iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IV **Month of publication:** April 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

Analysis of Machine Learning Algorithms for Ransomware Detection and Behavioural - Based Classification

Dr. Nirmala Hiremani[1], Ranjitha Bai A[2]
[1, 2]*Department of Computer Science and Engineering, VTU University*

Abstract: ***A ransomware is a type of malware, the original data is encrypted from the user file and block the access until the ransom is recovered. As the ransomware changes their behavior, detection techniques and traditional classification do not accurately detect the new variants of ransomware. To avoid detection of signature-based systems the attackers use metamorphic and polymorphic techniques, the potential to change their code as they generate. Based on their behaviour using different classification method to identify modified variants of ransomware. The two main overview of this survey paper is to get an idea of techniques for identifying and classifying an unknown malwares into the respected families using techniques of machine learning.***

Keywords: Ransomware, Malware Detection, Classification, Malware, Behavioural Analysis.

I. INTRODUCTION

Now a day's individuals and organization been facing a serious cyber threat of ransomware attack all around the globe. Ransomware is often spread through an E-mails which contain an malicious links or through drive by downloading an online files, downloading of file occurs due to unknowingly visit an infected website and automatically malware is been installed without an user knowledge. The ransomware attacks in major different sectors such as business, hospitals, education, research and etc. According to a recent survey, reports steady and immense increases growth in new ransomware attacks and samples. There are wide range of variants in ransomware like crypto Locker, Locky, Tesla Crypto, Bad Rabbit and etc.

Ransomware is mainly classified into two types namely: Crypto ransomware and Locker ransomware. The locker ransomware which infects the computers by virus and locks the files until user pays the fines, whereas crypto ransomware encrypts valuable files on a mobile or computer so that user cannot access those files until the money is paid to an attacker. The newest ransomware is Pure Locker, identified by Intezer and IBM in November 2019.

According to the survey report of cyber security in the year 2018, nearly 67% of Indian enterprises have been hit by ransomware. Ransomware attacks are not only rising at the double rate globally but are also highly varying in ransom, demand up to $5,300,000 and $1,032,460 on average. The India is highest ransomware infected countries in the world, here are some of the latest ransomware attacks in India. Firstly, in March 2018 ransomware attack on UHBVN (uttar haryana bijli vitran nigam) attacked by ransomware where the hackers access to the systems of the power company and the billing data of customers was stooled. The attacker demanded up to Rs1crore or $10 million to restore access to the data. Secondly, ransomware attack on AP, Telangana power utility sites in May 2019 hacked the information about two crore power consumers which demand for some ransom. Thirdly in 2019 according to Kaspersky security ransomware was attacked on municipalities with the three group of ransomware namely ryuk, purge and stop.

II. LITERATURE SURVEY

Ala Bahrani, Amir Jalaly Bidgly [1]. The database that contains 21 ransomware families. The Proposed system which describes the identification of ransomware using a process mining method which discovers the process model from event logs and extract the features from the process model. Classifiers are used for the classification of ransomware to their respected families. The report which show j48 and random forest which gets the best accuracy among the all classifiers with the accuracy rate of 95% in detecting a ransomware.An easy way to comply with IJRASET paper formatting requirements is to use this document as a template and simply type your text into it.

Subash Poudyal, Kul Subedi, Dipankar Dasgupta [2]**.** In this paper the dataset which contains 178 ransomware samples from 13 different ransomware families. They have been used machine learning, static and reverse engineering frame-work incorporating feature generation to detect the ransomware. The detection rate of ransomware is significantly greater for dataset with combination of feature dataset with a minimum 89.18% for the logistic regression and maximum 97.95% for random forest.

Udayakumar N, Vatsal J.Saglani, Aayush V.Gupta, Subbulakshmi T [3]. In this paper they use various algorithms of machine learning like decision trees, SVM and neural networks. The feature of dataset is debug size with score of 0.26. The SVM for

classification which gets the accuracy of 90.2% and for neural network which gets an accuracy rate of 98.94%. The performance for testing data in neural network which gives an accuracy rate of 99.33%.

Anam Fatima, Ritesh Maurya Kishore Dutta, Radim Burget and Jan Masek [4]**.** This paper aim is to detect android malware based on machine learning algorithms, which use to train the machine using genetic algorithm as a proposed methodology for getting most optimized feature and also have a potential to identify the malware before and after compared with selected feature. Machine learning classifiers which gives an accuracy more than 94% using neural networks and support vector machine.

Ravi Kiran Varma P, Kotari Prudvi Raj and K.V Subba Raju [5]. The system collects 3,258 samples of android app, where from every application they extract the features and train the models. They used different machine learning algorithms such as Random Forest, Navies Bayes, multilayer perceptron, multi class classifier and j48 for detecting the android malware and each algorithms evaluate their performance. The faster classification of malware set has been performed by Naive Bayes classifier.

N.Poonguzhali, T.Rajakamalam, S.Uma and R.Manju [6]. In this paper they have used deep leaning for the detection of malware. The feature extraction and identification is done by a CNN (Convolutional neural network). The malicious codes have been converted into a gray scale images and these images is been classified using support vector machine classifier. Also handles with imbalance of the data using bio-inspired optimization technique also called as bat algorithm. The result of the model which gives the accuracy of 94.01% and speed as when compared to other models of malware detection.

George Cabau, Magda Buhu and Ciprian Oprisa [7]. This paper which explains an automatic system which classifies an infected file based on a dynamic behavior of the file within the environment of controlled monitored. It use a support vector machine classifier that is further use to identify the malicious files. The result where classifier collects the runtime and discriminate between malicious and clean samples.

Phai Vu Dinh, Nathan Shone, Phan Huy Dung, Qi Shi, Nguyen Viet Hung and Tran Nguyen Ngoc [8]. Here the dataset which contains 2,068 malware samples from 8 different malware families (Dridex, Kelihos, Locky, Ramnit, Sality, Simda, Vwtrak and Zeus). The behavior-aware malware classification, they have use 4 assessment measurements and 7 classifiers such as DT, KNN, LR, ANN, NB, RF and SVM. They have used 60% dataset for training and 40% dataset for testing. The result have enhances the malware classification with a unigram approach.

Barath Narayanan, Ouboti Djaney-Boundjou and Temesguen M.Kebede [9]. In this paper they apply (PCA) principal component analysis for implementation. For identification of malware data into their classes respectively, they have used various algorithms such as K-Nearest Neighbours (KNN), Support vector machine (SVM) and Artificial neural network (ANN). The result obtained in this paper where the performance of classification clearly that indicates PCA transformation is ideal in this scenario and the KNN classifier which gives the best performance among the all that is 96.6%.

Xin Zhou, Jianmin Pang and Guanghui Liang [10]. The total database of malware which contain 15,781 samples. The development of computer vision and machine learning, with the combining malware detection and the image classification. THE Gist features are been extracted from grayscale images of the malware with the help of gabour filter. The knowledge of machine learning with the randomized trees, K-Nearest neighbors, ET and GBDT as the classifier that achieved to detect the malware. The report shows the best performance classification model is ET with the accuracy of 96.19%.

III.PROPOSED IDEA

The proposed idea for the classification of malwares based on their behavioral and identifying them, where the malwares are classified using machine learning algorithms based on their behaviors as we know till now algorithms such as K- Nearest Neighbor(KNN), Support Vector Machines(SVM), Random Forest, Naive Bayes, Decision Tree, Linear Regression and etc.

As the result till now shows that they have the maximum accuracy around 96% based on their algorithms and the datasets which have been taken.

The disadvantage of the current result is that where is does not classify the malware to their respected families, that has high false positive rate with the less accuracy and the time taken for training data is high.

So here we can improve the classification techniques with new dataset which is of new malwares which has polymorphic and metamorphic techniques namely Trojans horses, Adware, Rootkits, Zmist and etc.

Using this new type of malware we can improve the classification accuracy more than 99% and which of less false positive values.

Here the proposed system is done using the software tool python IDE 2.7 and the hardware requirements are 1TB Hard disk, 4GB Ram and Intel Core i5 system.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

IV. EXISTING MACHINE LEARNING CLASSIFICATION ALGORITHMS

A. KNN (K-Nearest Neighbours)

A k-nearest neighbors is used to solve both classification and regression problems which is a non parametric method. The KNN simplest algorithm that stores all the data cases and classifies the new data or case based on a similarity measure, which uses entire data set in its training states. It is a supervised learning algorithm and Euclidean distance is used to determine the neighbors in K-nearest neighbors. Below equation determines the function gi(X) of KNN for a given test data sample X. The final result is done based on the class that provides maximum gi(X).

$$gi(X) = \frac{No\ of\ KNN\ labeled\ class\ i}{i}$$

B. SVM (Support Vector Machines)

The support vector machine is a discriminative classifier formally defined by a separating hyper plane. This support vector machine is what's called as binary classifiers that separates only into two classes at a time. SVM is specific to supervised learning and it can solve linear and nonlinear problems. In two dimensional space this hyper plane is a line dividing a plane in two parts where in each class lay in either side.

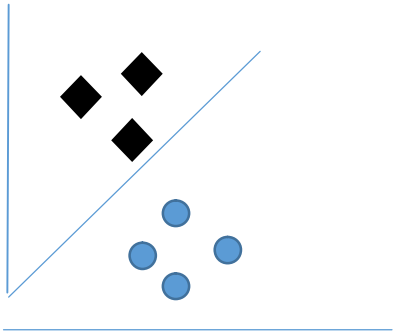


Fig. 1 separates the data into 2 classes with hyper line.

V. RANDOM FOREST

Random forest algorithm is used for regression and classification technique, it is a method that operates by constructing multiple decision trees during training phase.

Multiclass object detection is done using random forest algorithm. The random selection of features helps random forest to not only scale well when there exists many features per feature vectors, but also helps it in reducing the interdependence between the feature attributes and thus it is less vulnerable to inherent noise in the data.

Random forest that has higher accuracy and training time is less. Example they used in ETM Devices to acquire images of the earth's surface.

VI. NAIVE BAYES

The Naive Bayes classifier works on the principles of conditional probability as given by the Bayes theorem, it is the most effective classifier that requires a small amount of training data to estimate the necessary parameters. This Naive Bayes classifier is based on the supervised machine learning group based on the probabilistic logic. Naive Bayes classifier is extremely fast when compared to other sophisticated methods and also works in real world situations such as face recognition software and document classification.

VII. J48

The j48 algorithm was introduced by Ross Quinlan that is used to generate a decision tree. It is one of the best machine learning algorithm to survey the data continuously and categorically. This algorithm is used to find out the way the attributes-vector behaves for several of instances. The best attribute is to find on the basis of the present selection criterion and that attribute selected for branching. This j45 algorithm is almost used in the domains like machine learning, data mining and etc.

TABLE 1: Algorithms Comparison Matrix

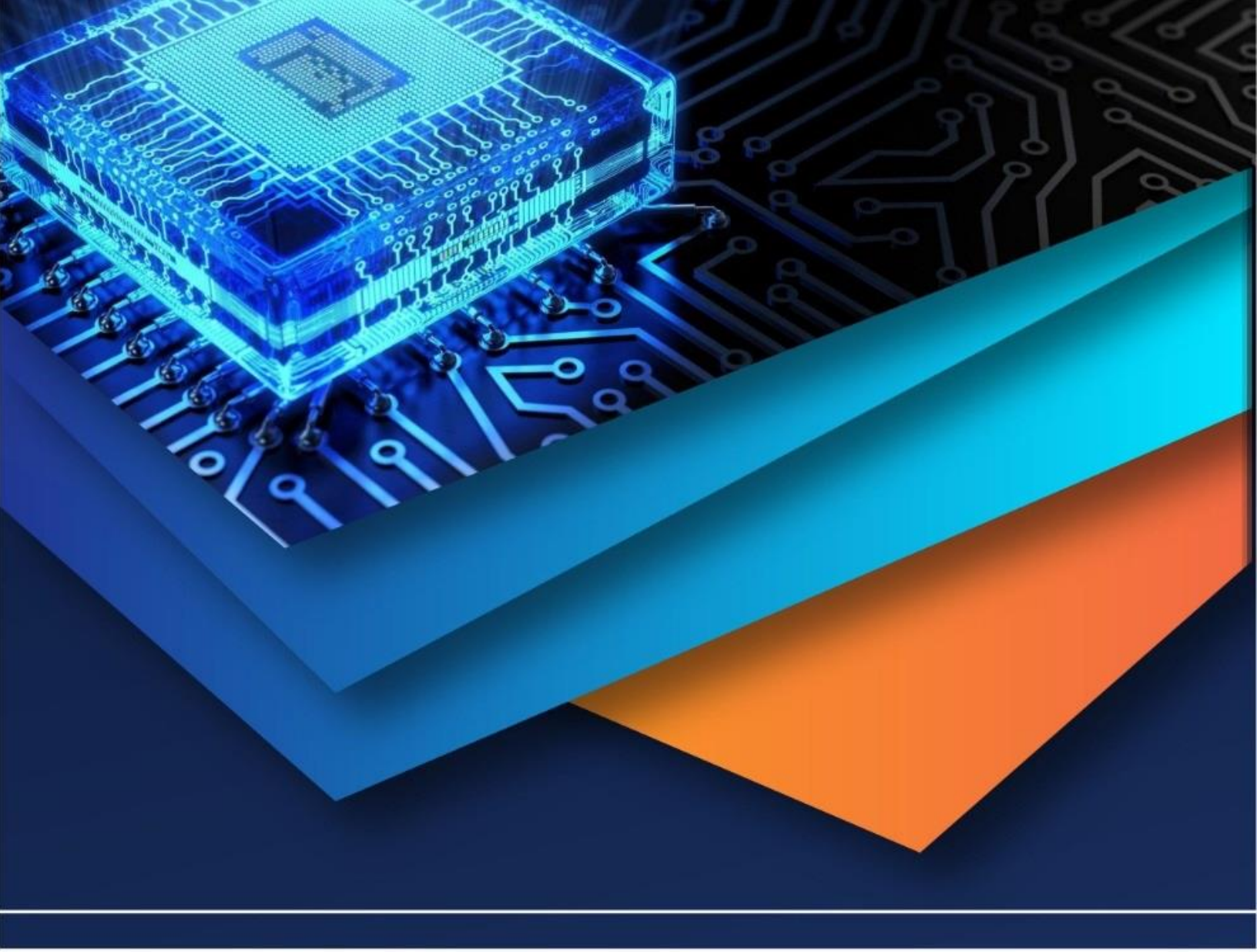
Sl No.	Classifier Algorithm	Author	Accuracy	Drawback
1.	K-Nearest Neighbor (KNN)	Barath Narayanan et al.[9]	96.60%	Computation cost is high.
2.	Support Vector Machine (SVM)	Udayakumar N et al.[3]	90.20%	Not provide probability estimates.
3.	Random Forest	Subash Poudyal et al.[2]	97.95%	Complex algorithm.
4.	Naive Bayes	N.Poonguzhali et al.[6]	97.11%	It is a bad estimator
5.	J48	Ala Bahrani et al.[1]	95.00%	Cost effective to classify new instance.

VIII. CONCLUSION

This survey is focused on the existing machine learning algorithms for Identifying Ransomware and Behavioral-Based Classifying. As we can see in table1 above the comparison of algorithms drawbacks are not be solved .Thus the proposed idea for this survey paper which can be further focused on getting more accuracy, training time, speed of classification with true positive values.

REFERENCES

[1] Ala Bahrani, Amir Jalaly Bidgly, “Ransomware detection using process mining and classification algorithms” IEEE, 2019.
[2] Subash Poudyal, Kul Subedi, Dipankar Dasgupta, “A Framework for Analyzing Ransomware using Machine Learning” ISBN:978-1-5386-9276-9, 2018 IEEE.
[3] Udayakumar N, Vatsal J.Saglani, Aayush V.Gupta, Subbulakshmi T, “Malware Classification Using Machine Learning Algorithms” IEEE, 2018 ISBN:978-1-5386-3570-4.
[4] Anam Fatima, Ritesh Maurya Kishore Dutta, Radim Burget and Jan Masek,”Android Malware Detection Using Genetic Algorithm based Optimized Feature Selection and Machine Learning” ISBN:978-1-7281-1864-2, 2019 IEEE.
[5] Ravi Kiran Varma P, Kotari Prudvi Raj and K.V Subba Raju, “Android Mobile Security by Detecting and Classification of Malware Based on Permissions using Machine Learning Algorithms” 2017, IEEE International Conference.
[6] N.Poonguzhali, T.Rajakamalam, S.Uma and R.Manju, “Identification of Malware using CNN and Bio-inspired Technique” 2019, IEEE.
[7] George Cabau, Magda Buhu and Ciprian Oprisa,”Malware Classification Based om Dynamic Behavior” DOI:10.1109/SYNASC.2016.51, IEEE.
[8] Phai Vu Dinh, Nathan Shone, Phan Huy Dung, Qi Shi, Nguyen Viet Hung and Tran Nguyen Ngoc, “Behaviour-aware Malware Classification: Dynamic Feature Selection” ISBN: 978-1-7281-3003-3, 2019 IEEE.
[9] Barath Narayanan, Ouboti Djaney-Boundjou and Temesguen M.Kebede, “Performance Analysis of Machine Learning and Pattern Recognition Algorithms for Malware Classification” 2016, IEEE, ISBN:978-1-5090-3441-3.
[10] Xin Zhou, Jianmin Pang and Guanghui Liang, “Image Classification for Malware Detection using Extremely Randomized Trees” 2017, IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY