International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Detection of Malware attacks in smart phones using Machine Learning

V. R. Niveditha, T. V. Ananthan

Abstract: ***In recent years, security has become progressively vital in mobile devices. The biggest security problems in android devices are malware attack which has been exposed to different threats. The volume of new applications by the production of mobile devices and their related app-stores is too big to manually examine the each and every application for malicious behavior. Installing applications which may leads to security vulnerabilities on the smart phones request access to sensitive information. There are various malwares can attack android device namely virus, worms, Botnet, Trojans, Backdoor and Root kits due to these attacks the users is compromised by privacy. Root kits and viruses in mobile phone and IoT devices improve along with smart device versions are very difficult to detect or to the least costly. There are 3 places where the trace of these root kits / virus is visible namely CPU, Baseband and Memory. In the new approach we will use machine learning to detect "anomaly" usage pattern and a remote (master server) will analyze and verify the presence of such threats. This research work aims to develop a pipeline to investigate if any application present in a smart device is a malware or not. This pipeline uses HMM algorithm to read anomaly in application behavior, deep learning with Deep Belief Networks (DBN) to classify application events, and bootstrapping algorithm using random forest to categorize the application itself after malware or benign.***

Keywords: Malware attacks, mobile phone, android device, Hidden Markov Model (HMM), Deep Belief Network (DBN), Random Forest (RF), Machine Learning (ML)

I. INTRODUCTION

In 2016, the Smartphone users will exceed by 2 billion and also in US population 65% will own a Smartphone [1]. In worldwide the most popular mobile device OS is android [2]. Also around the worldwide, more than 190 countries android is available and download more than 1.5 billion apps by the users from the Google play store in every month [3]. The approval of android in both official and unofficial app markets has brought the attention of malware apps. Many of the anti-virus software are to detect malware by using a signature-based method. Though, detecting the known malware and unknown malware which are highly active is not detected and the systems from the malware attacks can suffer. The increasing number of challenges we face today has lead for the development of complex malware detection approach for Android platform.

Fig.1. Device Infection Procedure

Figure 1 illustrated the flow from malware sharing to device infection. When the malicious form of the transportation app is installed, it authorizations whether the fake plugin is previously installed and, if not, downloads from the server and installs it. After that, it downloads and performs a further instinctive trojan binary which is alike to the trojan which is released by the fake plugin. After everything is done, it connects with the C2 servers and handles received commands. This research work to detect android malware which has been dynamically looking for various ways. Many researchers have been used classification of ML algorithms to classify the apps in order to identify if an android app is attacked by malware or not. The arbitrary sets are classified into various categories by using ML. In the malware case, the representations of vector features are malicious and benign files that can be extracted in various ways and most commonly from the application of binary code. Malware detection are commonly used by ML algorithms namely Naïve Bayes (NB), K-Nearest Neighbor (KNN), Decision Tree (DT) and Support Vector Machine (SVM) [4,5]. The traditional ML algorithms from the malware can potentially learn the behavior features. Unfortunately the performance of most ML algorithms depends on the extracted features accuracy.

Revised Manuscript Received on November 30, 2019.
* Correspondence Author
V. R. Niveditha*, Department of Computer Science and Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, India. Email: vrniveditha@gmail.com
T. V. Ananthan, Department of Computer Science and Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, India. Email: tvananthan@drmgrdu.ac.in

In addition for improving the performance of malware detection, it is often challenging to extract meaningful behavior features. Hence, malware detection using traditional ML algorithms are still somewhat unsatisfying [6]. Though the current anti-virus systems are effective against predetermined malware and improved for mobile devices, they cannot show the performance which has unknown signature against new malware [7].The analyzes of current attacks must be established for predicting the occurrence over forthcoming attacks with learning based methods and systems [8].This research work aims to develop a pipeline to investigate if an app in a smart device is a malware or not. In the first phase, introduces a method as Hidden Markov Model (HMM) that performs behavior patterns of apps in smart phones to detect anomaly. In second phase, introduces Deep Belief Network (DBN) on emerging an efficient framework for detecting malware and classify the event as the harmful or benign event using deep learning. This might permit the DBN for determining by either corresponding of input features to benign or malware events which act as binary classifier. In third phase, introduces classification of Random Forest to detect by examine the malware application behavior data if an android device has been compromised.

II. PROBLEM DEFINITION

Digitization of the world is growing at rapid speed where the smart homes devices mobile users are increasing day by day as the device can provide convenience, security, and energy efficiency to users. With the increasing need and use of mobile and smart devices in digital world or telecommunications, there is a raise in malware attacks in mobile devices and IoT (Internet of Things). There have been numerous topical instances of smart devices being hacked for violation of privacy and also misused so as to perform DDoS (Distributed Denial of Service) attacks. Presently, many researches are focused mainly on identification of anomalous activities that can arise in mobile's app and smart home environment. They are focused on vulnerable points like Application Program Interface (API) and user permission grand like authentication and authorization which run sandbox not in practical environment. In response to this problem, our study proposes to investigate several options for making smart phones and smart devices are more secure. Our work focuses on anomaly detection, analyzing the time series of all events from all application in the smart device and detection of malwares in the live practical environment.

III. LITERATURE REVIEW

Madhawa et. Al [9] provides the several network security enhancements based on Software Defined Network (SDN) in IIOT thereby reducing the property of consistency attacks. The anomaly detection is the model of intrusion detection mode with negligible false positives in the industrial process. Lu and Hou [10] described the model of android malware detection in order to detect the android malware effectively. This model is based on the declared permission. It can consist of two layers. The detection of enhanced RF algorithm is used to analyze the first layer whereas in second layer for analyzing the fuzzy sets produced by the detection of sensitive permission rules matching. Geneiatakis et al. [11] provide an efficient technique of permission certification. This paper, researcher had been combined by both static analysis and runtime information and then identified if they are follow the least restricted or over restricted to the profit mobile application. Min et.al [12] provides by original RF judgment while the decision tree count is 100 in RF is said to be better classification. Lashkariet.al [13] described to detect not only the masked apps, or malicious but also to identify them as particular malware or general malware on mobile device. Arora et.al [14] proposed with minimum number of features to arrange the network traffic features and analyzed for better processing time and detection accuracy. Chen et. Al [15] described on behavior destructive features for malware variant detection method by combining a static and dynamic analysis. Li et. Al [16] proposed to detect the android malware by deep learning methods and to detect the families of malicious applications by automatic detection engine. Martinelli et. al [17] described as a new field of ML from benign application to detect malware automatically. The malicious application detection is applied using deep learning. Yoeet.al [18] proposed using Convolutional Neural Network (CNN) for an automated malware detection technique. Baskaran and Ralescu [19] focused on the android OS based on ML algorithms that summarize the development of malware detection technique. Chumachenko [20] presents malware classification and detection that suggested the methods for ML. Additionally, the study performed with ML approach can be useful as a base for further research in the field of malware analysis.

IV. RESEARCH METHODOLOGY

In this new approach we will use machine learning to detect "anomaly" usage pattern and a remote (master server) will analyze and verify the presence of such threats. This research work aims to develop a pipeline to investigate if an app in a smart device is a malware or not. This pipeline uses HMM algorithm to read anomaly in application behavior, deep learning (with in DBN / CNN) to classify application events, and basic classification algorithm (with in random forest/SVM) to categorize the application itself after malware or benign. This research work considers comparative study of user DBN and CNN for deep learning, and random forest and SVM for classification of app. The overall block diagram is as shown in Figure 2.

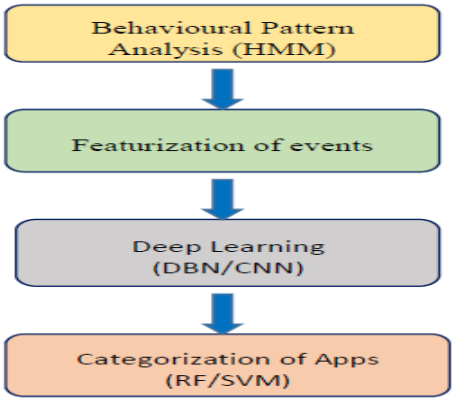

Fig. 2. Proposed Block diagram Detection of Malware attacks

Phase: 1
Apply HMM to process the events data.
Phase: 2
Detect anomaly from HMM output
Apply deep learning to classify the events

Phase: 3
Apply Random forest/SVM to classify the application as malware or benign from previous result.

A. In First Phase, apply Hidden Markov Models for Anomaly Detection to process the events data

In order to detect the behavior pattern changes present in the smartphone applications is said to be anomaly detection whereas it can be performed using statistical model construction which consists of metrics resulting from flagging and system operation as some instructive metrics observed that may have an essential deviation over statistical from the provided model. In this phase, anomaly detection will use machine learning to detect "anomaly" usage pattern and a remote (master server) to analyze and verify the presence of such threats. The process of discovering the malicious behavior which targeted network and its resources is said to be anomaly detection [21].However, this research focus on the performance of anomaly detection in various application field namely business news monitor, detection of hardware fault, monitor of network alarm and instruction detection. This research deal with anomaly detection issues includes wide range of time series data that may provide an important activities and entries number. Hence, the major focus of this process is to discover several fascinating and rare events with less delay and minimal amount of false alarms [22].In order to study and analyze the behavior patterns of apps in a smart phone and a smart TV using Hidden Markov model or one of its variants.

1. A Framework for Evolving an HMM

There are several parameters utilized in HMM namely
N = Number of states present in the model.
t = time for specific instant.
Si = {S_1,S_2,S_3…..S_n} whereas several individual state present in the model.
M= Number of individual symbol observation per state. This symbol represented the physical output to be modeled.
V = {V_1,V_2,…V_m} represent several individual symbol.
A = Probability distribution while transition state.
B = Probability distribution of the observational symbols
Π = Probability distribution of initial transition state.

However, there is different sequence of states represented as OO = OO1, OO2, OO3….OOT is said to be indirect observation of hidden state whereas 'T' is mentioned as total number of observation considered. Hence, the proposed methodology is considered depend on HMM by distance of behavioral consists of subsequent parameters. As an instant, this model with 'N' may be considered as hidden states whereas it is assumed as 5 and M is expressed as observation once the hidden state are accounted a SS1, SS2 and SS3 but the value is 2 for SS4 and SS5.

Once the step of parameter estimation is completed and followed by training HMM with forward procedure. The forward variable = B (OO1, OO2, OO3, OO4, OO5, qt = S_n| λ)

The 'B' represents partial observation sequence probability for OO1, OO2, OO3, OO4 and OO5 whereas the individual state is represented as S_n with a given time "t" for a model. However, the observation sequence "λ" for OO1, OO2, OO3, OO4 and OO5 have indicated the symbol of discrete observation of state S1, S2, S3, S4 and S5 correspondingly. Hence this case value of OO1, OO2 and OO3 has ranged from 1 to 6 and in the case of OO4 and OO5 the value has ranged as 1 or 2. Thus the involved steps in forward procedure are illustrated by equation (i),(ii) and (iii).

In order to initialize the value of forward variable is

$$\alpha t(n) = n^* b_n (OO1) \quad (1)$$

Where 1 <= n<= 5

Induction step in the Forward variable Process

$$\alpha_{(t+1)^p} = [\sum_{n=1}^{5} \alpha_t (n) * a_{np}]^* b_p (O_{t+1}) \quad (2)$$

Where 1<=t<=T-1 and 1<=p<=5.

Final step in the Forward variable Process

$$B (O \mid \lambda) = \sum_{n=1}^{5} \alpha_t (n) \quad (3)$$

Therefore, B (O | λ) is the sum of all the π t (n) values.

At present, the probability distribution of every state is manipulated and combined those probability distribution from every state for acquiring a mean probability distribution that can be correlated with every individual state based on anomalies behavior pattern of probability distribution are classified namely normal, medium and high as the anomaly type.

B. In Second Phase, to classify the events by applying Deep Learning

In this phase, apply deep learning to classify the event as the harmful or benign event. In ML, the most recent area of research is Deep learning. Geoffrey Hinton proposed in decision making to mimic the human brain. Deep Learning showed its supremacy during the last few years in many other fields such as image recognition, image processing and voice recognition. In this paper DBN is used for detecting the malware to classify the events. The power of DBN is implemented by layer-by-layer learning policy, which lies in their capability to renovate both the learning feature vectors and input vector. There are some certain layers have features that gets accomplished from the preceding layer which promises more sophisticated. DBN is otherwise known as reproductive graphical model. It consists of Restricted Boltzmann Machines (RBM) stacked number whereas the layer of RBM are connected together to the same layer without connecting the units is known as undirected graphical model. DBN has stacked RBM as unsupervised network that act as a sub network for hidden layer which is considered as a visible layer. After, the sub-networked hidden layer is treated as a visible layer for the subsequent layers and this process gets iterated. The every RBM present in a sub network is a contrastive divergence with layered and performs the layers as a feature detectors whereas this fine-tuned process gets applied after the process of pre-training for determining whether the input features relates to benign event or harmful by allow DBN using supervised learning to act as binary classifier.

Each RBN contains input layer that represents as visible layer (V) and hidden layer represents as (H) and associated with each other as a weight vector (W).
The following steps discussed in detailed given below:
1. For training the dataset, set the states of visible units.
2. Start with positive statistics by contrastive divergence(E_{ij}) = P$(H_j = 1|V)$. The hidden layer units estimated by individual activation probabilities which are indicated in

$$P(h_j = 1|V) = \sigma(b_j + \sum V_i W_{ij}) \quad (4)$$

Where "b" is represents the bias of the hidden layer

$$\sigma(x) = 1/(1+\exp(-x)) \quad (5)$$

σ(x) represents the function of logistic sigmoid.
3. For considering the negative phase from hidden to visible by calculating the statistics of negative for reconstructing the visible units by the same technique

$$(E_{ij}) = P(V_j = 1|h) \quad (6)$$

The probabilities of individual activation can be estimated for visible layer units by using the equation below

$$P(V_j = 1|h) = \sigma(a_j + \sum h_j W_{ij}) \quad (7)$$

Where"a" represents the bias of visible layer
Update the weights using equation

$$\Delta w_{ij} = \epsilon(< v_i h_j > data - (< v_i h_j > model)) \quad (8)$$

4. Repeat till the required threshold is attained for all the training samples.
5. After processing all the layers for attaining the finest classifier performance by fine tuning the obtained results.

This leads with android security to analyze either the input features corresponds to a harmful or benign event.

C. In Third Phase, apply Random forest to classify the application as malware

In this phase apply Random Forest algorithms to classify the application of the malware or benign from previous phase. The random forest based on decision trees that present's random attribute selection. The traditional decision trees in the attribute set choose best possible attribute of the recent node. The RF randomly chooses from the set of attributes by a division of the K attributes of the node. Then it chooses from the subset for the partition in an optimal attribute. RF is a classifier to determine by the number of individual categories of tree output which includes multiple decision trees. The proposal of this paper based on learner of RF by CART decision tree. For instance, the application of android is categorized into two namely malware and benign.
The following bootstrapped random forest training process is given below.

Step 1: The given representation of S represents the training of sample set, T represents the test of sample set, F indicates characteristics measurement, t represents the amount of CART decision tree, f indicates number of features per node and d defines the strength of every CART decision tree.

Step 2: For the i-th tree, S(i) represents the training set which is extracted from S that otherwise known as root node of sample and start from the root node for training.

Step 3: If it is reached by end condition and set it as the leaf node by the current node. The output predicted in the current node for the classification problem, sample set is the biggest number of c(j).
Step 4: Replicate the above process 2 to 3 till entire nodes have been trained or categorized as leaf nodes.
Step 5: Replicate the steps above process 2 to 4 until all the "t" trees have trained.

This leads to categorize the application itself after malware or benign from the previous result and communicate result back to mobile device.

V. RESULT AND DISCUSSION

We evaluated our framework in terms of detection accuracy, precision and recall of rewritten applications. 1,260 malware samples from the Mal-Genome project were used in the experiments. An experiment for detect the behavior pattern changes present in the smartphone applications is said to be anomaly detecting. Our results are promising to obtain an accuracy for events data is 0.93, precision for events data is 0.968 and recall for events data 0.95 to identify the malware family. In DBN has been developed. The pro-posed framework achieves 99.1% accuracy for the present dataset to classify the events from mal-ware and good ware applications. The overall accuracy of the model is 0.9526 within the bounds within the bounds of [0.9411, 0.9514]. The resulting accuracies have shown that Random Forest model achieved a 0.9526 classification accuracy for the given malware dataset are shown in table 1 and figure 3.

Table-I: Performance evaluation based on classify malware behavior, events and application

Algorithm	Accuracy	Precision	Recall
HMM	0.93	0.96	0.95
DBN	0.99	0.97	0.96
RF	0.95	0.99	0.95

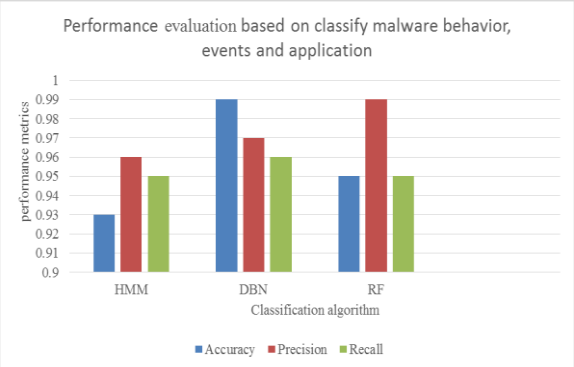

Fig. 3. Performance evaluation for accuracy, precision and recall

VI. CONCLUSION

Mobile devices including smart home accessories are ubiquitous and become common targets of malware attack. Several solutions are exits that tried to detect malware, but not successful yet. This research work proposes a 3 stage pipeline the combination of ML algorithm to detect malware form its behavioral pattern including requesting of permission. Although the methodology proposed here provides efficient HMMs model for behavior patterns of apps in a smart phone. After detect anomaly from HMM output apply an efficient computational framework for classify the event as the harmful or benign depend upon Deep Belief Network (DBN) has been developed and categorize the application itself after malware or benign using Random forest algorithm. This pipeline is used to detect a malware in a smart device with acceptable QoS.

REFERENCES

1. Emarketer. (2014) 2 billion consumers worldwide to get smart (phones) by 2016. [Online]. Available: http://www.emarketer.com/Article/2-Billion-Consumers-Worldwide-Smartphones-by-2016/1011694
2. M. Senthil Kumar et. al, Experimental investigations on mechanical and microstructural properties of Al_2O_3/SiC reinforced hybrid metal matrix composite, IOP Conference Series: Materials Science and Engineering, Volume 402, Number 1, PP 012123. (https://doi.org/10.1088/1757-899X/402/1/012123)
3. Android. (2015) Android, the world's most popular mobile platform. [Online]. Available: http://developer.android.com/about/index.html
4. M. Fan, J. Liu, X. Luo et al., "Android malware familial classification and representative sample selection via frequent subgraph analysis," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pp. 1890–1905, 2018.
5. Z. Lin, X. Fei, S. Yi, Y. Ma, C.-C. Xing, and J. Huang, "A secure encryption-based malware detection system," *KSII Transactions on Internet and Information Systems*, vol.12, no.4, pp. 1799–1818, 2018.
6. Fei Xiao, Zhaowen Lin ,Yi Sun and Yan Ma, "Malware Detection Based on Deep Learning of Behavior Graphs", *Mathematical Problems in Engineering*, Volume 2019, Article ID 8195395, 10 pages, https://doi.org/10.1155/2019/8195395.
7. Kumar, M. Senthil, et al. "Processing and characterization of AA2024/Al_2O_3/SiC reinforces hybrid composites using squeeze casting technique." Iran. J. Mater. Sci. Eng 16.2 (2019): 55-67.
8. Surendar Madhawaa, P. Balakrishnanb and Umamakeswari Arumugama, "Employing invariants for anomaly detection in software defined networking based industrial internet of things", *Journal of Intelligent & Fuzzy Systems* xx (20xx) x–xx DOI: 10.3233/JIFS-169670, 2018.
9. S.Yogeshwaran, R.Prabhu, Natrayan.L, Mechanical Properties of Leaf Ashes Reinforced Aluminum Alloy Metal Matrix Composites, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 13, 2015.
10. Tianliang Lu and Su Hou, "A Two-Layered Malware Detection Model Based on Permission for Android", 2018, *IEEE International Conference on Computer and Communication Engineering Technology (CCET)*.
11. L.Natrayan et al. Optimization of squeeze cast process parameters on mechanical properties of Al_2O_3/SiC reinforced hybrid metal matrix composites using taguchi technique. Mater. Res. Express; 5: 066516. (DOI: 10.1088/2053-1591/aac873,2018).
12. Liu Min, Lang Rongling, Cao Yongbin. "Number of trees in random forest", *[J]. Computer Engineering and Applications*. 2015, 51(5):126- 131.
13. Arash Habibi Lashkari, Andi Fitriah A. Kadir, Hugo Gonzalez, Kenneth Fon Mbahand Ali A. Ghorbani, "Towards a Network-Based Framework for Android Malware Detection and Characterization", 2017, *15th Annual Conference on Privacy, Security and Trust*.
14. L.Natrayan , P. Sakthi shunmuga sundaram, J. Elumalai. Analyzing the Uterine physiological With MMG Signals using SVM, International journal of Pharmaceutical research, 2019, 11(2); 165-170.
15. Natrayan, L., M. Senthil Kumar, and Mukesh Chaudhari. "Optimization of Squeeze Casting Process Parameters to Investigate the Mechanical Properties of AA6061/Al_2O_3/SiC Hybrid Metal Matrix Composites by Taguchi and Anova Approach." Advanced Engineering Optimization Through Intelligent Techniques. Springer, Singapore, 2020. 393-406.
16. Sundaram, P. Sakthi Shunmuga, N. Hari Basker, and L. Natrayan. "Smart Clothes with Bio-Sensors for ECG Monitoring." International Journal of Innovative Technology and Exploring Engineering 8.4 (2019): 298-301.
17. Martinelli F, Mercaldo F, Saracino A. BRIDEMAID: An Hybrid Tool for Accurate Detection of Android Malware[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ACM, 2017: 899-901.
18. L. Natrayan and M. Senthil Kumar. Study on Squeeze Casting of Aluminum Matrix Composites-A Review. Advanced Manufacturing and Materials Science. Springer, Cham, 2018. 75-83. (https://doi.org/10.1007/978-3-319-76276-0_8.)
19. Balaji Baskaran and Anca Ralescu, "A Study of Android Malware Detection Techniques and Machine Learning", MAICS 2016.
20. Kateryna Chumachenko, "Machine Learning Methods for Malware Detection And Classification", 2017.
21. Jemili Farah, Zaghdoud M., and Ben Ahmed M.A framework for an adaptive intrusion detection system using Bayesian network. Intelligence and Security Informatics, IEEE, 2007.
22. Singh Satnam, Donat William, Pattipati Krishna, and Willet Peter. Anomaly detection via feature-aided tracking and hidden markov model. Aerospace Conference, IEEE, pages 1–18, March 2007.

AUTHORS PROFILE

Ms. V. R. Niveditha has completed B.E in Comp. Sci. Engineering in PB college of Engineering and M.Tech Information Security and Cyber forensics in Dr. M.G.R Educational and Research Institute. She also published two papers in International journals and six papers in National conferences.

Dr. T.V. Ananthan is currently working as Professor in Dept. of Comp. Sci. & Eng. and Inf. tech., Dr. M.G.R. Educational and Research Institute, Chennai. He has 26 years of teaching experience. He completed his Doctorate of Philosophy in the year 2012. He has published more than 40 papers in various International and National Journals and Conferences. His area of research is Networking, Wireless Communication, Mobile Communication and Network Security. He is a member in IEEE and IEI. He is also playing the role of reviewer in International Journals.