International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Android Malware Detection Using Genetic Algorithm based Optimized Feature Selection and Machine Learning

Tharun Kumar Bajineni[1], Venkata Sai Ramadugu[2], Sri Charan Reddy Konatham[3], Surya Prakash Amballa[4], Somashekar Yaggadi[5], G. Vijaya Lakshmi[6]

[1, 2,3,4,5] UG students, Dept of CSE, ANURAG Engineering College, Ananthagiri, Suryapet, TS, India.

[6]Assistant Professor, Dept of CSE, ANURAG Engineering College, Ananthagiri, Suryapet, TS, India.

ABSTRACT:

Android is an open source free operating system and it has support from Google to publish android application on its Play Store. Anybody can developed an android app and publish on play store free of cost. This android feature attract cyber-criminals to developed and publish malware app on play store. If anybody install such malware app then it will steal information from phone and transfer to cyber-criminals or can give total phone control to criminal's hand. To protect users from such app we using machine learning algorithm to detect malware from mobile app. To detect malware from app we need to extract all code from app using reverse engineering and then check whether app is doing any mischievous activity such as sending SMS or copying contact details without having proper permissions. If such activity given in code then we will detect that app as malicious app. In a single app there could be more than 100 permissions (examples of permissions are transact, API call signature, on Service Connected, API call signature, bind Service, API call signature, attach Interface, API call signature, Service Connection, API call signature, android.Os.Binder, API call signature, SEND_SMS, Manifest Permission, java.lang.Class.getCanonicalName, API call signature etc.) which we need to extract from code and then generate a features dataset, if app has proper permission then we will put value 1 in the features data and if not then we will value 0. Based on those features dataset app will be mark as malware or good ware.

Keywords: API, Malware, DL, ML.

1. INTRODUCTION:

Malware is a major threat to the security of computer users which can cause huge financial losses to firm. With increasing applications of Internet of Things (IoT), this made attackers to target them. Malware has different names such as adware, rootkit, backdoor, ransomware, trojans, worms, spyware etc. i.e depending on the behavior, thus detecting these malwares became as an evolving problem for researchers. There are two types of malware analysis and detection mechanisms: static analysis and dynamic analysis. Examining and Extracting information from theexecutable file without running is Static Analysis. Running the malware and observing its behavior on the system is Dynamic analysis. With a new variant of malware, experts generally analyze the sample manually or create a program that can match with similarity of this class of malware. Recently, image classification has been improved a lot with the development of deep

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

learning techniques. Convolutional Neural Networks demonstrated better performance. Here feature engineering, feature learning and feature representation are automatically acquired.

2. LITERATURE SURVEY

In literature [1], the paper was published in the year 2018. They have performed the malware detection with the help of 300 malware files and 300 benign apk files, also they managed to generate only 183 malware and 300 benign gray-scale images. The other 117 malware samples were unable to generate into images because the apk files were corrupted or either that files did not contain classes.dex file. Also, the accuracy was much less in all the algorithms they used. They have detected with the help of three different classifier techniques namely the k nearest neighbour(KNN), Random Forest (RF), and Decision Tree(DT).

In literature[2], the paper was published in the year 2017. They had used different machine learning algorithms such as Naive Bayes, j48, random forest, Multiclass classifier and multilayer perceptron to detect android malware and evaluate the performance of each algorithm. Here they implemented a framework for classifying android applications with the help of the machine learning techniques to check whether it is a malware or normal application. For validating their system they have collected 3258 samples of android apps and those have to be extracted for every application, extract their features and have to train the models going to be evaluated with the help of classification accuracy and time taken for the model.

In literature [3], the paper was published in the year 2016. They have proposed a Robotium program in an Android sandbox that can trigger any android application automatically and monitor its behaviour. The program has a UI Identification automatic trigger program that can click the mobile applications in a meaningful order. The program was able to perform large-scale experiments. They also tried to build a decision model using behaviour that has collected with the help of the random forest algorithm. It has been able to determine whether the unknown application is malware and also shows its confidence value. They could store the result and also the confidence value of the unknown apk file in their database.

In literature [4], the paper was published in the year 2018. They have proposed the android malware detection system with the help of permissions, APIs, and also with the presence of different key apps information such as, the dynamic code, Reaction code, native code, cryptographic code, database, etc. as the feature to train and build classification model.

EXISTING SYSTEM

In the existing system, the application permissions are extracted to detect the malware and executed through the command prompt. A proper GUI was not provided to execute the tasks. All the commands were run through the command prompt. It was difficult for the non-technical user to use the system. And also Semantic analysis was not implemented.

DRAWBACKS

It is time taking as it is extracting manifest file and also it doesn't have GUI. Some malware sample could not be generated into images because the APK files are either corrupted or they did not class.dex file. They have mainly classified

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

data using Random Forest no other algorithm is used.

PROPOSED SYSTEM

Two set of Android Apps or APKs: Malware/Goodware are reverse engineered to extract features such as permissions and count of App Components such as Activity, Services, Content Providers, etc. These features are used as feature vector with class labels as Malware and Goodware represented by 0 and 1 respectively in CSV format.

To reduce dimensional of feature-set, the CSV is fed to Genetic Algorithm to select the most optimized set of features. The optimized set of features obtained is used for training two machine learning classifiers: Support Vector Machine and Neural Network. In the proposed methodology, static features are obtained from AndroidManifest.xml which contains all the important information needed by any Android platform about the Apps. Androguard tool has been used for disassembling of the APKs and getting the static features.

ADVANTAGES

Proposed a novel and efficient algorithm for feature selection to improve overall detection accuracy. Machine-learning based approach in combination with static and dynamic analysis can be used to detect new variants of Android Malware posing zero-day threats.

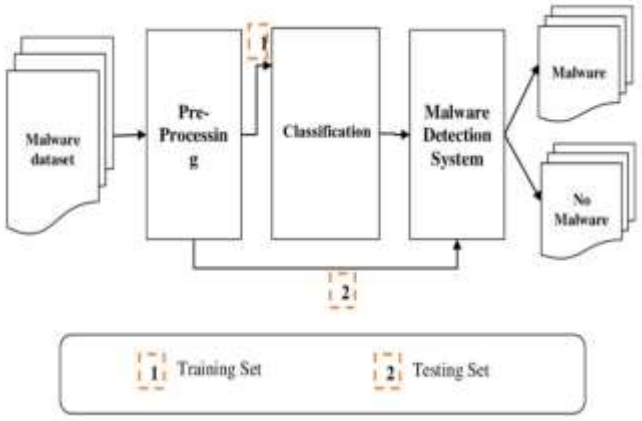

Fig.2. System Diagram.

Feature selection is an important part in machine learning to reduce data dimensionality and extensive research carried out for a reliable feature selection method. For feature selection filter method and wrapper method have been used. In filter method, features are selected on the basis of their scores in various statistical tests that measure the relevance of features by their correlation with dependent variable or outcome variable. Wrapper method finds a subset of features by measuring the usefulness of a subset of feature with the dependent variable. Hence filter methods are independent of any machine learning algorithm whereas in wrapper method the best feature subset selected depends on the machine learning algorithm used to train the model. In wrapper method a subset evaluator uses all possible subsets and then uses a classification algorithm to convince classifiers from the features in each subset. The classifier considers the subset of feature with which the classification algorithm performs the best. To find the subset, the evaluator uses different search techniques like depth first search, random search, breadth first search or hybrid search. The filter method uses an attribute evaluator along with a ranker to rank all the features in the dataset. Here one feature is

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

omitted at a time that has lower ranks and then sees the predictive accuracy of the classification algorithm. Weights or rank put by the ranker algorithms are different than those by the classification algorithm. Wrapper method is useful for machine learning test whereas filter method is suitable for data mining test because data mining has thousands of millions of features.

Fig.1. Data set upload.

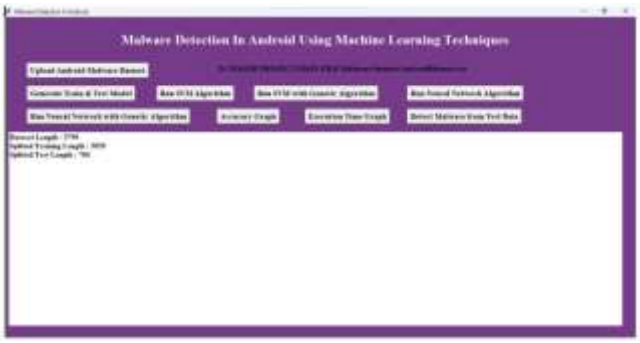

Fig.2. Data loaded.

Fig.3. Algorithm applied.

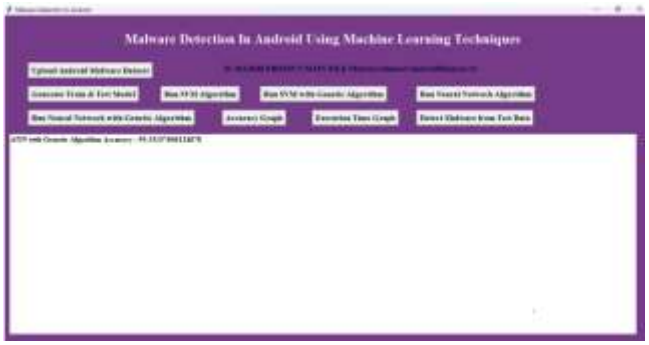

FIg.4. ANN accuracy display.

Fig.5. OUTPUT graphs.

Fig.6. Graph output.

CONCLUSION

As the number of threats posed to Android platforms is increasing day to day, spreading mainly through malicious applications or malwares, therefore it is very important to design a framework which can detect such malwares with accurate results. Where signature-based approach fails to detect new

variants of malware posing zero-day threats, machine learning based approaches are being used. The proposed methodology attempts to make use of evolutionary Genetic Algorithm to get most optimized feature subset which can be used to train machine learning algorithms in most efficient way.

REFERANCES

[1] D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, and K. Rieck, ―Drebin: Effective and Explainable Detection of Android Malware in Your Pocket,‖ in Proceedings 2014 Network and Distributed System Security Symposium, 2014.

[2] N. Milosevic, A. Dehghantanha, and K. K. R. Choo, ―Machine learning aided Android malware classification,‖ Comput. Electr. Eng., vol. 61, pp. 266–274, 2017.

[3] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-An, and H. Ye, ―Significant Permission Identification for Machine-Learning-Based Android Malware Detection,‖ IEEE Trans. Ind. Informatics, vol. 14, no. 7, pp. 3216–3225, 2018.

[4] A. Saracino, D. Sgandurra, G. Dini, and F. Martinelli, ―MADAM: Effective and Efficient Behaviorbased Android Malware Detection and Prevention,‖ IEEE Trans. Dependable Secur. Comput., vol. 15, no. 1, pp. 83–97, 2018.

[5] S. Arshad, M. A. Shah, A. Wahid, A. Mehmood, H. Song, and H. Yu, ―SAMADroid: A Novel 3- Level Hybrid Malware Detection Model for Android Operating System,‖ IEEE Access, vol. 6, pp. 4321–4339, 2018.