v. 2, n. 25, 2022

Journal of
Engineering
Research

RANSOMWARE CLASSIFICATION BY MACHINE LEARNING AND DIMENSIONALITY REDUCTION

George Tassiano Melo Pereira

Claudomiro de Souza de Sales Júnior
http://lattes.cnpq.br/4742268936279649

DOI 10.22533/at.ed.3172252201116

Abstract: Ransomware is a type of malware that aims to take control of systems that host or encrypt data - until a ransom is paid. This threat, which has been seen as a new type of terrorism, is a difficult task given the rapid spread and changes developers apply to encryption techniques. Given this, Machine Learning (ML) classifier algorithms have been reported as promising tools for classifying ransomware. This work explores 7 ML techniques in order to make 5 types of approaches, along with 2 dimensionality reduction techniques. The Gaussian Process presented the best performance, as it proved to be effective in four approaches.

INTRODUCTION

According to a report by the cybersecurity company *Cybersecurity Ventures*, it was predicted that in 2021 the *ransomware* will bring global damage estimated at $6 trillion [Freedman 2020]. *Ransomware* gangs are increasingly improving and reaching a high number of victims, whether common end users, business organizations, federal, banks, city halls, doctors, etc.

Ransomware is malware designed to restrict access to a system or data until a ransom amount demanded by the cybercriminal is satisfied. It can be classified into four stages. In the first phase, there is an attempt to invade through malicious users and applications, spam, phishing, etc. In the second phase, there is communication with the hacker's command and control server. In the third phase there is destruction, which can be of two types: encrypting victims' files or locking the machine, preventing victims from accessing their systems. In the fourth phase there is extortion, where the hacker demands a ransom payment to release files or access to the system.

The constant evolution of ransomware as well as applied encryption, the lack of a universal solution to prevent ransomware attacks the internal vulnerabilities, delayed updates, the effects of pandemic conditions, ransomware attacks operated by humans, among other challenges, are some open issues in a research on ransomware. Furthermore, ransomware generates a large amount of high-dimensional data, and making a classification without using computational intelligence consumes a high cost of time and processing. Processing and analyzing high-dimensional data have become a challenge for researchers working in various disciplines, especially machine learning and data mining [Cusack et al. 2018].

Therefore, dimensionality reduction can be used to extract attributes, reducing the dimensional space, given that, in classification problems with a finite number of samples and with a very high amount of attributes, a series of effects can occur. negatives known in the literature as the curse of dimensionality. In this research, Principal Component Analysis (PCA) and Kernel PCA (KPCA) techniques were used.

PCA is an unsupervised linear transformation algorithm that produces new features, called principal components, by determining the maximum variance of the data. The PCA projects the highly dimensional dataset to a new subspace where the orthogonal axes are considered as the directions of maximum variation of the data, while the Kernel PCA (KPCA) maps the input data to a higher dimensional space before to reduce dimensionality. This causes nonlinearities to be incorporated into the input data. With Kernel PCA, the only requirement of the method is to build the Kernel array from the input data. The solution consists in finding the eigenvectors associated with the largest eigenvalues of the Kernel matrix.

Machine learning (ML) has been effective in classifying ransomware as shown in the authors' research [Adamu and Awan 2019], [Abbasi et al. 2020], and [Fernando et al. 2020]. When compared to static code analysis techniques, machine learning techniques can analyze the similarity of malware behavior in vector space, with clustering approaches as well as different classification algorithms, and with that new malware variants can be identified, as shown in the study by [Zakaria et al. 2017].

In this article, we explore 7 ML classification algorithms, such as: Logistic Regression (RL), Support Vector Machine (SVM), Gaussian Process (GP), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), and Multi-layer Perceptron (MLP), along with the application of PCA and PCA kernel (KPCA) techniques for dimensionality reduction data in order to find the best classification algorithms for the problem, approaching five classification methods used in the literature.

STATE OF ART

Machine learning has been very effective in detecting malware, as shown in the work of [Fernando et al. 2020], [Milosevic et al. 2017] [Hwang et al. 2020], [Alenazi et al. 2019] and [Adamu and Awan 2019]. Ransomware has been a very active research topic and several researchers have proposed research that focuses on different aspects of ransomware research.

For a better analysis of ransomware and all its evolution, the work of [Oz et al. 2021] conducted a comprehensive survey of ransomware and defense solutions in relation to PCs/workstations, mobile devices and Cyber-Physical Systems (CPS) and Internet of Things (IoT) platforms. The survey covered 137 studies during the period 1990-2020, presented a detailed overview of ransomware evolution, comprehensively analyzed the main building blocks of ransomware, presented a taxonomy of the most notable ransomware families and listed a number of open questions for future ransomware research.

The work of [Fernando et al. 2020] conducted research on the evolution of ransomware detection using machine learning and deep learning techniques. The article evaluated 19 works, using the algorithmic approach, the resource engineering process, as well as the evaluation of each result. The article is extremely relevant as it explores the new directions of ransomware and how ransomware is expected to evolve in the coming years.

The work of [Abbasi et al. 2020] proposed a feature selection method that uses particle swarm optimization (PSO) for ransomware detection and classification using high-dimensional ransomware and goodware (benign software) behavior analysis data. The article results show that the model's ranking performance depends on the number of features selected from each of the feature groups in the dataset. The authors achieved an average accuracy of 50% for ransomware families.

The work of [Borah et al. 2021] performed a classification called ERAND (Ensemble Ransomware Defense) for defense against ransomware. The authors used the NSGA-II to calculate the weights of five classifiers (ExtraTree, Gradient Boosting, AdaBoost, XGBoost and Random Forest) and achieved high accuracy, finding accuracies for each family above 95%. However, the methodology used by the authors became quite obscure, and caused several different interpretations.

The work by [Adamu and Awan 2019] performed a ransomware prediction using supervised learning with Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT), Naive Baye (NB), Artificial Neural Network (AN) algorithms), Logistic Regression (RL). The research achieved a binary classification performance of 88.2% with SVM, 65.7% with Logistic Regression, 84% with Random Forest, 52.5% with Naive Bayes and 86% with MLP.

The authors' research [Silva et al. 2020] aimed to determine which machine learning algorithm performs best in classifying ransomware, using static and dynamic analysis techniques. The authors used the Naive Bayes, SVM and Decision Tree algorithms in the classification and reached an accuracy of 81.65%, 70.79% and 75.66%, respectively.

The work of [Cusack et al. 2018] proposed a ransomware detection model based on machine learning methods using network traffic data. The researchers monitored the network communication between the victim's machine and the command and control (C&C) to detect and prevent the delivery of the encryption key needed to encrypt the victim's files without which the encryption process did not start. The authors used dimensionality reduction techniques to find the eight attributes that most contribute to the detection of ransomware in network traffic. However, the solution suffers from having a 12.5% false positive rate, which can generate many false alarms. We will use for our comparisons the works of [Adamu and Awan 2019], [Silva et al. 2020] and [Abbasi et al. 2020] and in which the authors worked with the same dataset and with some algorithms that we used for our experiments.

MATERIALS AND METHODS

DATA COLLECTION

Ransomware samples from the dataset used for this dissertation were downloaded from VirusShare, a website that maintains a continuously updated database of malware for various operating systems. They were made available by [Sgandurra et al. 2016] in which he proposed an anti-ransomware solution, called Elderan as mentioned in the literature review. The samples were run for 30 seconds in a dynamic analysis environment to record their behavior in terms of the operations of the program used. In total, 30,967 attributes categorized into seven groups were recorded. The dataset was analyzed at the end of February 2016, and consists of 582 ransomware working samples belonging to 11 different classes and 942 goodware. Table 1 shows the summary of instances from the original dataset and the instances used after preprocessing.

Family	**Original Dataset Instances**	**Instances after Pre'-Processing**
Citroni	50	34
CryptLocker	107	100
CryptoWall	46	37
Kollah	25	20
Kovter	64	57
Locker	97	96
Matsnu	59	46
Pgpcoder	4	4
Reverton	90	56
TeslaCrypt	6	6
Trojan-Ransom	34	28
Goodware	942	928
Total	1524	**1412**

Table 1. Summary of instances used.

DATA PRE-PROCESSING

First, duplicate instances of the same class were removed, where 108 instances were found. Then the duplicated instances of different classes were removed, in which 4 were found, thus totaling 112 removals of instances from the dataset. This process was made to not use redundant data and to avoid possible conflicts, since it is not possible to differentiate what a classifier must consider as a class. Table 2 presents a summary of the feature group in the dataset.

Input data was encoded as binary values, where 0 denotes no execution and

1 returns execution of a system call. As each classification technique can have its highest accuracy with different number of components used, an adjustment was made to the number of components used per technique defined randomly from 1 component to a maximum total for all techniques.

Group	Description	# Features
API	API calls	232
DROP	Removed file extensions	346
REG	Key Registry Operations	6622
FILES	File Operations	4141
FILES EXT	Manipulated file extensions	935
DIR	File directory operations	2424
STR	built-in strings	16267
Total	-	30967

Table 2. Summary of the dataset's feature group.

HYPERPARAMETER ADJUSTMENTS

Some techniques for adjustments, such as cross-validation (Cross-Validation) and Grid Search, with hyperparameter adjustments, were applied to reduce the possibility of over-fitting. For binary classification, Cross-Validation of the Stratified type with 10-fold was used and in the other classifications 4-fold was adopted, due to the existence of only 4 instances in the family class (Pgpcoder), as shown in Table 1. Cross-Validation of the Stratified type maintains the same proportion of classes in all folds. Also implemented in Scikit-learn is the" Randomized Search CV" library (randomized search cross-validation), in which the algorithm chooses the most successful version of the model seen after training N different versions of the model with different combinations of randomly selected hyperparameters, leaving with a model trained on a near-ideal set of hyperparameters.

METHODOLOGY OVERVIEW

The database, composed of a group of features (API, DROP, REG, FILES EXT, DIR, STR) with 942 goodwares and 585 ransomwares, undergoes data pre-processing in order to remove duplicates and irrelevant data. Then cross-validation is performed to obtain training and test data for training and validating the models used.

The construction of the model responsible for classifying the data is performed by combining seven machine learning algorithms and two dimensionality reduction techniques that transform the model's input data into principal components. With the test results, the adjustment of the hyperparameters is carried out, where the best configuration found in the model is stored. Finally, there is a comparison of the results obtained from the models built and optimized to obtain the best model found. The illustration of the applied methodology is defined according to (Figure 1).

EVALUATION OF RESULTS

In the evaluation of the results, the accuracy was used to evaluate the performance of each classifier method. It measures the hit ratio among all data samples, and it can be calculated from equation 1.

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

on what:
TP = True Negative
TN = True Positive
FN = False Negative
FP = False Positive

RESULTS AND DISCUSSIONS

The ransomware issue represents a significant challenge for Information Technology security researchers and experts. With rapid dissemination, and the

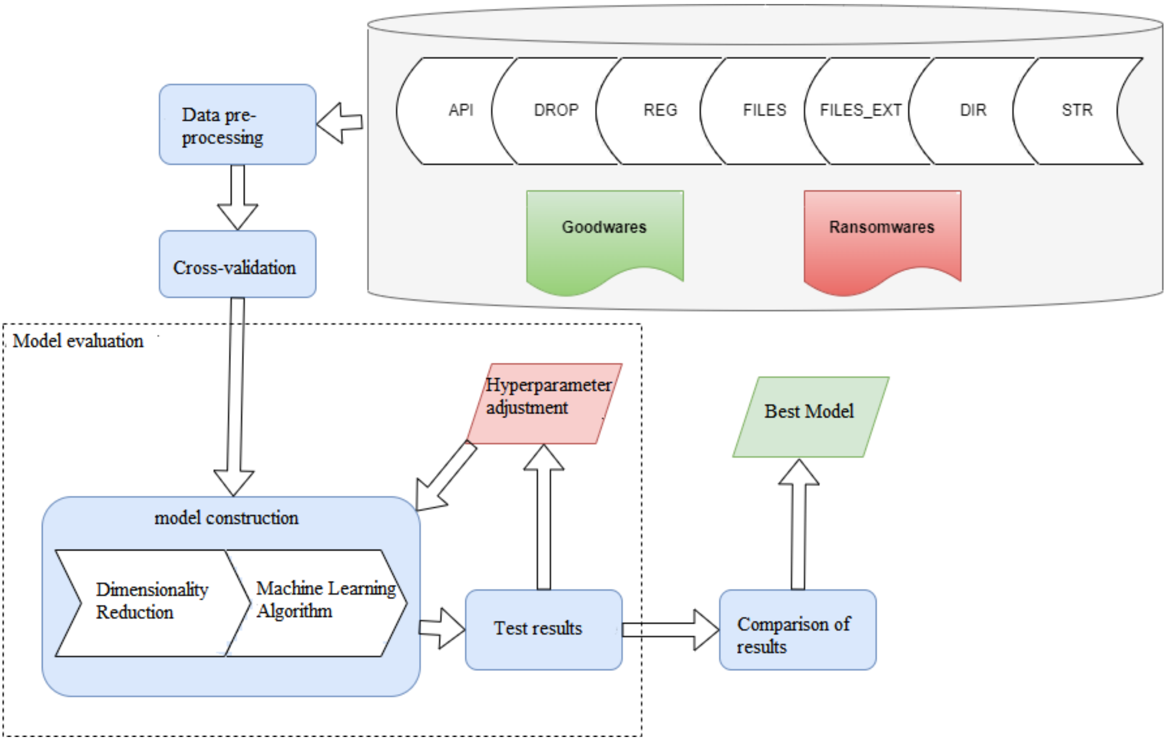

Figure 1. Illustration of the applied methodology. Source: Author.

development of sophisticated encryption techniques, studies related to the fight against ransomware require researchers to present constant approaches to follow evolving techniques.

In this context, machine learning techniques based on predictive models have been investigated in the literature as a promising tool for detecting malicious code.

This article assists in this effort, presenting a dynamic approach and classifying ransomware considering five approaches, seven machine learning classifier algorithms, taking into account two dimensionality reduction techniques.

Binary classification is a supervised machine learning task in which it is used to predict which of two classes (categories) a data instance belongs to. In this perspective, the binary classification scenario used in this work is determining whether a given instance is *a ransomware or goodware*, represented respectively by the integers 1 or 0.

Analyzing the performance of the binary classification, the methodology of this article stands out, as it achieved an accuracy of 97.73% as shown in Table 3, while the authors [Adamu and Awan 2019], [Silva et al. 2020], [Sgandurra et al. 2016] and [Abbasi et al. 2020], reached 86%, 82.40%, 96.34% and 97.34, respectively.

The purpose of the goodware family classification is to determine if a given instance belongs to one of the eleven ransomware families (described in Table 1), represented by an integer from 1 to 11, or if the instance belongs to the ransomware class. goodwares represented by the number 0. This multiclass classification achieved 84.56% accuracy with the Random Forest, SVM and Gaussian Process algorithms.

In the classification of families without goodwares, instances belonging to the goodwares class are not used, only using the 11 ransomware classes, where we sought to determine to which class each instance would

	Classifications					
	Binary		Family with goodwares		Family without goodwares	
Algorithms	PCA	KPCA	PCA	KPCA	PCA	KPCA
RF	96.17	96.24	**84.56**	83.64	**66.80**	**68.84**
SVM	96.45	87.81	**84.56**	77.69	60.33	20.86
GP	**97.37**	97.59	84.06	**84.56**	47.72	59.09
DT	93.34	91.92	73.01	76.84	38.63	42.14
NB	83.78	85.13	70.67	71.88	49.17	48.55
MLP	97.30	**97.73**	69.33	68.90	61.57	55.99
RL	97.02	93.91	82.15	81.56	60.53	58.05

Table 3. Accuracy of binary, family with goodware and family without goodware classifications, respectively.

belong. This methodology also surpassed the result of [Abbasi et al. 2020], as it achieved a performance of 55.81% accuracy, while this article reached 68.84% using the KPCA, as shown in Table 3.

In the classification by family class, instances of only a certain ransomware class were used, together with instances of the goodware class, thus performing a binary classification between ransomware and goodware. This classification aims at the specialization of classifiers in detecting a ransomware class. In this experiment, the Gaussian Process classifier obtained a 98.91% weighted average accuracy using the KPCA, as shown in Table 4. This result surpassed that of [Borah et al. 2021], given that it reached an average of 98.07%.

In variant classification, all instances of a given ransomware class are removed from being used during the classifier training stage, and then they are used separately for their validation. The classification is performed in a binary way to determine whether an instance is ransomware or goodware. This classification aims to validate whether the classifiers are capable of detecting new ransomware classes that may still arise. This experiment obtained 93.07% weighted average accuracy using the MLP classifier and the KPCA reduction technique, however [Borah et al. 2021] obtained 98.2% weighted average accuracy.

Analyzing all the proposed experiments, it is observed that the algorithm of Gaussian Process machine learning performed best in four out of five proposed approaches. The results were superior to the articles by [Adamu and Awan 2019], and [Silva et al. 2020], for example, due to the performance of hyperparameter adjustment, and the inclusion of PCA and kPCA techniques as options for dimensionality reduction.

	RF		SVM		GP		DT		NB		MLP		RL	
	PCA	KPCA	PCA	KPCA	PCA	KPCA	PCA	KPCA	PCA	KPCA	PCA	KPCA	PCA	KPCA
F1	98.33	98.75	99.16	98.64	99.37	99.27	97.71	98.12	98.54	98.44	98.96	99.06	99.06	99.16
F2	97.17	97.27	97.85	95.23	98.34	98.44	94.55	95.23	95.33	95.91	97.95	98.15	98.24	97.85
F3	97.71	97.82	98.54	97.30	98.96	99.06	96.58	97.61	98.13	98.23	98.96	98.86	98.86	97.20
F4	98.73	98.62	98.62	97.89	99.15	99.47	97.67	98.10	99.47	99.26	99.26	100	99.26	97.89
F5	97.56	98.17	98.88	97.56	99.28	99.39	95.73	95.73	98.07	98.57	98.88	99.39	99.28	99.18
F6	97.16	96.97	97.94	90.62	98.24	98.24	94.04	94.43	95.60	95.31	97.94	98.04	97.55	98.33
F7	97.43	97.12	98.57	97.22	98.87	99.07	96.20	96.09	97.12	97.63	98.66	98.46	97.84	98.35
F8	99.57	99.57	99.57	99.57	99.57	99.57	99.57	99.57	99.57	99.67	99.57	99.57	99.57	99.57
F9	97.66	98.27	98.98	94.30	98.98	99.08	96.74	94.61	97.86	97.76	99.08	99.05	99.08	97.35
F10	99.46	99.57	99.67	99.35	99.46	99.57	99.35	99.35	99.46	99.67	99.25	99.57	99.57	99.57
F11	97.80	98.22	99.05	97.38	98.74	100	97.07	97.28	98.53	98.43	98.84	98.84	98.43	98.74
MP	97.57	97.72	98.47	95.39	98.76	98.91	95.75	95.85	97.05	97.2	98.54	98.65	98.47	98.22

Table 4. Classification accuracy by family class.

	RF		SVM		GP		DT		NB		MLP		RL	
	PCA	KPCA	PCA	KPCA	PCA	KPCA	PCA	KPCA	PCA	KPCA	PCA	KPCA	PCA	KPCA
F1	78.67	61.76	94.11	97.05	100	100	72.79	88.23	44.11	41.17	100	94.11	94.11	97.05
F2	93	92	92.50	0	98.24	97	87	88.50	88	86	98	98	95	100
F3	81.08	81.75	93.24	0	95.94	97.29	79.05	81.08	68.24	48.64	97.29	97.29	86.48	97.29
F4	90	90	68.75	0	70	70	71.24	70	56.25	50	75	70	70	70
F5	61.84	63.59	98.24	0	96.49	96.49	62.28	52.63	51.31	47.36	91.22	94.29	96.49	96.49
F6	84.63	85.67	86.19	92.70	91.66	91.66	85.41	87.50	59.63	52.08	91.22	91.14	90.88	92.18
F7	89.67	95.65	97.82	95.65	90.21	93.47	88.58	77.17	70.10	58.69	92.93	96.19	90.21	97.82
F8	18.75	50	100	0	87.50	75	68.75	68.75	75	75	100	100	81.25	100
F9	73.21	72.76	74.55	92.85	83.48	83.33	81.69	89.28	61.16	62.50	80.80	85.71	84.37	67.85
F10	83.33	83.33	100	83.33	83.33	83.33	83.33	83.33	83.33	83.33	83.33	83.33	83.33	83.33
F11	82.14	91.07	100	100	99.10	96.42	81.25	75.89	75.89	78.57	100	100	100	100
MP	81.66	82.02	90.30	51.86	92.81	92.70	80.52	80.78	66.47	61.36	92.61	93.07	90.96	92.25

Table 5. Variant classification accuracy.

REFEREENCES

Abbasi, M. S., Al-Sahaf, H., and Welch, I. (2020). Particle swarm optimization: A wrapper-based feature selection method for ransomware detection and classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12104 LNCS, pages 181–196. Springer.

Adamu, U. and Awan, I. (2019). Ransomware prediction using supervised learning algo- rithms. In *Proceedings - 2019 International Conference on Future Internet of Things and Cloud, FiCloud 2019*, pages 57–63. Institute of Electrical and Electronics Engine- ers Inc.

Alenazi, F., Hindi, K., and AsSadhan, B. (2019). Fine-tuning na¨ıve bayes for imbalanced datasets. *International Conference on Data Science*.

Borah, P., Bhattacharyya, D. K., and Kalita, J. K. (2021). Cost effective method for ransomware detection: An ensemble approach. volume 12582 LNCS, pages 203–219. Springer Science and Business Media Deutschland GmbH.

Cusack, G., Michel, O., and Keller, E. (2018). Machine learning-based detection of ran- somware using sdn. In *SDN-NFVSec 2018 - Proceedings of the 2018 ACM Inter- national Workshop on Security in Software Defined Networks and Network Function Virtualization, Co-located with CODASPY 2018*, volume 2018-January, pages 1–6. Association for Computing Machinery, Inc.

Fernando, D. W., Komninos, N., and Chen, T. (2020). A study on the evolution of ran- somware detection using machine learning and deep learning techniques. *IoT*, 1:551– 604.

Freedman, L. F. (2020). Ransomware attacks predicted to occur every 11 seconds in 2021 with a cost of $20 billion — data privacy + cybersecurity insider.

Hwang, J., Kim, J., Lee, S., and Kim, K. (2020). Two-stage ransomware detection using dynamic analysis and machine learning techniques. *Wireless Personal Communicati- ons*, 112:2597–2609.

Milosevic, N., Dehghantanha, A., and Choo, K. K. R. (2017). Machine learning aided android malware classification. *Computers and Electrical Engineering*, 61:266–274.

Oz, H., Aris, A., Levi, A., and Uluagac, A. S. (2021). A survey on ransomware: Evolu- tion, taxonomy, and defense solutions.

Sgandurra, D., Mun˜oz-Gonza´lez, L., Mohsen, R., and Lupu, E. C. (2016). Automated dynamic analysis of ransomware: Benefits, limitations and use for detection.

Silva, A., Guelfi, A., Azevedo, M., and Sergio, K. (2020). Aplicac¸a˜o de algoritmos de aprendizado de ma´quina na unificac¸a˜o das te´cnicas de ana´lise esta´tica e dina^mica na classificac¸a˜o de ransomware. *PESQUISA EDUCAC¸ A˜O A DISTA^NCIA*.

Zakaria, W. Z. A., Abdollah, M. F., Mohd, O., and Ariffin, A. F. M. (2017). The rise of ransomware. In *Proceedings of the 2017 International Conference on Software and E- Business*, ICSEB 2017, page 66–70, New York, NY, USA. Association for Computing Machinery.