightarrow phi})} \times 100 \quad (6)$$

Recall is the number of true positives divided by the total number of instances that really belong to the positive class.

$$Recall = \frac{N_{phi \rightarrow phi}}{(N_{phi \rightarrow phi} + N_{phi \rightarrow ben})} \times 100 \quad (7)$$

Weighted F-Measure (WFM) is the harmonic mean of Precision and Recall and is given by:

$$WFM = 2 \frac{Precision \cdot Recall}{Precision + Recall} \quad (8)$$

The experimental results as indicated in Fig. 2 suggest that Random Committee and Random Forest classifiers offer better Prediction Accuracy out of all classifier algorithms. Based on these findings, we decided to use an Ensemble approach [8] involving these top performing Classifiers on all the three Data sets. The Ensemble approaches that were experimented by us included Stacking, Grading, Voting, and StackingC. Our findings are plotted in Fig. 3. The points along the Y-axis in both Fig. 2, Fig. 3 and Fig. 4 indicate Prediction Accuracy [9] in terms of Percentage recorded with respect to various Classifiers, Ensemble Models and the existing Models respectively. It may be noticed from Fig. 4 that ERCRFS used by us perform better compared to EKRV [2] and HEFS [5]. Unfortunately, none of the Ensemble schemes behave uniformly on all the data sets. So, it was decided to use such Ensemble Scheme that behaves better on all the chosen data sets. Based on the results, we chose the Ensemble Scheme involving RC and Random Forest using StackingC technique as our Proposed Model.

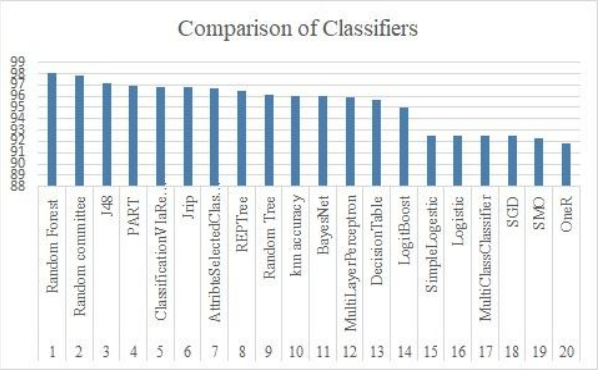

Figure 2: Comparison of Classifiers

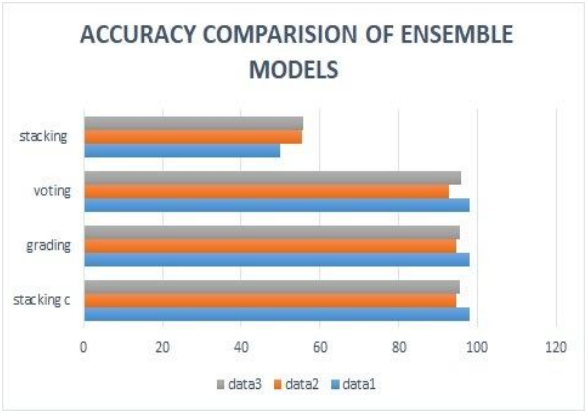

Figure 3: Comparison of Ensemble Schemes

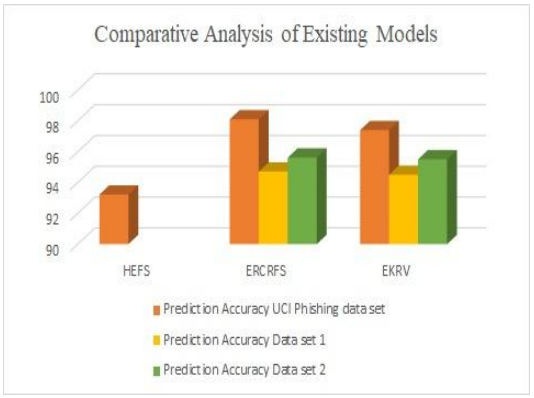

Figure 4: Comparison of Existing Models

5. CONCLUSION

Feature Selection using MMR, MSDR and FSF are applied to select significant features from the data set as a part of the Pre-Processing phase[10]. The proposed FSF returned a subset of 22 features in case of UCI Phishing data set[11] and 13 features in case of Data set1 and 11 features from Data set2. This amounts to a total reduction of features by 30% on UCI Phishing data set, 56.66% on Data set1 and 63.33% on Data set2 [12] averaging at 63.33 % of Feature Reduction without any noticeable reduction in the performance. The proposed ERCRFS outperforms EKRV and HEFS in terms of Prediction Accuracy. A tenfold cross validation if performed before recording the performance metrics. The ERCRFS model is required to be tested on real time data sets and the time required to carry out the entire process must be reduced so that online classification of the Phishing samples may be carried out on their onset.

.REFERENCES

1. Niranjan A, Nitish A, P Deepa Shenoy and Venugopal K R. **Security in Data Mining-a Comprehensive Survey**, Global Journal of Computer Science and Technology, vol. 16, no. 5, 2017, pp. 52-73.
2. Niranjan A, D K Haripriya, R Pooja, S Sarah, P Deepa Shenoy, Venugopal K R. **EKRV: Ensemble of kNN and Random Committee Using Voting for Efficient Classification of Phishing** Progress in Advanced Computing and Intelligent Engineering, Springer, Singapore, e-ISBN : 978-981-13-1708-8, p-ISBN : 978-981-13-1707-1, pp. 403-414, 2019. DOI: https://doi.org/10.1007/978-981-13-1708-8_37
3. Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, and Banu Diri. **Machine learning based phishing detection from URLs, Expert Systems with Applications**(Elsevier), vol.117 (2019), pp.345–357, 2019. https://doi.org/10.1016/j.eswa.2018.09.029

4. S Priyaa, and S Selvakumar. **Classifier Performance Evaluation of Phishing Detection Model on Optimal Number of Clusters** ,Proceedings of International Conference on Sustainable Computing in Science, Technology & Management (SUSCOM-2019), pp. 406--415, 2019. https://doi.org/10.2139/ssrn.3352333
5. Kang Leng Chiew, Choon Lin Tan, KokSheik Wong, Kelvin S.C. and Yong, Wei King Tiong,.**A New Hybrid Ensemble Feature Selection Framework for Machine Learning-based Phishing Detection System** . Information Sciences, PII: S0020-0255(19)30076-3, 2019. DOI: https://doi.org/10.1016/j.ins.2019.01.064
6. Gan Kim Soon, Liew Chean Chiang, Chin Kim On, Nordaliela Mohd Rusli and Tan Soo. **Fun,Comparison of Ensemble Simple Feedforward Neural Network and Deep Learning Neural Network on Phishing Detection ,**Computational Science and Technology. Lecture Notes in Electrical Engineering, vol 603, pp.595--604, Springer, Singapore, 2020. DOI : https://doi.org/10.1007/978-981-15-0058-9_57
7. Rao, R.S. & Pais, A.R., **Detection of phishing websites using an efficient feature-based machine learning framework** , Neural Computing & Applications, 2019. DOI: https://doi.org/10.1007/s00521-017-3305-0
8. S. Wu, P. Wang, X. Li, Y. Zhang, **Effective Detection of Android Malware Based on the Usage of Data Flow APIs and Machine Learning**, Information and Software Technology, vol.75, pp. 17-25, 2016, ISSN 0950-5849.
9. Prakash Kolla Bhanu, Dorai Rangaswamy M A, **Content Extraction of Biological Datasets Using Soft Computing Techniques** Journal of Medical Imaging and Health Informatics, Volume 6, Number 4, August 2016, pp. 932-936(5). https://doi.org/10.1166/jmihi.2016.1931
10. Asha S Manek, Samhitha M R, Shruthy S , Veena H Bhat, P Deepa Shenoy, M. Chandra Mohan, Venugopal K R, L M Patnaik **RePID-OK: Spam Detection using Repetitive Pre-processing**" IEEE CUBE 2013 Conference, ISBN : 978-1-4799-2234-5, pp. 144-149, November 15-16, 2013.
11. Niranjan A, Akshobhya K M, P Deepa Shenoy, Venugopal K R,**EKMC: Ensemble of kNN using MetaCost for Efficient Anomaly Detection** , Volume 4, Issue 5, Pages 401-408, OCT 2019, ASTESJ, https://dx.doi.org/10.25046/aj040552
12. Rami M Mohammad, **Phishing Websites Features**, School of Computing and Engineering, University of Huddersfield, Huddersfield, UK, 2015. http://eprints.hud.ac.uk/24330/