Enhancing Ransomware Classification with Multi-Stage Feature Selection and Data Imbalance Correction

Faithful Chiagoziem Onwuegbuche[1,2][0000−0001−9580−4260], Anca Delia Jurcut[2][0000−0002−2705−1823], and Liliana Pasquale[2][0000−0001−9673−3054]

[1] SFI Center for Research Training in Machine Learning (ML-Labs)
faithful.chiagoziemonwuegb@ucdconnect.ie
[2] School of Computing, University College Dublin, Ireland
{anca.jurcut, liliana.pasquale}@ucd.ie

Abstract. Ransomware is a critical security concern, and developing applications for ransomware detection is paramount. Machine learning models are helpful in detecting and classifying ransomware. However, the high dimensionality of ransomware datasets divided into various feature groups such as API calls, Directory, and Registry logs has made it difficult for researchers to create effective machine learning models. Class imbalance also leads to poor results when classifying ransomware families. To tackle these challenges, in this paper we propose a three-stage feature selection method that effectively reduces the dimensionality of the data and considers the varying importance of the different feature groups in the classification of ransomware families. We also applied cost-sensitive learning and re-sampling of the training data using SMOTE to address data imbalance. We applied these techniques to the Elderan ransomware dataset. Our results show that the proposed feature selection method significantly improves the detection of ransomware compared to other state-of-art studies using the same dataset. Furthermore, the data balancing techniques (cost-sensitive learning and SMOTE) were effective in the multi-class classification of ransomware.

Keywords: Ransomware detection · Malware classification · Machine learning · Feature analysis.

1 Introduction

Ransomware has rapidly become a serious threat to today's society and has affected several critical sectors, including healthcare, critical infrastructure, education and finance. For example, in May 2017 the UK National Health System (NHS) was attacked by the WannaCry ransomware, resulting in the loss of patients' records, delays in non-urgent surgeries and cancellation of 19,000 patient appointments [13, 14]. The rise of ransomware can be attributed to the financial gains accrued using cryptocurrencies as a payment mechanism [17], the COVID-19 working from home paradigm resulting in some workers adopting poor security

practices [7], and the popularity of ransomware-as-a-service, which allows novice attackers to launch ransomware with pre-built software and platforms [20]. Ransomware is a type of malware developed to facilitate different malicious activities such as blocking access to a computer system, encrypting files, exfiltrating files or even damaging files unless a ransom is paid [7,21]. Traditionally, ransomware detection methods rely on signatures, which can be easily evaded by generating new variants and using obfuscation techniques [24,25]. To overcome these limitations, machine learning is now used for ransomware detection and classification. However, there are issues in the current literature regarding this approach that this study aims to address.

First, the high dimensionality of ransomware datasets, obtained through dynamic analysis, poses a major challenge in developing effective machine learning models for ransomware classification [1,22,25]. To avoid the curse of dimensionality and reduce required computational resources, researchers have proposed different feature selection methods, generally classified into four categories: Filter, Wrapper, Embedded, and Hybrid [5]. However, finding the best feature selection method or combination of methods for a specific task is still an open problem [8,10]. Moreover, most feature selection methods used in the literature on ransomware classification ignore the varying importance of different feature groups, which can lead to suboptimal results and limit the effectiveness of the models developed using these methods [1]. In this study, we propose a three-stage feature selection method that significantly improves the classification of ransomware and considers the different feature groups that are present in the data.

Second, multi-class ransomware classification poses a class imbalance problem, that leads to poor performance when minority class examples are classified [9, 11, 19, 27]. Most studies (e.g., [16, 22, 25]) on ransomware detection and classification using machine learning have not considered the class imbalance problem. Those that considered the multi-class classification (e.g., [1]) did not consider the effect of different data imbalance correction techniques in the multi-class and binary classification of ransomware. To address the data imbalance problem in the classification of ransomware, we adopt two approaches: resampling the training dataset using Synthetic Minority Oversampling Technique (SMOTE) and cost-sensitive machine learning methods. The SMOTE algorithm generates synthetic data for the minority class(es) based on their feature space similarities using nearest neighbours [19], while cost-sensitive learning methods modify machine learning models to bias toward classes with fewer examples in the training dataset.

In this paper, we provide the following contributions:

- We propose a three-stage feature selection method that significantly improves the classification of ransomware, reduces the dimensionality of the dataset and considers the different feature groups involved in the data.
- We adopt two approaches to address class imbalance in the classification of ransomware: resampling the training dataset using Synthetic Minority Oversampling Technique (SMOTE) and cost-sensitive machine learning methods.

We compared the performance of various machine learning models, i.e. eXtreme Gradient Boosting (XGBoost), Logistic Regression (LR), Random Forest (RF), Decision Trees (DT) and Support Vector Machine (SVM), in detecting and classifying ransomware using the Elderan ransomware dataset [25]. We used balanced accuracy as the primary evaluation metric. Our evaluation results show that the proposed feature selection method improves ransomware detection significantly compared to previous studies, and that cost-sensitive learning and SMOTE improve the ability to classify different ransomware families.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents the research approach used in this paper. Section 4 discusses the experimental results obtained for binary and multi-class classification. Finally, Section 5 concludes the paper.

2 Related Work

Previous work has used machine learning to detect and classify ransomware using the Elderan dataset. Abbasi et al. [1] proposed a two-stage feature selection method for machine learning-based ransomware detection. In the first stage, an equal number of top-ranked features is selected for each feature group using Mutual Information. In the second stage, swarm particle optimization removes the redundant features identified during the first stage. Using several machine learning models and balanced accuracy as the metric for evaluation, Abbasi et al. observed that their proposed feature selection method performs significantly better for multi-class classification but showed comparable performance for binary classification when compared with the feature selection method used by Sgandurra et al. [25]. Moreira et al. [22] utilised six machine learning models such as Naive Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM) to analyse ransomware attacks. Using balanced accuracy as the metric for evaluating the model performance, they found out that the random forest model outperformed other models in detecting ransomware. Also, using their newly developed metric for feature group relevance, they concluded that Application Programming Interface (API) calls are the most relevant feature group to distinguish ransomware from goodware. Khan et al. [16] proposed a machine learning-based digital DNA sequencing engine for detecting and classifying ransomware called DNAact-Ran. In the preprocessing stage, DNAact-Ran used Multi-Objective Grey Wolf Optimization (MOGWO) and Binary Cuckoo Search (BCS) algorithms to select key features and then applied design constraints of DNA sequence and k-mer frequency vector to the selected features to generate the digital DNA sequence. Differently from previous work, we propose a three-stage feature selection method that significantly improves ransomware classification.

However, most studies did not consider the data imbalance problem that arises mainly in the multi-class classification of ransomware families. Data imbalance can result in a learning model's poor prediction of the minority class

samples [27]. Previous research has focused on correction of imbalanced data of Android ransomware [2] or general malware detection [4,15,23]. The approaches proposed in the literature can be divided into two categories: resampling techniques and cost sensitive learning [11]. Resampling techniques involve modifying the distribution of samples in the dataset by oversampling the minority class or undersampling the majority class or both [23]. Cost-sensitive learning aims to adjust the classification threshold to account for the cost of misclassification for each class [30]. The approach to address class imbalance may also depend on the characteristics of the data.

Thus, the techniques used to address data imbalance for malware detection may not be effective for ransomware detection because the characteristics of ransomware samples may be different from those of other types of malware samples. We use cost-sensitive learning and SMOTE to deal with the data imbalance problem.

Sgandurra et al. [25] applied the Regularized Logistic Regression to analyse and classify ransomware. To achieve this aim, they developed a ransomware dataset called Elderan by performing dynamic analysis of ransomware and goodware samples in the Cuckoo sandbox (a controlled environment for safely executing potentially malicious software). Sgandurra et al. applied Mutual Information to select the top 400 features out of 30,967 features contained in the dataset. This feature selection method did not consider the varying importance of the different feature groups to improve the classification of ransomware families.

3 Research Approach

We used the Elderan ransomware dataset [25] that was created using dynamic analysis on ransomware and goodware samples. The lack of a publicly available dataset for ransomware classification is a known problem [3,7,12]. Although the Elderan dataset is not large, it is one of the most comprehensive ransomware datasets publicly available.

We adopted two feature selection methods in this study: (1) mutual information as used by [25] and (2) our proposed multi-stage feature selection method. We use these methods to train machine learning models (XGBoost, LR, RF, DT and SVM) for both the binary and multi-class classification of ransomware. The binary classification problem discriminates ransomware samples from goodware samples, while the multi-class classification problem aims to distinguish between the different ransomware families. After that, we used SMOTE to separately resample the training data for the binary and multi-class classification. Using the re-sampled data, we adopted the same machine learning models for the binary and multi-class ransomware classification. Similarly, we also used cost-sensitive machine learning models for the binary and multi-class ransomware classification.

3.1 Research Questions

In this study, we aim to answer the following research questions:

1. What is the difference in performance between the proposed three-stage feature selection method and other state-of-the-art studies in the classification of ransomware?
2. Which technique for addressing the data imbalance problem is more effective in detecting and classifying ransomware?

3.2 Dataset Description

The Elderan dataset includes 1524 software samples. Out of those samples, 942 were classified as goodware, and 582 were classified as ransomware. The ransomware families included in the dataset are Citroni, CryptoLocker, Kovter, Locker, Matsnu, Pgpcoder, Reveton, TeslaCrypt, Trojan-Ransom, CryptoWall, and Kollah.

The total number of features in the dataset is 30,967, grouped into 7 feature groups namely Application Programming Interface invocations (API), Extensions of the dropped files (DROP), Registry key operations (REG), File operations (FILE), Extension of the files involved in file operations (FILES EXT), File directory operations (DIR), and Embedded strings (STR). Every feature has a value (0 or 1) representing the absence or presence of the corresponding operation. The distribution of the different feature groups is shown in Table 1.

3.3 Feature Selection Technique

Feature selection is a crucial data preprocessing step in the fields of data mining and machine learning. It aims to reduce the number of features used in the analysis, making the models simpler, more interpretable, and computationally efficient while avoiding the curse of dimensionality [18].

The study by Sgandurra et al. [25] employed Mutual Information (MI) [26] as a feature selection technique. MI quantifies the discrimination power of each feature in the classifier. Sgandurra et al. [25] showed that for the Elderan dataset, the maximum performance was achieved by selecting the top 400 features based on mutual information. However, this approach did not consider the varying importance of different feature groups as they selected just the top 400 features irrespective of the feature group they belong.

We propose a three-stage feature selection method that addresses this limitation by taking into account the varying significance of different feature groups. Our method involves splitting the data into feature groups and applying three different feature selection techniques within each group to select the relevant data.

Stage I: In the first stage, we used chi-square (CHI2) to select the top 200 features from each feature group, resulting in 1400 features from the seven groups. We considered 200 features because the smallest feature group has 233 features. CHI2 is a statistical hypothesis test that compares observed and expected frequencies of a categorical variable. The test assumes observed frequencies follow

a chi-squared distribution and calculates a test statistic to determine the significance of differences between observed and expected frequencies [29].

Stage II: In stage II, we used the Duplicated Features (DUF) method to select features from stage I. Duplicate features are redundant and provide no extra information. They can also cause problems with machine learning algorithms and lead to overfitting. It's advised to remove them before creating a model to prevent clutter and make analysis easier. Keeping duplicate features causes multicollinearity.

Stage III: In stage III, we applied Constant Features (COF) filter feature selection method to the remaining features from stage II. Constant features are those that have only one value for all entries in the dataset. These constant features can hinder the performance of a machine learning model and, thus, should be removed.

Table 1: Features remaining after applying each stage of the proposed feature selection method

Feature Groups	**API**	**DROP**	**REG**	**FILES**	**FILES_EXT**	**DIR**	**STR**	**Total**
All Features	233	346	6622	4141	935	2424	16267	30968
Stage I: CHI2	200	200	200	200	200	200	200	1400
Stage II: DUF	192	126	116	90	131	44	144	843
Stage III: COF	151	20	112	27	59	31	66	**466**

3.4 Machine Learning Models

The machine learning models that we employed to detect and classify ransomware were logistic regression, random forest, eXtreme Gradient Boosting (XGBoost), Decision Trees, and Support Vector Machine (SVM). These models have been shown to be effective in the detection of ransomware [1, 22].

3.5 Data Imbalance Techniques

Machine learning models typically assume balanced classes in datasets, so highly imbalanced data can cause classifier performance to decrease [27]. The machine learning community has addressed data imbalance in two ways: by resampling the dataset (through oversampling the minority class, undersampling the majority class, or a combination of both); or by cost-sensitive learning (assigning different costs to training examples) [11].

In this study, we used two methods to address data imbalance. The first is SMOTE [11], which generates synthetic examples of minority classes to balance

the number of samples with majority classes. The second is cost-sensitive machine learning, which adjusts models to prioritize classes with fewer examples in the training dataset [9]. Misclassification costs are assigned to instances to minimize the total misclassification cost instead of optimizing accuracy. A cost matrix assigns a cost to each cell in the confusion matrix, with weights based on the inverse proportions of class frequencies in the input data [19]. These methods were chosen for their success in correcting imbalanced data [6, 28].

3.6 Evaluation Metrics

To determine the best machine learning model for detecting and classifying ransomware, we evaluated their performance using balanced accuracy as a more suitable metric than accuracy for imbalanced datasets. Balanced accuracy is the arithmetic mean of specificity and sensitivity. Sensitivity measures the proportion of real positives that are correctly predicted while specificity measures the proportion of correctly identified negatives. We also used other evaluation metrics such as precision, recall, F1, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

3.7 Setup

The clean dataset used required no preprocessing. The data was divided into two sets - 80% for training and 20% for independent testing. A stratified train-test split was used to maintain class proportions, which is more effective for imbalanced datasets. Repeated stratified k-fold cross-validation was employed with a split of 3 and a repeat of 4, as it is better for imbalanced datasets [9]. Gridsearch was used to tune hyperparameters, and the model's performance was evaluated on the 20% independent test set.

4 Experimental Results and Discussion

Tables 2 and 3 summarise binary and multi-class classification results for ransomware, comparing the balanced accuracy of different machine learning models and data balancing techniques. We report the results for both the cross-validation test (CV test) and the independent test (ID test).

The proposed three-stage feature selection method outperformed Sgandurra et al.'s method significantly (p-value < 0.05), with an average improvement of 10% for binary and 21.79% for multi-class classification. For binary classification, XGBoost with cost-sensitive learning and SMOTE performed best, while for multi-class classification, the random forest model using cost-sensitive learning achieved the highest balanced accuracy of 61.94%. This study achieved better performance than other state-of-the-art studies as shown in Figure 2, especially in multi-class classification, which addressed the severe class imbalance problem using cost-sensitive learning and SMOTE. The comparison for multi-class was

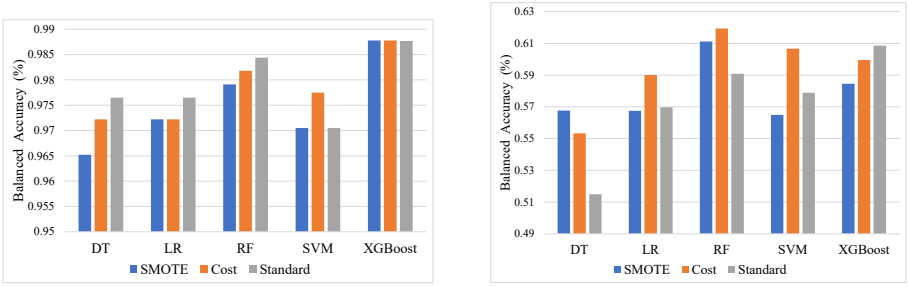
Table 2: Balanced Accuracy for 4 Repeats Stratified 3-fold Cross-validation with Standard Deviation and Independent Test for binary classification

Type	Classifiers	Sgandurra [25] FS Method		Our FS Method	
		CV Test	ID Test	CV Test	ID Test
Standard Approach	DT	0.8627 ± 0.0153	0.8838	0.9550 ± 0.0120	0.9765
	LR	0.8709 ± 0.0136	0.8648	0.9666 ± 0.0099	0.9765
	RF	0.8879 ± 0.0115	0.9001	0.9737 ± 0.0113	0.9844
	SVM	0.8848 ± 0.0107	0.9126	0.9650 ± 0.0073	0.9705
	XGBoost	0.8703 ± 0.0116	0.8841	0.9716 ± 0.0097	0.9877
	Baseline	-	0.8579	-	0.9678
SMOTE	DT	0.8796 ± 0.0144	0.8859	0.9612 ± 0.0083	0.9652
	LR	0.8805 ± 0.0160	0.8808	0.9741 ± 0.0072	0.9722
	RF	0.8992 ± 0.0169	0.9096	0.9753 ± 0.0080	0.9791
	SVM	0.8996 ± 0.0151	0.8948	0.9683 ± 0.0059	0.9705
	XGBoost	0.8893 ± 0.0146	0.8818	0.9761 ± 0.0072	0.9878
Cost Sensitive Learning	DT	0.8605 ± 0.0142	0.8686	0.9685 ± 0.0096	0.9722
	LR	0.8754 ± 0.0138	0.8739	0.9685 ± 0.0096	0.9722
	RF	0.8846 ± 0.0115	0.9080	0.9742 ± 0.0092	0.9818
	SVM	0.8833 ± 0.0102	0.9017	0.9668 ± 0.0139	0.9775
	XGBoost	0.8761 ± 0.0095	0.8914	0.9716 ± 0.0097	0.9878

Table 3: Balanced Accuracy for 4 Repeats Stratified 3-fold Cross-validation with Standard Deviation and Independent Test for multi-class classification

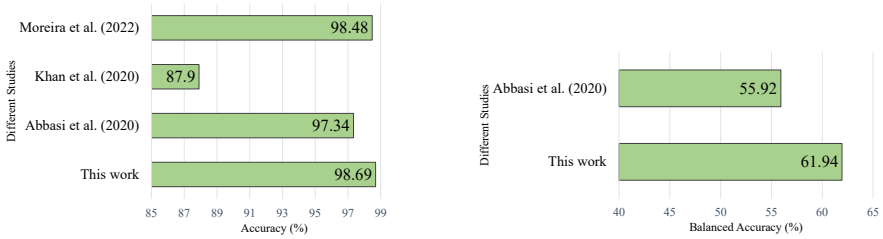
Type	Classifiers	Sgandurra [25] FS Method		Our FS Method	
		CV Test	ID Test	CV Test	ID Test
Standard Approach	DT	0.3707 ± 0.0198	0.3193	0.4631 ± 0.0364	0.5149
	LR	0.3641 ± 0.0254	0.2833	0.4984 ± 0.0198	0.5697
	RF	0.3842 ± 0.0246	0.3237	0.4999 ± 0.0363	0.5909
	SVM	0.3703 ± 0.0179	0.3237	0.4890 ± 0.0207	0.5789
	XGBoost	0.3750 ± 0.0145	0.3334	0.4950 ± 0.0193	0.6086
	Baseline	-	0.2833	-	0.5828
SMOTE	DT	0.6817 ± 0.0093	0.3887	0.8979 ± 0.0058	0.5677
	LR	0.6399 ± 0.0081	0.3514	0.8966 ± 0.0053	0.5675
	RF	0.6866 ± 0.0093	0.4097	0.9069 ± 0.0046	0.6112
	SVM	0.6858 ± 0.0088	0.3882	0.9037 ± 0.0044	0.5649
	XGBoost	0.6847 ± 0.0086	0.4088	0.9060 ± 0.0048	0.5846
Cost Sensitive Learning	DT	0.4107 ± 0.0236	0.4057	0.4543 ± 0.0390	0.5533
	LR	0.4007 ± 0.0204	0.3430	0.5167 ± 0.0387	0.5901
	RF	0.4218 ± 0.0251	0.4084	0.5013 ± 0.0469	0.6194
	SVM	0.4086 ± 0.0235	0.3935	0.5003 ± 0.0333	0.6067
	XGBoost	-	0.3973	-	0.5995

only done with the work of [1], as it was the only study among the compared studies that specifically addressed the multi-class classification problem.



(a) Binary Classification (b) Multi-class Classification

Fig. 1: Performance across the 3 approaches using the different algorithms using our FS method



(a) Binary Classification (b) Multi-class Classification

Fig. 2: Comparison of the highest classification accuracy of ransomware in [1, 16, 22] and the present study

4.1 Threats to Validity

The study's internal validity relies on the choice of feature selection, data balancing, and machine learning models, but other techniques could produce different outcomes. The results are limited to using SMOTE and cost-sensitive learning, and the dataset used for analysis only includes older ransomware families. Construct validity was upheld by reviewing and testing the codes multiple times, with care taken when creating and evaluating synthetic samples.

5 Conclusion

In this paper, we propose a three-stage feature selection method to reduce data dimensionality while considering the composition of feature groups, improving ransomware detection compared to previous studies. We address data imbalance using SMOTE and cost-sensitive machine learning for binary and multi-class detection and classification. Our method outperforms state-of-the-art results by 6.02% in multi-class classification. The best binary classifier is XGBoost with cost-sensitive learning and SMOTE (98.78%), followed by XGBoost with the standard approach (98.77%). The best multi-class classifier is the random forest model using cost-sensitive learning (61.94%). Improvements could be made by using other data balancing techniques and creating a comprehensive dataset with current ransomware samples such as Wannacry and Conti. Furthermore, future work can be done by analysing the ransomware features to investigate their contributions to the model outcome.

Acknowledgements This work was funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183).

References

1. Abbasi, M.S., Al-Sahaf, H., Welch, I.: Particle swarm optimization: A wrapper-based feature selection method for ransomware detection and classification. In: International Conference on the Applications of Evolutionary Computation (Part of EvoStar). pp. 181–196. Springer (2020)
2. Almomani, I., Qaddoura, R., Habib, M., Alsoghyer, S., Al Khayer, A., Aljarah, I., Faris, H.: Android ransomware detection based on a hybrid evolutionary approach in the context of highly imbalanced data. IEEE Access **9**, 57674–57691 (2021)
3. Almousa, M., Basavaraju, S., Anwar, M.: Api-based ransomware detection using machine learning-based threat detection models. In: 2021 18th International Conference on Privacy, Security and Trust (PST). pp. 1–7. IEEE (2021)
4. Aurangzeb, S., Anwar, H., Naeem, M.A., Aleem, M.: Bigrc-eml: Big-data based ransomware classification using ensemble machine learning. Cluster Computing **25**(5), 3405–3422 (2022)
5. Avila, R., Khoury, R., Pere, C., Khanmohammadi, K.: Employing feature selection to improve the performance of intrusion detection systems. In: International Symposium on Foundations and Practice of Security. pp. 93–112. Springer (2022)
6. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter **6**(1), 20–29 (2004)
7. Beaman, C., Barkworth, A., Akande, T.D., Hakak, S., Khan, M.K.: Ransomware: Recent advances, analysis, challenges and future research directions. Computers & Security **111**, 102490 (2021)
8. Bolón-Canedo, V., Alonso-Betanzos, A.: Ensembles for feature selection: A review and future trends. Information Fusion **52**, 1–12 (2019)
9. Brownlee, J.: Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning. Machine Learning Mastery (2020)

10. Cai, J., Luo, J., Wang, S., Yang, S.: Feature selection in machine learning: A new perspective. Neurocomputing **300**, 70–79 (2018)
11. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
12. Chen, Q., Bridges, R.A.: Automated behavioral analysis of malware: A case study of wannacry ransomware. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 454–460. IEEE (2017)
13. Collier, R.: Nhs ransomware attack spreads worldwide (2017)
14. Cyber Security Policy: Securing cyber resilience in health and care: October 2018 progress update (2018), https://www.gov.uk/government/publications/securing-cyber-resilience-in-health-and-care-october-2018-update
15. Goyal, M., Kumar, R.: Machine learning for malware detection on balanced and imbalanced datasets. In: 2020 International Conference on Decision Aid Sciences and Application (DASA). pp. 867–871. IEEE (2020)
16. Khan, F., Ncube, C., Ramasamy, L.K., Kadry, S., Nam, Y.: A digital dna sequencing engine for ransomware detection using machine learning. IEEE Access **8**, 119710–119719 (2020)
17. Kshetri, N., Voas, J.: Do crypto-currencies fuel ransomware? IT professional **19**(5), 11–15 (2017)
18. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection: A data perspective. ACM computing surveys (CSUR) **50**(6), 1–45 (2017)
19. Ma, Y., He, H.: Imbalanced learning: foundations, algorithms, and applications (2013)
20. McIntosh, T., Kayes, A., Chen, Y.P.P., Ng, A., Watters, P.: Ransomware mitigation in the modern era: A comprehensive review, research challenges, and future directions. ACM Computing Surveys (CSUR) **54**(9), 1–36 (2021)
21. Meland, P.H., Bayoumy, Y.F.F., Sindre, G.: The ransomware-as-a-service economy within the darknet. Computers & Security **92**, 101762 (2020)
22. Moreira, C.C., de Sales Jr, C.d.S., Moreira, D.C.: Understanding ransomware actions through behavioral feature analysis. Journal of Communication and Information Systems **37**(1), 61–76 (2022)
23. Pang, Y., Peng, L., Chen, Z., Yang, B., Zhang, H.: Imbalanced learning based on adaptive weighting and gaussian function synthesizing with an application on android malware detection. Information Sciences **484**, 95–112 (2019)
24. Rieck, K., Trinius, P., Willems, C., Holz, T.: Automatic analysis of malware behavior using machine learning. Journal of computer security **19**(4), 639–668 (2011)
25. Sgandurra, D., Muñoz-González, L., Mohsen, R., Lupu, E.C.: Automated dynamic analysis of ransomware: Benefits, limitations and use for detection. arXiv preprint arXiv:1609.03020 (2016)
26. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
27. Thabtah, F., Hammoud, S., Kamalov, F., Gonsalves, A.: Data imbalance in classification: Experimental evaluation. Information Sciences **513**, 429–441 (2020)
28. Thai-Nghe, N., Gantner, Z., Schmidt-Thieme, L.: Cost-sensitive learning methods for imbalanced data. In: The 2010 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2010)
29. Urdan, T.C.: Statistics in plain English. Routledge (2011)
30. Wu, D., Guo, P., Wang, P.: Malware detection based on cascading xgboost and cost sensitive. In: 2020 International Conference on Computer Communication and Network Security (CCNS). pp. 201–205. IEEE (2020)