TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 3034 - 3041

A Heuristic Research on Detecting SuspiciousMalware Pattern in Mobile Environment

G Maria Jones, L Ancy Geoferlaand, S Godfrey Winster

Article Info
Volume 82
Page Number: 3034 - 3041
Publication Issue:
January-February 2020

Article History

Article Received: *14 March 2019*
Revised: *27 May 2019*
Accepted: *16 October 2019*
Publication: *19 January 2020*

Abstract

Abstract—with the increasing trend of network technology in mobile platform is becoming more easily accessible and available device to people which an effect of spending valuable amount of time in online social networks which leads to increase of crime. However, due to the increasing technology advanced and popularity of mobile devices, cyber criminals are targeting on mobile device platforms for potential information from victims smart phones. There are billions of malware attacks taking place in every environment (Mobile, IoT, Wireless sensor Network, Cloud). The victims are tempted to use more internets where the device can be compromised by phishing websites and also by malware propagation. In this work, we compared the performance measures by using four machine learning algorithms and one neural network algorithm with aim of identifying mobile malware. With the goal of achieving detection of malware capability, we evaluated the accuracy of the system in terms of Accuracy, Precision, Recall and F1-Score with algorithms. The result shows that neural network algorithm achieved more malware detection accuracy compared with others

Keywords: *Malware, Phishing, Machine Learning, Mobile malware.*

I. INTRODUCTION

There has been increasing rise of technology in smartphones for personal and business use. These devices have wide variety of social applications to gain the private/ sensitive information by compromising the entire device. Social networking application makes everyone to share the personal information which attempts to compromise the data integrity. Around 1.48 billion smartphone has been shipped during 2018. These smartphones acquired with major risk as because devices are capable of accessing contacts, mails. Messages, Google map for accessing location, documents, phone calls, net banking like digital wallet. All the information is stored in volatile and non-volatile memory of electronic devices. Instant messaging applications like messenger, vibe, line, whatsapp, instagram act as a communication medium which makes everyone to share, exchange the information.

This leads to mobile forensics artifacts of former or deleted messages provided with all necessary information for investigation. While existence of all these personal information is never in digital wallet, sometimes it makes compromise the date through malware entry.

There is a risk mechanism associated in android environment which warn us before installing every application about the permission. Various play stores from android, iOS, windows; blackberry has many applications which open the door for malware attack. Malicious software can also damage and manipulate the normal behaviours. Ransomware is one type of malicious software which has a major threat in recent times. Among all existing platforms for malware, windows preferred as a target one for attackers, since mobile malware is an ever growing threat representation. This type of malware steals sensitive personal information, important documents and make demand for ransom (money for release the files). It not only infects our computer system but also mobile phones through downloading unknown files, documents and software's etc. According to 2016 report malware type PhishMe has 13.93% of phishing emails. The well-known data breach of RSA network occurred in 2011 by sending the

January - February 2020
ISSN: 0193 - 4120 Page No. 3034 - 3041

malicious excel file through email to company people. When they opened the excel file, the malware propagated through network and exploited the vulnerability to gain controls.

Even after the installation of application from Google play, Google Play Protect automatically scans the applications in order to check whether the applications remain safe or not. It has been reported that 50 billion applications are scanned daily by Google protect. In china, more than one million smart phones are infected by malware during 2011. The Machine Learning (ML) techniques is having more efficient for detecting and analyzing the malware and benign. There are three major classification of ML. They are: Supervised learning where the labeled output will be present, unsupervised learning, there is no labeled output and Reinforcement learning where it will act based on the environment. In this paper, supervised, unsupervised learning and back propagation which belongs to neural network are used. A major aim of mobile malware analysis is to capture and detect the malware and also with additional properties to be included for improving and taking preventive security measures as much as possible by using technology. Machine learning techniques are commonly used in every application to reduce the human action. Many novel researchers work this domain with a variety of applications, different objectives and with higher accuracy rate.

The rest of this paper is organized as following manner. Section II presents the literature work of malware analysis with detailed explanation. Section III describes the machine learning algorithms used in this paper. Section IV presents the experimental analysis. Section V describes the evaluation Metrices and finally section VI gives the conclusions and future work.

II. RELATED WORK

Avaid et.al presented a new technique called Structural feature extraction methodology (SFEM) with machine learning classifiers for detecting malicious documents which contains 830 malicious and benign files of 16,180. The random forest algorithm achieved highest accurate rate of 99%. Yao-Saint and Hung-Min presented a methodology which can detect Android malware using convolutional neural network by visualizing the code's importance value on an image with accuracy of 92%. Zahoor-Ur et.al presented a hybrid approach method for detecting malware in android Apps with various machine learning classifier algorithms for identifying it as legitimate or malicious and achieved highest accuracy of 85.5% with Support Vector Machine with standard deviation of 4.37. Sen Chen proposed KUAFUDET approach which learns malware in mobile by adversarial detection and showed that it can reduce false negative rates and it can increase the accuracy with 15%. Daniele presented a literature survey on existing on malware in context of windows platforms, what algorithm can be used and also with number of challenges and problem in this field. Shanshan et.al performed a framework for detecting android malware in server side which includes network traffic analysis with machine learning algorithm C4.5 which achieved detection rate of 97.89% and also revealed the mobile malware behavioural characteristics. Vasileios et.al contributed open source tool which collects the malware behavioural pattern from android mobile and evaluated the detection model with favourable results by using various machine learning algorithms. Zhenlong et.al implemented automatically detecting an online deep-learning-based Android malware detection engine (DroidDetector) in android aps and achieved 96.76% detection accuracy. Shanshan et.al introduced a technique for detecting malware in mobile phones using network traffic flows using NLP string analysis with detection rate for malicious flows reached 99.15% whereas the misjudgment rate for benign traffic is only 0.45%. Saba et.al proposed SAMADroid for malware detection model in android operating systems with novel 3-level hybrid by combining the advantages of both Static and Dynamic Analysis; Local and Remote Host; and Machine Learning Intelligence and also showed the result with high accuracy detection malware model

January - February 2020
ISSN: 0193 - 4120 Page No. 3034 - 3041

by ensuring the efficiency in terms of power and storage consumption. Jin et.al designed Significant Permission IDentification (SigPID) for allowing limited number of permission for detecting and analyzing mobile malware through pruning technique and supervised algorithm to detect the malware with the result of 93% of malware and 91% of new malware. George et.al presented a practical approach to increase accuracy in intrusion detection techniques using deep learning and mathematical models for robotic vehicles. Lei et.al proposed model based on logistic regression for android malware detection and explored various issues in feature granularity, representation, feature selection, and regularization. Elana et.al introduced multi-objective optimization which is capable of detecting android malware to discover dependencies, accuracy, time and power consumption which increase the efficiency of the system. Guanjun et.al evaluated real-time Twitter spam detection in terms of scalability, performance, capability by using machine learning and also calculated accuracy. Shweta et.al presented a detection tool called SWORD based on injection evasion technique which applied on system-calls and obtained the accuracy of 94.2%. Li et.al presented a classifier approach of parallel passive aggressive for malware detection in android for achieving higher accuracy and also concluded with future work is based on in run-time analysis. Christoforos et.al investigated and evaluated the recover authentication of mobile applications from the volatile memory of Android mobile devices by using open-source and free forensic tools. Giang et.al proposed a machine learning technique to detect suspicious behaviours activities based on logs from mobile devices by using Pre-processing, sliding windows, feature selection and lemmatization which is capable of in applying to the large-scale mobile environment. Ram proposed a technique to identify the malware in cloud environment by using weighted K means clustering algorithm with auto associate neural networks. Diana et.al developed an efficient framework based on deep belief network for android malware with high level static and dynamic analysis with accuracy of 99.1%. Nickson et.al proposed a framework for cyber forensics by using deep learning which solves problems and also discussed the challenges of evidences which should be considered during legal proceedings. Amato et.al proposed system architecture and semantic methods for analyzing and retrieving the digital evidences.

III. ALGORITHM FOR MALWARE DETECTION

Machine learning techniques have the capability of training and testing the data by extracting information from raw data. Five types of machine learning algorithms have been developed for detecting and targeting malware application scenarios from real time. In this paper, the authors has used four supervised and unsupervised learning algorithms and one neural network algorithm for detecting mobile malware from mobile device environment. The selected algorithms are categorized into three groups as classification, clustering and neural network which shown in table 1. These algorithms are used in many fields for variety of purposes. The authors would like to look over the performance of these algorithms with respect to malware datasets. The reasons for choosing these algorithms are also follows:

Categories	Algorithm
Classification	Logistic Regression
	Naïve Bayes
	Fuzzy Knn
Clustering	K-Means Clustering
Neural Networks	Back Propagation

Table.1 Algorithms used to perform

- ✓ Logistic Regression is a classification algorithm used as predictive analysis which determines the presence of malware.
- ✓ Naïve Bayes technique is used as a classifier algorithm which classify spam messages, mails; malware filter; text analysis and also in medical diagnosis.
- ✓ Fuzzy Knn algorithm is mainly used as a classifier in supervised learning.

January - February 2020
ISSN: 0193 - 4120 Page No. 3034 - 3041

- ✓ K Means clustering algorithm is unsupervised learning used to solve clustering problems.
- ✓ The Machine learning, neural network and deep learning is the popular algorithm for analyzing object detection, automatic handwriting generation, computer vision, and speech recognition.

IV. EXPERIMENTAL ANALYSIS

Research papers based on malware analysis for detecting malicious worms, virus, Trojans etc., from android, and iOS mobile phones, malware in network are mentioned in the related work session. The mobile devices artifacts can be either compromised by malware propagation or serve as a condition of crime attacks which leads to compromise entire mobile network. The current paper discuss about the malware and performance measures by using machine learning techniques. The author selected mobile malware dataset for current examination. The pre-processing, training and testing, feature selection was performed with the state and availability of details concerning the malware dataset associated with the mobile phones. However, the proposed methodology has certain limitation and implemented with limited number of algorithms. In order to proceed for detection of mobile malicious state, each mobile phone's modus operandi has to be found out. There is multiple infection factors are involved for delivering the malicious code to smartphones. It includes sms, mms, internet, usb and wireless (Bluetooth).

Generally, the dataset has been collected from various repositories which include: UCI repository, Kaggle, Github, Gov Data etc. Some users from worldwide voluntarily publish their data by installing application agent. Windows 7 operating system was used and Jupiter notebook from anaconda navigator is used for implement the machine learning algorithm with necessary and suitable libraries installed in it. The packages includes matplot library for plotting the points, numpy is a numerical python used for mathematical calculations, pandas, pickle and sklearn are used to perform. All dataset documents can be either stored it as a Comma Separated Value (.csv), Excel, Json, etc., file for importing into respective environments.

There are about 100000 data regarding malicious malware dataset are collected. During implementation, it's not necessary to have all data, so noise data can be filtered and redundancies can be removed which can be done by pre-processing technique. During training stage, Machine Learning and Neural network techniques are applied to train the dataset. The aim of our paper is to generate the best data model with higher accuracy to distinguish the conventional data from suspicious data in the surrounding of malware. In pre-processing, the certain amount of ratio has been taken for training and testing the data set. The data types used are malware with their respective attributes. Since the dataset contains large number of samples, the 10 fold cross-validation is used during the implementation. With the same hardware and software configuration, the evaluation of performance measures of each algorithm under same set of malware dataset is taken place. The 2:3 ratios, which means 40% of samples, are used for training and remaining 60% of samples, are used for training of malware is used. Once the five types of machine learning algorithms are used, the system can capable of separating the samples of safe and at risk state of malware as shown in fig.1. The safe state indicates that there is no malware contain in mobile environment, whereas at risk state, there is a malware propagation which can able to compromise the entire environment through one by one stage propagation.

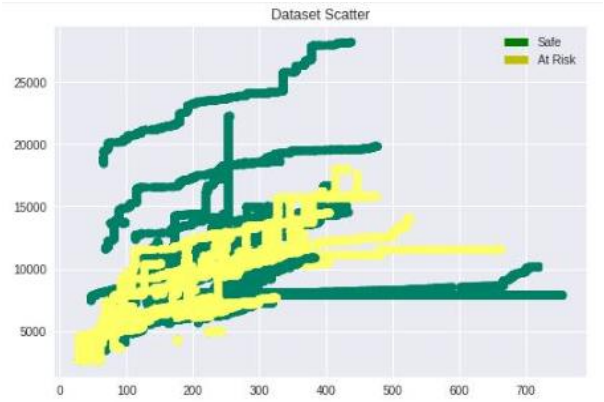

Fig.1 Malware at safe and risk state

January - February 2020
ISSN: 0193 - 4120 Page No. 3034 - 3041

After the evaluation of all five algorithms, the malware state indicates that the samples are distinguished as equal amount of malware presence and benign (good nature) are at equal level as shown in fig. 2

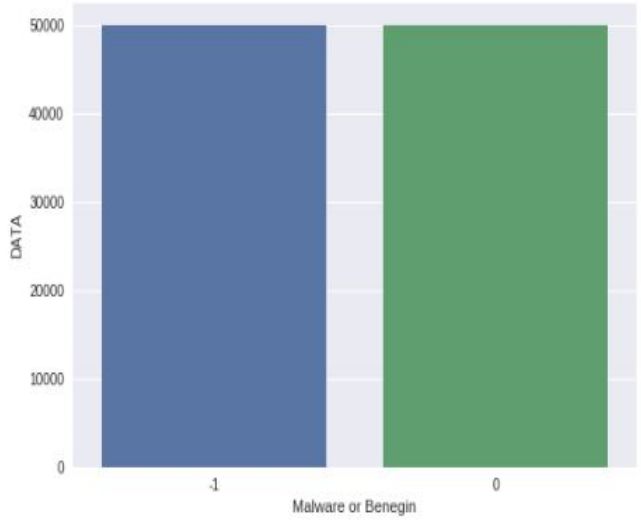

Fig. 2 Malware and Benign

V. EVALUATION METRICES

In this section, the performance measures are calculated to find the accuracy of selected 6 algorithms to check the performance of the malware detection as mentioned in table 1. We compare the performance for malware samples taken by logistic regression, naïve bayes, K – means clustering, Knn and back propagation algorithm.

A. PERFORMANCE MEASURES

The true positive rate (TPR), false positive rate (FPR), Precision, Recall and F-measure of the system is used for performance measure which is meant for calculating the accuracy of malware from mobile environment data set. The following (1), (2), (3) and (4) equations are used in the system as below:

$$\text{Accuracy (A)} = TP / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision (P)} = TP / (TP + FP) \quad (2)$$

$$\text{Recall (R)} = TP / (TP + FN) \quad (3)$$

$$\text{F1 Score} = 2 * PR / (P + R) \quad (4)$$

The performances measures are used to calculated the accuracy and verify the efficiency of the algorithm by using True Positive (TP), False Negative (FN), True Negative (TN) and False Positive (FP). The percentage of accuracy is the ratio of correctly identified true positive rate in malware over the total amount of malware sample as shown in the equation (1). The precision rate is calculated by the ratio of correctly identified malware by total sum of True Positive and False Positive as shown in the equation (2). The recall sample is calculated as same like precision with the difference of ratio of TP over the sum of True Positive and False Negative as given in equation (3) and finally F Score is another measurement to calculate the prediction accuracy by using P and R. Fig. 2 describes the mobile malware propagation in the detection of process performed on the reserved virtual memory by its time duration. Where some process executes in reserved state for certain period of time. The malware and benign are also classified based on the time.

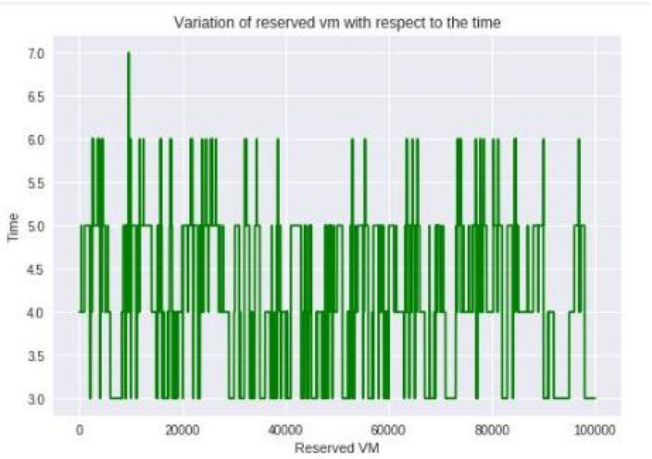

FIG. 3 DETECTION PROCESS

B. VALIDATION TECHNIQUE

In the experimental analysis of mobile malware, five different types of machine learning algorithms are used, namely Logistic Regression, Naïve Bayes, Fuzzy K nearest neighboring, K – means Clustering and back propagation for single malware set. The 10-fold cross validation technique is used. Table 2 shows the best performance measures with malware and benign with 1 and 0 respectively. For each algorithm implementation, cross validation has done in order to gain the accuracy. Neural network shows the highest accuracy among the entire five algorithms. The diagrammatic representation of all five algorithms accuracy is shown in the fig. 5. The risk indicator of mobile malware was presented in fig. 4 where the malware in mobile environment can damage the entire environment which is in risk state.

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 3034 - 3041

S.No	Algorithm	State	Accuracy	Precision	Recall	F1-Score	Support
1	Logistic Regression	1 (Malware)	81.86%	0.84	0.79	0.81	50,000
		0 (Non-Malware)		0.80	0.82	0.82	50,000
2	Naïve Bayes	1 (Malware)	61.46%	0.64	0.52	0.57	50,000
		0 (Non-Malware)		0.60	0.71	0.65	50,000
3	Fuzzy Kn	1 (Malware)	87.33%	1.00	1.00	1.00	50,000
		0 (Non-Malware)		1.00	1.00	1.00	50,000
4	K – Means Clustering	1 (Malware)	43.24 %	0.00	0.00	0.00	50,000
		0 (Non-Malware)		0.32	0.13	0.18	50,000
5	Neural Network	1 (Malware)	98.2%	-	-	-	50,000
		0 (Non-Malware)					50,000

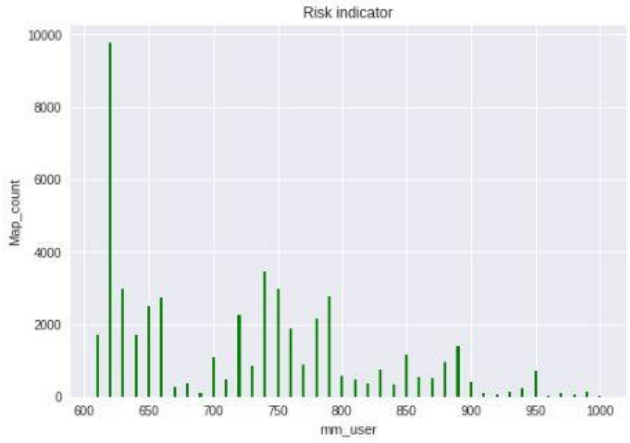

FIG.4 RISK INDICATOR

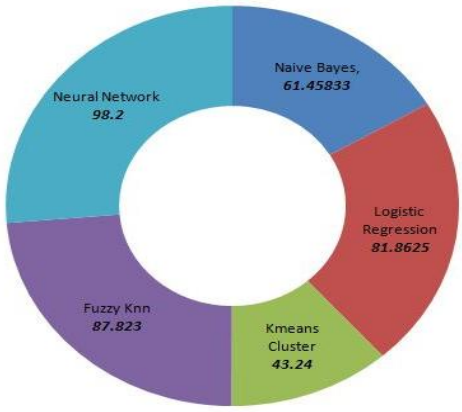

FIG. 5 ACCURACY FOR USED ALGORITHMS

VI. CONCLUSION AND FUTURE WORK

Malware detection is an important task in every mobile environment platform as it is fast growing market of today's world. This paper propose a malware detection model by using machine learning approach with a learning model based on regularized logistic regression, Naïve Bayes, KNN, K-Means Clustering, Back Propagation (Deep Learning) which achieves high accuracy detection results. The evaluation compares several machine learning approaches. The obtained results confirm that neural Networks algorithm is a strong competitor to these algorithms in malware detection according to the presented dataset. In our model, we acquired 98.2 % accuracy using neural network of back propagation whereas other algorithm gives lower accuracy when compared to other used algorithms.

VII.REFERENCES

[1] PierangAviad Cohen, NirNissim, LiorRokac, Yuval Elovic, "SFEM: Structural feature extraction methodology for the detection of malicious office documents using machine learning methods" Expert Systems with Applications, Vol. 63, 2016, PP. 324 – 343.

[2] Yao-Saint Yen and Hung-Min Sun, "An Android mutation malware detection based on deep learning usingvisualization of importance from codes", Microelectronics Reliability, Vol. 93, 2019, PP. 109 – 114

[3] Zahoor-Ur Rehman, Sidra Nasim Khan, Khan Muhammad, Jong Weon Lee, ZhihanLv, Sung

January - February 2020
ISSN: 0193 - 4120 Page No. 3034 - 3041

WookBaik, Peer Azmat Shah, Khalid Awan, IrfanMehmood, "Machine learning-assisted signature and heuristic-based detection of malwares in Android devices" Computers and Electrical Engineering, Vol. 69, 2018, PP. 828–841.

[4] Sen Chen, MinhuiXue, Lingling Fan, ShuangHao, LihuaXu, Haojin Zhu, Bo Li, "Automated poisoning attacks and defences in malware detection systems: An adversarial machine learning approach", computer & security, vol. 73, 2018, PP. 326 – 344.

[5] Daniele Ucci, Leonardo Aniello, and Roberto Baldoni, "Survey of machine learning techniques for malware analysis" computer & Security, Vol 81, 2019, PP. 123 – 147.

[6] Shanshan Wang, Zhenxiang Chen, Qiben Yan, Bo Yang, LizhiPeng, ZhongtianJia, "A mobile malware detection method using behavior features in network traffic", Journal of Network and Computer Applications, 2018.

[7] Vasileios Kouliaridis1, Konstantia Barmpatsalou2, Georgios Kambourakis1, and Guojun Wang, "Mal-warehouse: A data collection-as-a-service of mobile malware behaviowsral patterns", 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, ScalableComputing& Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovations, PP. 1503- 1508.

[8] Zhenlong Yuan, Yongqiang Lu and YiboXue, "Droiddetector: android malware characterization and detection using deep learning", Tsinghua Science and Technology, Vol. 21, Issue: 1, 2016. PP. 113 – 123.

[9] Shanshan Wang, Qiben Yan, Zhenxiang Chen , Bo Yang, Chuan Zhao, And Mauro Conti, "Detecting Android Malware Leveraging Text Semantics Of Network Flows", IEEE Transactions On Information Forensics And Security, Vol. 13, No. 5, May 2018, PP. 1096 – 1108.

[10] Saba Arshad, Munam A. Shah, Abdul Wahid, AmjadMehmood, Houbing Song and Hongnian Yu, "SAMADroid: A Novel 3-Level Hybrid Malware Detection Model for Android Operating System" IEEE Access, Feb 2018, PP. 4321 – 4339.

[11] Jin Li, Lichao Sun, Qiben Yan , Zhiqiang Li, WitawasSrisa-an , and Heng Ye, "Significant Permission Identification for Machine-Learning-Based Android Malware Detection",IEEE Transactions On Industrial Informatics, Vol. 14, No. 7, July 2018, PP. 3216 – 3225.

[12] George Loukas, Tuan Vuong, Ryan Heartfield , Georgia Sakellari, Yongpil Yoon, And Diane Gan, "Cloud-Based Cyber-Physical Intrusion Detection for Vehicles Using Deep Learning" IEEE Access, Vol. 6,2018, 3491 – 3506.

[13] Lei Cen, Christoher S. Gates, Luo Si, and Ninghui Li, "A Probabilistic Discriminative Model for Android Malware Detection with Decompiled Source Code" IEEE Transactions On Dependable And Secure Computing, VOL. 12, NO. 4, July/August 2015, PP. 400 – 412

[14] Elena Milosevic, MiroslawMalek and Alberto, "Time, Accuracy and power consumption tradoff in mobile malware detection system", Computers & systems, 2019.

[15] Guanjun Lin, Nan Sun1, Surya Nepal, Jun Zhang, Yang Xiang and Houcine Hassan, "Statistical Twitter Spam Detection Demystified: Performance, Stability and Scalability", IEEE Access, Special Section on Big Data Analytics in Internet of Things and Cyber-Physical Systems, Vol. 5, 2017, PP. 11142 – 11154.

[16] ShwetaBhandari, RekhaPanihar, Smita Naval, Vijay Laxmi, AkkaZemmari, Manoj Singh Gaur, "SWORD: Semantic aWareandrOidmalwaRe Detector" , Journal of Information Security and Applications, Vol. 42, 2018, PP. 46–56.

[17] Android e Google Play Protect, 2017. Google. https://www.android.com/playprotect/. (Accessed 21 August 2017).

[18] AV-TEST. Security Report 2016/17. 2017.https://www.av-test.org/fleadmin/pdf/security_report/ AV-TEST_Security_Report_2016-2017.pdf

[19] Li Zhang, Vrizlynn L.L. Thing and Yao Cheng, " scalable and extensible framework for android malware detection and family attribution", Computers & Security, Vol. 80, 2018, PP. 120133.

[20] ChristoforosNtantogian, DimitrisApostolopoulos, GiannisMarinakis, Christos Xenakis, "Evaluating the privacy of Android mobile applications under

January - February 2020
ISSN: 0193 - 4120 Page No. 3034 - 3041

forensic analysis", computers & security, Vol.42, 2014, PP. 66-76.

[21] Giang Nguyen, Binh Minh Nguyen, Dang Tran and LadislavHluchy, "A heuristics approach to mine behavioural data logs in mobile",Data and Knowledge Engineering, Vol. 115,2018, PP. 129 – 151.

[22] Ram Mahesh Yadav, "Effective Analysis Of Malware Detection In Cloud Computing" , Computers & Security, dec 2018.

[23] Aviad Cohen, NirNissim,YuvalElovici, " Novel set of general descriptive features for enhanced detection of malicious emails using machine learning methods", Expert systems with applications, May 2018.

[24] http://hellanicus.lib.aegean.gr/handle/11610/18300

[25] Dina Saif, S.M. El-Gokhy and E. Sallam, "Deep Belief Networks-based framework for malware detection in Android systems", Alexandria Engineering Journal, vol. 57,2018, 2018, PP. 4049–4057.

[26] https://www.firstpost.com/tech/news-analysis/global-smartphone-shipments-to-cross-1-49-billion-units-in-2018-4527291.html

[27] NicksonM.Karie, Victor R. Kebabde and H.S. Venter, "Diverging deep learning cognitive computing techniques into cyber forensics" Forensics Science International: Synergy. Vol.1, 2019, PP. 61 - 67

[28] F. Amato, G. Cozzolino, V.Moscato and F. Moscato, "Analyse digital forensics evidences through a semantic-based methodology and NLP techniques", Future Generation Computer Systems, Vol.98, 2019,

AUTHORS PROFILE

G Maria Jones,Research Scholar in computer science and engineering department at Saveetha engineering college, Chennai, India. She carries out her research activities since 2018. Her research activities involve digital forensics, Machine Learning and malware propagation.

L AncyGeoferla, Assistant Professor (Gr. II) in the department of information technology at RMK Engineering College, Chennai, India. She has 11 years of teaching experience and also presented various papers in national and international conferences.

S Godfrey Winster, Professor in Computer Science and Engineering department at Saveetha Engineering College, Chennai, India. He received his Ph.D in Information and Communication Engineering from Anna University. He has 16 years of experience in techining and published several papers in reputed journals. His area of specialization includes data mining, big data and cloud computing.