10th International Conference on Computer and Knowledge Engineering (ICCKE2020) October 29-30, 2020 – Ferdowsi University of Mashhad - Iran

Ransomware Detection with Semi-Supervised Learning

Fakhroddin Noorbehbahani
Faculty of Computer Engineering
University of Isfahan
Isfahan, Iran
noorbebahani@eng.ui.ac.ir
noorbehbahani@gmail.com

Mohammad Saberi
Intellicom Inc.
Isfahan, Iran
saberi@intellicom.ir
mohammad.saberi72@gmail.com

***Abstract*—Today, ransomware is one of the most harmful cybersecurity threats that organizations and people face. Hence, there is a vital need for developing effective ransomware detection methods. Machine learning methods can be very useful for ransomware detection if there is sufficient labeled data for training. However, labeling data is time-consuming and expensive while a huge amount of unlabeled data exists. To cope with this problem, semi-supervised learning can be employed that exploits a few labeled data and a lot of unlabeled data for learning. To our best knowledge, there is no research investigating semi-supervised learning methods for ransomware detection. In this paper, we analyze different feature selection and semi-supervised classification methods applied to the CICAndMal 2017 dataset. Our findings suggest that the wrapper semi-supervised classification method using the random forest as a base classifier and OneR or Chi-squared as a feature selection method outperforms the other semi-supervised classification methods for ransomware detection.**

Keywords—ransomware; machine learning; semi-supervised learning; feature selection; malware detection

I. INTRODUCTION

In recent years, ransomware is known as one of the most dangerous malwares that encrypts or locks the victim's data until the ransom money is paid. Ransomware is categorized as the locker and the crypto that blocks access to devices and files respectively.

According to Cybersecurity Ventures [1], ransomware costs have been raised from $325 million in 2015 to $11.5 billion in 2019 (about 25X increase in four years). It is predicted that ransomware cost will reach $20 billion by 2021 and every 11 seconds ransomware will attack a business by the end of 2021.

The most prominent ransomwares with respect to the percentage of reported incidents are CryptoLocker (66%), WannaCry (49%), CryptoWall (34%), Locky (24%), Petya (17%), CryptXXX (14%), and notPetya (12%) according to North American MSPs reporting 2019. The most common ransomware delivery methods are spam/phishing emails, lack of cybersecurity training, weak password and access management, poor user practices/gullibility, and malicious websites/web ads (MSPs report).

Today, machine learning-based ransomware detection has received a great amount of attention from researchers. Most of the works focused on this approach employ supervised machine learning techniques to build a detection model [2]. However, supervised learning methods require a huge amount of labeled data that are difficult to obtain. Insufficient training data yields a weak detection model. In this regard, semi-supervised learning methods can be applied to learn a model with a limited number of labeled training data and a huge amount of unlabeled data. In this paper, we analyze different semi-supervised classification methods to find which of them is more effective for ransomware detection.

The rest of the paper is structured as follows. Section II reviews the researches on semi-supervised malware detection. Section III describes the research methodology and Section IV presents evaluation results. Finally, in Section V the paper is concluded and future works are discussed.

II. RELATED WORK

Ransomware detection techniques can be categorized as behavior-based, I/O request packet monitoring, network traffic monitoring, API call monitoring, at the storage level, in android devices, and other techniques. Most ransomware detection methods rely on the machine learning approach [2]. There are several pieces of research on the application of supervised machine learning methods for ransomware detection such as [3] [4] [5] [6] [7] [8]. Supervised learning techniques require sufficient labeled training data that is expensive and time-consuming to obtain them. To cope with this limitation, semi-supervised learning methods are applied to learn a model utilizing limited labeled data and a large dataset of unlabeled data. To our best knowledge, there is no research on semi-supervised ransomware detection. Hence in this section, we describe the works on semi-supervised malware detection methods.

In [9], a new method of unknown malware detection adopting a semi-supervised learning approach has been proposed. The authors have employed the Learning with Local

978-1-7281-8566-8/20/$31.00 ©2020 IEEE

and Global Consistency (LLGC) technique which is a semi-supervised classification algorithm. Their proposed method is based on learning a classifier using a set of labeled (malware and legitimate software) and unlabeled data. Exploiting 50% labeled data their method could achieve an accuracy of 0.86.

Zhang and his colleagues [10], has presented a new semi-supervised algorithm called ColSVM for malware detection. Their proposed method is based on collaborative learning that divides the training set into 2 ones by splitting feature sets. Next, the SVM classifier is trained on each training sets and the resulting model is applied to classify the unlabeled test set. Afterward, some accurate classified test data are added to the training set to form a new training set and the algorithm starts again with a new training set. The authors have been shown that the application of ColSVM using a few labeled data leads to high accurate classification.

In [11], a new malware detection method has been proposed based on the weighted word embedding vector (WEV) and clustering. The authors applied API syscalls modeled as a text. Next, TF-IDF and Word2Vec techniques are employed to represent syscalls WEVs. Finally, WEVs are clustered with the *k*-means clustering algorithm to form 115 clusters and reduced to 4 final categories namely Modification, Stealing, Evasion, and Disruption.

In [8], the LLGC semi-supervised learning method has been applied to classify the apps as benign or malware. In this work, four feature selection methods have been examined as preprocessing. The results show that the Rough Set Analysis (RSA) feature selection together with the LLGC learning method is effective for malware detection.

Santos and his colleagues [12], developed a malware detection method adopting the LLGC semi-supervised learning approach. Their proposed method achieves an accuracy of 0.85 using 90% labeled data.

A malware detection framework applying model-based semi-supervised (MBSS) classification on dynamic Android API has been proposed in [13]. The authors have been shown that their framework is competitive compared to the SVM, *k*NN (*k*-nearest neighbor algorithm), and LDA malware detection classifiers.

III. Research Methodology

As shown in Fig. 1, the research methodology consists of four steps. In the first step called the dataset preparation, the ransomware datasets from CICAndMal2017 dataset were selected including 10 ransomware families. Next, the same number of benign instances were added to each ransomware dataset as the ransomware dataset size. It is notable that the benign instances were chosen randomly from the benign dataset. Therefore, 10 ransomware-family/benign balanced datasets (DS_1 to DS_{10}) were formed. To create the DS_M dataset, we merged all DS_i ransomware family datasets and relabeled each family label as "Ransomware". Hence DSM is a 2-label ransomware/benign dataset. Moreover, in this step, some irrelevant features were eliminated from the datasets (flow ID, Source IP, Destination IP, Time Stamp). The feature "fwd header length" was also eliminated because it was duplicated in the datasets. Finally, all continues features were normalized to [0,1].

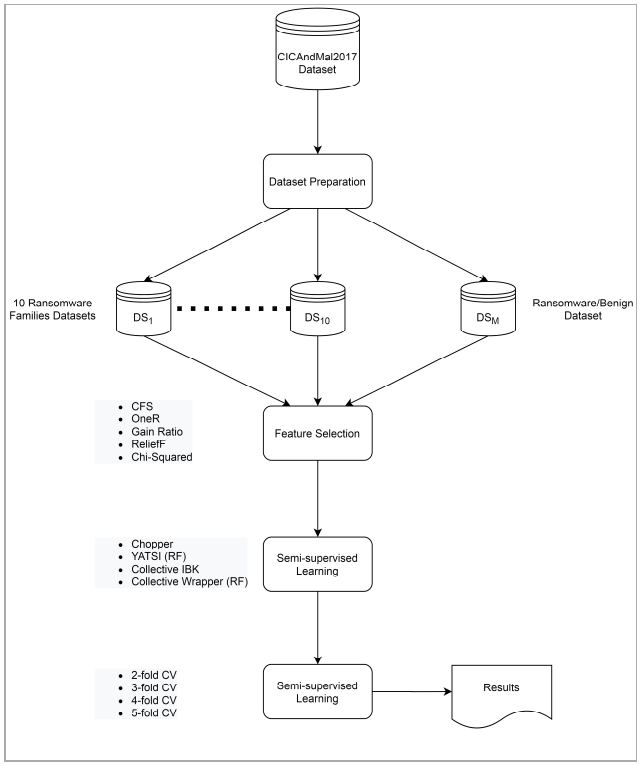

Fig. 1. Research methodology

In the second step, we examined five feature selection algorithms on 10 ransomware families separately as well as ransomware/benign datasets. The selected feature selection algorithms are CFS (Correlation-based Feature Subset Selection) [14], OneR [15], Gain Ratio [16], ReliefF [17], and Chi-squared [18].

In the fourth step, four semi-supervised learning methods were investigated namely Chopper [19], YATSI [20], Collective IBK [21], and Collective Wrapper [19].

Chopper [19] applies a classifier to label the test data after training on the train set. The test set is then ranked based on the difference between two class confidences. Next, the fold with the highest ranking is added to the training set and the classifier is trained again on a new training set.

YATSI (Yet Another Two-Stage Idea) is a collective classifier that exploits the base classifier to train on the training set and labeling the unlabeled data (called pre-labeled data). Next, with the application of the *k*NN using the actual training set and pre-labeled data, the test instances are classified [20].

Collective IBK [21] employs the IBK algorithm (a *k*NN based algorithm), to find the best *k* in the training set. Then it

978-1-7281-8566-8/20/$31.00 ©2020 IEEE

finds *k*-nearest instances for each test instance from the training and test sets that sorted based on their distances to the test instance. The neighborhoods are ranked based on their differences in class occurrences. Then each unlabeled test instance ranked highest is classified using the majority vote.

The wrapper-based semi-supervised methods exploit a supervised base classifier. First, the classifier is trained on the labeled data, then the predictions of the classifier are applied to generate additional labeled data to re-train the classifier [22]. Finally, the results are evaluated and analyzed using *k*-fold cross-validation and the accuracy measure.

IV. Evaluation Results

To conduct the experiments, Weka 3.8.4 and Collective package was applied. Because the ReliefF, OneR, Gain Ratio and, Chi-squared feature selection methods rank the features, first, the CFS feature selection method is applied, then for other feature selection methods, the same number of features (as CFS) are chosen from the ranked feature list. The evaluation details are described in the following sections.

A. Dataset Description

CICAndMal2017 android malware dataset consists of four malware categories namely Adware, Ransomware, Scareware, and SMS Malware and 80 traffic features. The dataset includes 5065 benign apps and more than 43 families of malware. The dataset is fully labeled and contains network traffic, logs, API/SYS calls, phone statistics, and memory dumps of malware families. The Ransomware category includes 10 families and 101 captured samples. Table I displays the details of ransomware families [23].

TABLE I. Dataset Description

Label	Family	Year	# of Records	# of Features	Captured Samples
RANSOMWARE	Charger	2017	79090	84	10
	Jisut	2017	51344	84	10
	Koler	2015	89410	84	10
	LockerPin	2015	50618	84	10
	Pletor	2014	9430	84	10
	PornDroid	2016	92167	84	10
	RansomBO	2017	79712	84	10
	Simplocker	2015	72682	84	10
	Svpeng	2014	108552	84	11
	WannaLocker	2017	65402	84	10
BENIGN	Benign	2017	6500	84	600

B. Evaluation Scheme

To evaluate the results, the accuracy measure and *k*-fold cross-validation are employed. The accuracy measure is a well-known evaluation metric in machine learning. However, it is not a proper measure when the dataset is imbalanced. Because all datasets in our experiment are balanced, the accuracy is a good measure for evaluation. The accuracy is calculated by (1).

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

In (1), *TP*, *FP*, *TN*, and *FN* are the number of true positive, false positive, true negative and, false negative predictions, respectively.

The *k*-fold cross-validation is effective and useful for the performance evaluation of the classifiers. This is because it validates the data multiple times consequently reduces the evaluation bias.

When *k*-fold cross-validation is applied the dataset is split into *k* parts. In each run, the classifier is trained on *k-1* parts and tested on the remained part. This process runs *k* times and the average accuracy is calculated. But how can cross-validation be applied for semi-supervised learning? In a semi-supervised setting, in each run, the test set is considered as unlabeled data. In other words, in each fold, the classifier is trained on *k-1* parts as labeled data and the remained part as unlabeled data. Next, the accuracy is calculated using the test part. It is worth noting that because we want to use a limited number of labeled data, the folds should be swapped. That is training and testing folds are inverted. For example, in the 10-fold CV the classifier is trained on 10% data and tested against 90% instead of the normal 90/10. Fig. 2 displays *k*-fold cross-validation with swap for semi-supervised classification evaluation.

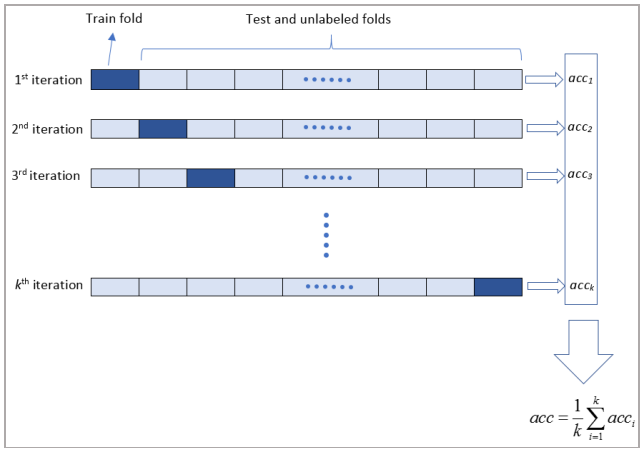

Fig. 2. *k*-fold cross-validation in semi-supervised setting

C. Evaluation Results

Fig. 3.a to Fig. 3.j display the classifier accuracies obtained for each ransomware family. To analyze the classifiers different feature selection schemes were examined and in the following diagrams, the best feature selection scheme is shown for each classifier. In [6], it is shown that the best supervised classification method for ransomware detection is Random Forest (RF). Hence, in the case of Wrapper and YATSI semi-supervised learning techniques, the random forest was employed as a base classifier.

978-1-7281-8566-8/20/$31.00 ©2020 IEEE

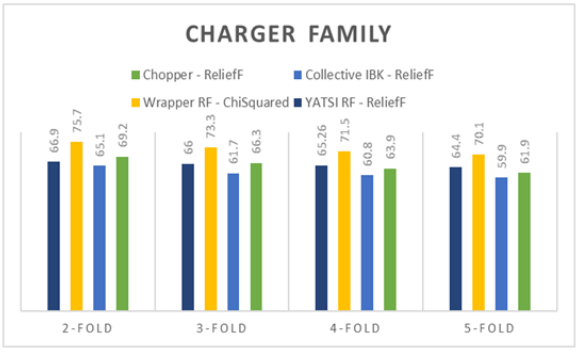

Fig. 3.a. The accuracy of the classifiers for Charger family

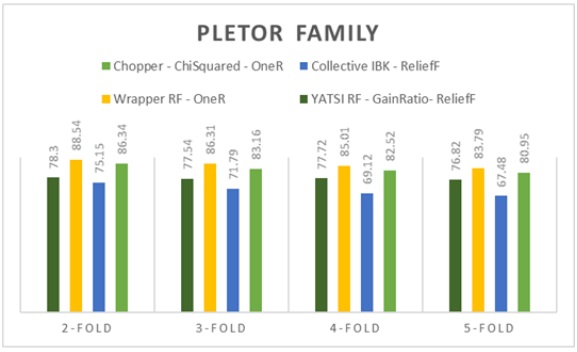

Fig. 3.e. The accuracy of the classifiers for Pletor family

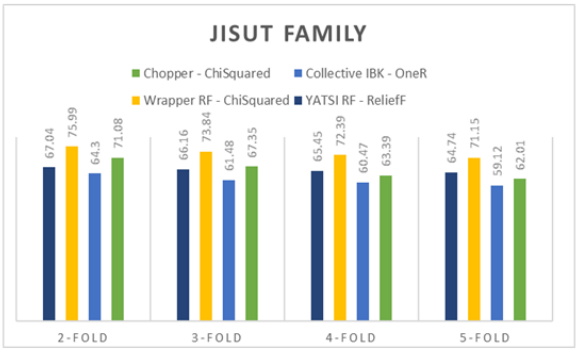

Fig. 3.b. The accuracy of the classifiers for Jisut family

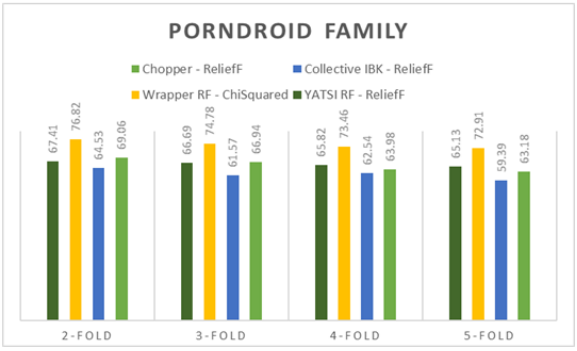

Fig. 3.f. The accuracy of the classifiers for PornDroid family

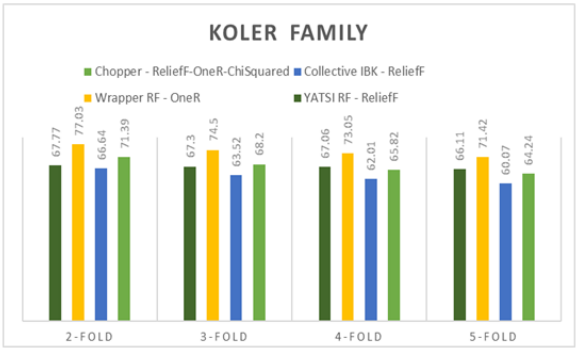

Fig. 3.c. The accuracy of the classifiers for Koler family

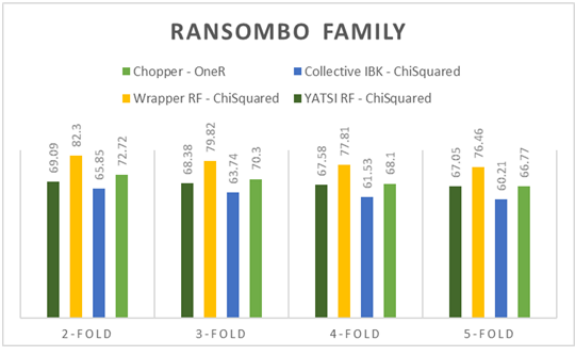

Fig. 3.g. The accuracy of the classifiers for RansomBO family

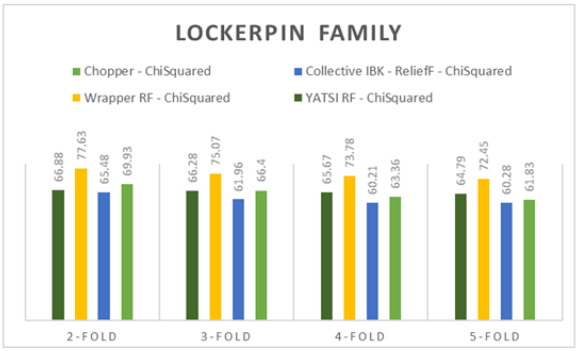

Fig. 3.d. The accuracy of the classifiers for LockerPin family

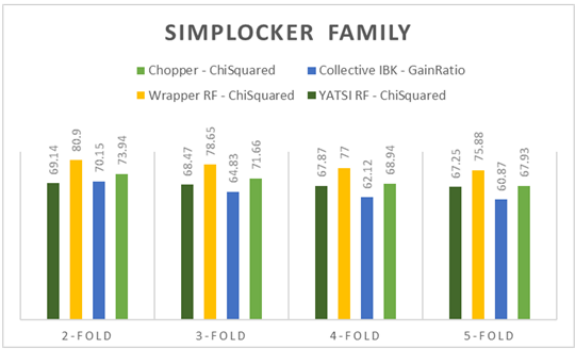

Fig. 3.h. The accuracy of the classifiers for Simplocker family

978-1-7281-8566-8/20/$31.00 ©2020 IEEE

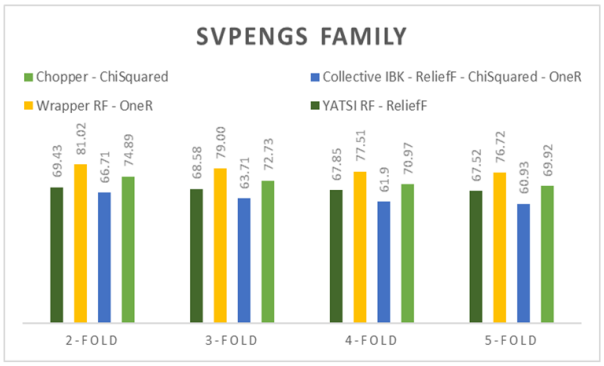

Fig. 3.i. The accuracy of the classifiers for Svpeng family

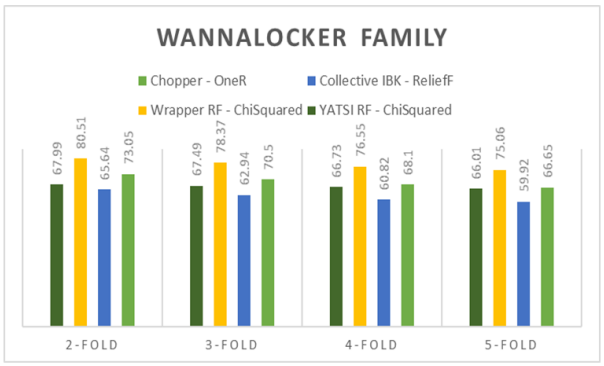

Fig. 3.j. The accuracy of the classifiers for WannaLocker family

As shown in figures the wrapper semi-supervised learning with the random forest as a base classifier outperforms the other methods and the worst classifier is collective IBK. Among feature selection methods the Chi-squared and OneR are the best. In Fig. 4, the best accuracy obtained for each ransomware family exploiting wrapper semi-supervised learning with random forest and the best feature selection method are depicted. The best results pertain to 2-fold cross-validation.

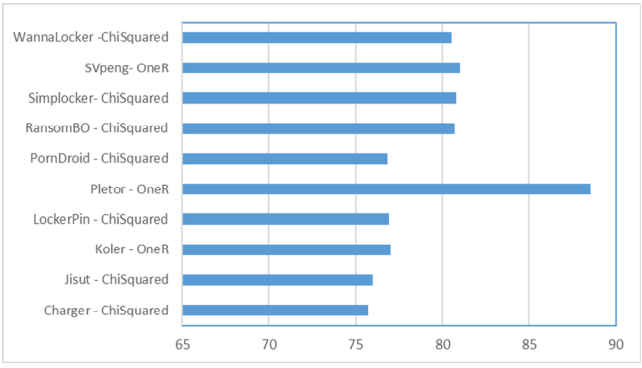

Fig. 4. The best accuracy for each family obtained by Wrapper-RF semi-supervised learning together with the best feature selection method.

Table II summarizes the evaluation results of binary classifiers and the best feature selection method using the merged 2-class dataset (DS_M). It should be highlighted that due to the weak results of collective IBK for ransomware families classification, we didn't apply this method for binary classification. As shown in Table II, again Wrapper RF outperforms the other methods. However, the results imply that ransomware detection using family datasets separately is more effective than binary classification.

TABLE II. BINARY CLASSIFIER ACCURACY

Classifier	Cross-validation			
	2-fold	*3-fold*	*4-fold*	*5-fold*
Chopper	62.74 (OneR)	62.05 (OneR)	61.78 (GainRatio)	61.28 (GainRatio)
Wrapper RF	69.50 (OneR)	67.39 (OneR)	65.79 (OneR)	64.73 (OneR)
YATSI RF	64.49 (GainRatio)	64.10 (GainRatio)	64.02 (GainRatio)	63.75 (GainRatio)

V. CONCLUSION AND FUTURE WORK

In this paper, different semi-supervised ransomware detection and feature selection methods were examined. The results show that the Wrapper RF classification and Chi-squared or OneR feature selection methods are very effective in semi-supervised ransomware detection. The main limitation of this research is that the applied feature selection methods are supervised. We applied the Simplified Silhouette Filter (SSF) [24] unsupervised feature selection, but the results were very poor. Therefore, it is necessary to investigate and propose a semi-supervised feature selection method for ransomware detection in future works.

REFERENCES

[1] "Global Ransomware Damage Costs Predicted To Reach $20 Billion (USD) By 2021." [Online]. Available: https://cybersecurityventures.com/global-ransomware-damage-costs-predicted-to-reach-20-billion-usd-by-2021/. [Accessed: 13-Jun-2020].

[2] C. V Bijitha, R. Sukumaran, and H. V Nath, "A Survey on Ransomware Detection Techniques," in *International Conference On Secure Knowledge Management In Artificial Intelligence Era*, 2019, pp. 55–68.

[3] B. A. S. Al-rimy, M. A. Maarof, Y. A. Prasetyo, S. Z. M. Shaid, and A. F. M. Ariffin, "Zero-Day Aware Decision Fusion-Based Model for Crypto-Ransomware Early Detection," *Int. J. Integr. Eng.*, vol. 10, no. 6, 2018.

[4] D. Morato, E. Berrueta, E. Magaña, and M. Izal, "Ransomware early detection by the analysis of file sharing traffic," *J. Netw. Comput. Appl.*, vol. 124, pp. 14–32, 2018.

[5] O. M. K. Alhawi, J. Baldwin, and A. Dehghantanha, "Leveraging machine learning techniques for windows ransomware network traffic detection," in *Cyber Threat Intelligence*, Springer, 2018, pp. 93–106.

[6] F. Noorbehbahani, F. Rasouli, and M. Saberi, "Analysis of Machine Learning Techniques for Ransomware Detection," in *2019 16th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC)*, 2019, pp. 128–133.

[7] N. B. Harikrishnan and K. P. Soman, "Detecting Ransomware using GURLS," in *2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, 2018, pp. 1–6.

[8] A. Mahindru and A. L. Sangal, "Feature-Based Semi-supervised Learning to Detect Malware from Android," in *Automated Software Engineering: A Deep Learning-Based Approach*, Springer, 2020, pp. 93–118.

978-1-7281-8566-8/20/$31.00 ©2020 IEEE

[9] I. Santos, J. Nieves, and P. G. Bringas, "Semi-supervised learning for unknown malware detection," *Adv. Intell. Soft Comput.*, vol. 91, no. January 2011, pp. 415–422, 2011.

[10] K. Zhang, C. Li, Y. Wang, X. Zhu, and H. Wang, "Collaborative Support Vector Machine for Malware Detection.," in *ICCS*, 2017, pp. 1682–1691.

[11] H. L. Duarte-Garcia *et al.*, "A Semi-supervised Learning Methodology for Malware Categorization using Weighted Word Embeddings," in *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2019, pp. 238–246.

[12] I. Santos, B. Sanz, C. Laorden, F. Brezo, and P. G. Bringas, "Opcode-sequence-based semi-supervised unknown malware detection," in *Computational Intelligence in Security for Information Systems*, Springer, 2011, pp. 50–57.

[13] L. Chen, M. Zhang, C. Yang, and R. Sahita, "POSTER: semi-supervised classification for dynamic android malware detection," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 2479–2481.

[14] M. A. Hall, "Correlation-based feature subset selection for machine learning," *Thesis Submitt. Partial fulfillment Requir. degree Dr. Philos. Univ. Waikato*, 1998.

[15] J. Novaković, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugosl. J. Oper. Res.*, vol. 21, no. 1, 2016.

[16] N. Arora and P. D. Kaur, "A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment," *Appl. Soft Comput.*, vol. 86, p. 105936, 2020.

[17] I.-H. Lee, G. H. Lushington, and M. Visvanathan, "A filter-based feature selection approach for identifying potential biomarkers for lung cancer," *J. Clin. Bioinforma.*, vol. 1, no. 1, p. 11, 2011.

[18] M. Trabelsi, N. Meddouri, and M. Maddouri, "A new feature selection method for nominal classifier based on formal concept analysis," *Procedia Comput. Sci.*, vol. 112, pp. 186–194, 2017.

[19] B. Pfahringer, K. Driessens, and P. Reutemann, "Collective and Semi-supervised classification," *Univ. Waikato*, pp. 1–21, 2014.

[20] K. Driessens, P. Reutemann, B. Pfahringer, and C. Leschi, "Using weighted nearest neighbor to benefit from unlabeled data," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2006, pp. 60–69.

[21] C. Laorden, B. Sanz, I. Santos, P. Galán-García, and P. G. Bringas, "Collective classification for spam filtering," in *Computational Intelligence in Security for Information Systems*, Springer, 2011, pp. 1–8.

[22] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[23] A. H. Lashkari, A. F. A. Kadir, L. Taheri, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark android malware datasets and classification," in *2018 International Carnahan Conference on Security Technology (ICCST)*, 2018, pp. 1–7.

[24] T. F. Covões and E. R. Hruschka, "Towards improving cluster-based feature selection with a simplified silhouette filter," *Inf. Sci. (Ny).*, vol. 181, no. 18, pp. 3766–3782, 2011.

978-1-7281-8566-8/20/$31.00 ©2020 IEEE