IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

Malicious Android Application Detection Method using Machine Learning

Divya Chaudhari[1], Arati Chaure[2], Shreyash Dhadke[3], Tushar Dhanawate[4], Prof. Shraddha Shirsath[5]
Students, Department of Computer Engineering[1,2,3,4]
Professor, Department of Computer Engineering[5]
Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

Abstract: *With the increasing popularity of the Android platform, we have seen the rapid growth of malicious Android applications recently. Considering that the heavy use of applications on mobile phones such as games, emails, and social network services has become a crucial part of our daily life, we have become more vulnerable to malicious applications running on mobile devices. This paper demonstrates on the problem of detecting malicious applications in the mobile internet, which is of great importance for personal information security and privacy security. We convert the android internet malicious application detection problem to a classification problem, and utilize the SVM classifier to solve it. Finally, we conduct an experiment to test the performance of the proposed method. Experimental results that the proposed can detect android internet malicious application with higher accuracy, true positive rate, and lower false positive rate.*

Keywords: SVM, Android internet, malicious application.

I. INTRODUCTION

In this technological era, smartphone usage and its associated applications are rapidly increasing due to the convenience and efficiency in various applications and the growing improvement in the hardware and software on smart devices. It is predicted that there will be 4.3 billion smartphone users by 2023. Android is the most widely used mobile operating system (OS). As of may 2021, its market share was 72.2% .The second highest market share of 26.99% is owned by apple iOS, while the rest of the 0.81% is shared among samsung, kaiOS, and other small vendors. Google play is the official app store for android-based devices. The number of apps published on it was over 2.9 million as of may 2021. Of these, more than 2.5 million apps are classified as regular apps, while 0.4 million apps are classified as low-quality apps by appbrain. Android's worldwide popularity makes it a more attractive target for cybercriminals and is more at risk from malware and viruses. Studies have proposed various methods of detecting these attacks, and ml is one of the most prominent techniques among them. This is because ml techniques are able to derive a classifier from a (limited) set of training examples. The use of examples thus avoids the need to explicitly define signatures in developing malware detectors. Defining signatures requires expertise and tedious human involvement and for some attack scenarios explicit rules (signatures) do not exist, but examples can be obtained easily. Numerous industrial and academic research has been carried out on ml-based malware detection on android, which is the focus of this review paper.

The first Android based smart device was developed in November 2008, and Android system rapidly goes into the smart system in the world. Two years later, Android acquired 48% world's smart device market. Meanwhile, as a famous development platform, Android has become the first choice of mobile application developers. In 2017, the number of android devices will exceed one billion and its mobile application downloading may be exceed 5 billion times. We can find that more and more Android's malicious applications will upload to various application platforms and bring great threats to users.

However, smart devices may save rich privacy information and user's personal data. For example, mobile payment has been popular by more and more people, and then users; personal information are memorized in mobile devices. But, once the smart phone is lost, it will bring a great threaten at people's privacy. Therefore, security problems are becoming more and more serious, and it is for great importance to detect mobile Internet

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

malicious applications.

Android has over one billion active users for all their mobile devices with a market impact that is influencing an increase in the amount of information obtained from different users, facts that have motivated the development of malware by cybercriminals to solve the problems caused by malware. Android implements a different architecture and security controls, such as unique user ID for each application, system permissions and its distribution platform google play.

We use android phone dataset to detect malicious application. We give Android phone data set as input. Then data set go to preprocessing, Segmentation phase after both phase done we use SVM algorithm to classification the Detect the malicious application.

II. RELATED WORK

1. PengTian and Xiaojun Huang, "A Malicious Application Detection Model to Remove the Influence of Interference API Sequence"

This paper proposes a new model for Detecting Android Malicious applications. The model obtains the API call sequences of APP runtime, and extracts features from them. These features have the highest correlation with malicious attributes detection, and have the characteristics of small redundancy between each other and noticed that API subsequences generated by normal behavior that may exist in a malicious application can interfere with the training of the detector. We use VSM and K-means combined with GBDT algorithm to eliminate this interference and improve the detection accuracy. Experiments show that this method can effectively eliminate the influence of interference API sequence and obtain higher detection accuracy.

2. Fei Chen, Yan Fu, "Dynamic Detection of Unknown Malicious Executables Based on API Interception"

In this paper, we propose a new approach for the dynamic detection of malicious executables on the platform of windows. Our approach extracts signatures of malicious executable's behaviors by using API (Application Program Interface) interception technique which makes possible the detection of unknown malicious executables. The dynamic detection of unknown malicious executables is achieved in three major steps: getting the sequence of API function calls of the executable, processing the API sequence to generate a vector, calculating the similarity between the vector and the feature library constructed by security policies to verify if the executable is malicious. The experiment confirms that this approach is effective in detection of unknown malicious executables

3. Yingbo Li, Jing Fang and Cheng Liu*, "Study on the Application of Dalvik Injection Technique for the Detection of Malicious Programs in Android".

With the increasing popularization of smart phones in life, malicious soft- ware targeting smart phones is emerging in an endless stream. As the phone system possessing the highest current market share, Android is facing a full-scale security challenge. This article focuses on analyzing the application of Dalvik injection technique in the detection of Android malware. Modify the system API (Application Program Interface) through Dalvik injection technique can detect the programs on an Android phone directly. Through the list of sensitive API called by malicious programs, eventually judge the target program as malicious or not.

The studies conducted in static analysis techniques used for Android applications from 2011 to 2015. The tools that can be used to perform Android code analysis using static analysis techniques were also summarised. Abstract representation, taint analysis, symbolic execution, program slicing, code instrumentation, and type/model checking were identified as fundamental analysis methods. Though this review correctly identified the most widely used approach to detect privacy and security related issues, the applicability of static analysis techniques for malware detection was not discussed. Apart from that, it did not take into account the recent research where novel analysis methods and malware detection methods were suggested. The study conducted in [35] provided a good systematic review mainly about static analysis techniques that can be used in Android malware detection. Four methods were identified as characteristic-based, opcode-based, program graph-based and symbolic execution-based. After that, it evaluated the capabilities of static analysis based Android malware detection methods on those four methods using the existing literature. The paper has identified ML and statistical models as possible methods by which Android malware

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

can be identified. However, ML-based machine learning methods have not been thoroughly reviewed as the main focus is only on the static analysis techniques.

III. SYSTEM ARCHITECTURE

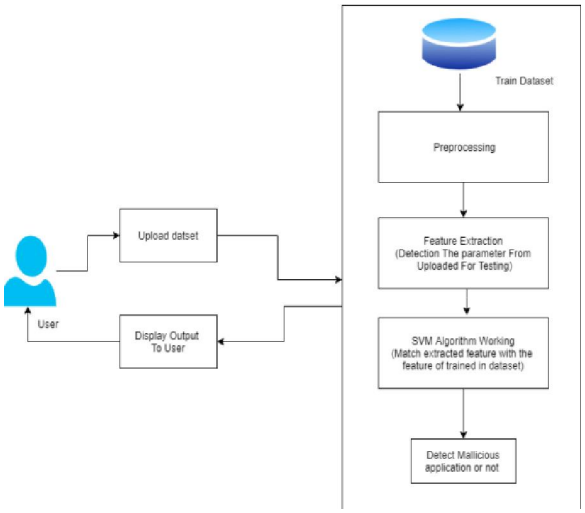


Figure: System Architecture

Malware attacks are the most common case that can be identified as a threat to Android. There are various definitions for malware given by many researchers depending on the harm they cause. The ultimate meaning of the malware is any of the malicious application with a piece of malicious code which has an evil intent to obtain unauthorised access and to perform neither legal nor ethical activities while violating the three main principles in security: confidentiality, integrity, and availability.

Step 1: Admin- the Admin has to log in by using valid user name and password.

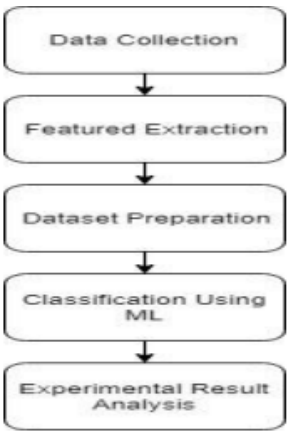
Step 2: View and Authorize Users the admin can view the user's details such as, user name, email, address and admin authorizes the users. Step 3: View Charts Results- View all products search ratio,view all keyword search results,view all product review rank results..

Step 4: Ecommerce User- There are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database.

Step 5: End User- There are n numbers of users are present. User should register before doing any operations. Once user registers, their details will best or to the database

IV. METHODOLOGY

This section discuss the methodology used to classify the Android applications into malicious and benign on the basis of permission analysis.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

A) Data Collection

We downloaded multiple Android application samples both malicious and benign. We have collected 2500 malicious Android applications from various websites like VirusShare, zeltseretc and 1500 benign application from Google‟s Android play store and other trusted website.

B) Feature Extraction

Firstly, we studied about the malware and the permissions and the different categories of permissions that an application seeks during its functioning. Android permissions are divided in to several protection levels:

1. Normal Permissions are those permissions which have very little risk of user‟s privacy. These permissions do not require user‟s involvement; these are granted by the Android system directly. eg. BLUETOOTH, INTERNET, SET_ALARM, etc.
2. Signature Permissions are permission granted by the system at the install time. These are only granted to an app when it is signed by the same certificate as the app that defines the permission. eg. BIND_DEVICE_ADMIN, BIND_NFC_SERVICE, etc.
3. Special Permissions -The permissions that doesn‟t comes under the normal and dangerous are the special permissions SYSTEM_ALERT_WINDOW and WRITE_SETTINGS are particularly sensitive, if an application wants to access these it must declare it in the manifest and access those with the help of intents.
4. Dangerous permissions -These are the permissions which require access to users private data. For granting these permissions a message is prompt on the screen asking about the user‟s permissions. eg. READ_CALANDER, CAMERA, EAD_CONTACTS, etc. The permissions used in any Android application are found in the manifest.xml file of the zipped. apk package. We developed a Python script to extract these permissions and list them in a CSV file.

C) Dataset Preparation

In this paper, we have used an approach which is aimed at uncovering the already known malware families and also the unknown malware to reduce chances of malware in the android community from escaping detection from scanners. For this we created a dataset using multiple Android .apk samples downloaded from both google play and VirusShare and other trusted sites providing malware samples. We got a collection of 4000 samples of both malicious and benign android application samples. As the permissions required by a particular application is inside the android manifest file of the android sample, we have created a script in python which reads and processes multiple samples at the same time and accesses the manifest.xml file and extract permissions and compile the permissions into a CSV file format which could be further used as an input file to the machine learning algorithms.

Machine learning algorithms are loosely classified into supervised and unsupervised learning algorithms. [29] Now we will discuss about different ML algorithms:

- **Logistic Regression**- Logistic regression (LR) is an algorithm which uses the statistical concepts and models a relationship between the input and output numerical values.
- **K-Nearest Neighbour (KNN)**- K-Nearest Neighbours (KNN) is a type of algorithm which can be used both for regression and classification problems but is mostly used in classification problems. This algorithm is easy in interpretation and requires very low calculation time and thus is a widely used ML algorithm. The K in this algorithm is the number of neighbours which are defined by the user.In this algorithm we use the Euclidean distance to measure the K nearest neighbours of the data point and predict the output according to its neighbours.Euclidean distance function: In Cartesian coordinates, if p = (p1, p2,..., pn) and q = (q1, q2,..., qn) are two points in Euclidean n-space, then the distance (d) from p to q, or from q to p.
- **Decision Tree**- Decision tree (DT) algorithm is a type of supervised learning algorithm in which a data structure is used to solve a problem. In this case the leaf node is referred to as the class label and the internal nodes of the tree represent the attributes.
- **Gaussian Naïve Bayes**- Gaussian Naive Bayes (GNB) theorem is a type of classification algorithm which can be used for both binary and multi class classification problems. This theorem is called so because it has its roots of Bayes theorem. Naïve Bayes is often represented by probabilities.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

- **Support Vector Classifier**- SVC is a supervised machine learning algorithm which is commonly used for both regression and classification problems.

D) Experimental Result Analysis

Experimental results are discussed in the following section i.e section IV.

V. EXPERIMENTAL RESULTS

This section discusses the experimental results obtained after applying the following ML algorithms on the created dataset as explained in previous section.

Table 1 Confusion Matrix

		Predicted Class	
Actual Class		Yes	No
	Yes	True Positive	False Negative
	No	False Positive	True Negative

V. CONCLUSION

In this paper, we propose an effective method to detect malicious applications in Android platform. The APK file directory structure is analyzed in detail to build up the feature vector for Android malicious applications detection. Main idea of this paper lies in that we regard the mobile Internet malicious application detection problem as a classification problem. Experimental results prove the effective of the proposed method.

REFERENCES

[1]. Number of Mobile Phone Users Worldwide from 2016 to 2023 (In Billions). Available online: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/ (accessed on 19 May 2021).

[2]. Mobile Operating System Market Share Worldwide. Available online: https://gs.statcounter.com/os-market-share/mobile/worldwide/ (accessed on 19 May 2021).

[3]. Number of Android Applications on the Google Play Store. Available online: https://www.appbrain.com/stats/number-of-android-apps/ (accessed on 19 May 2021).

[4]. Gibert, D.; Mateu, C.; Planes, J. The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. J. Netw. Comput. Appl. 2020, 153, 102526. [Google Scholar] [CrossRef]

[5]. Khan, J.; Shahzad, S. Android Architecture and Related Security Risks. Asian J. Technol. Manag. Res. [ISSN: 2249–0892] 2015, 5, 14–18. Available online: http://www.ajtmr.com/papers/Vol5Issue2/Vol5Iss2_P4.pdf (accessed on 19 May 2021).

[6]. Platform Architecture. Available online: https://developer.android.com/guide/platform (accessed on 19 May 2021).

[7]. Android Runtime (ART) and Dalvik. Available online: https://source.android.com/devices/tech/dalvik (accessed on 19 May 2021).

[8]. Cai, H.; Ryder, B.G. Understanding Android application programming and security: A dynamic study. In Proceedings of the 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME), Shanghai, China, 17–22 September 2017; pp. 364–375. [Google Scholar] [CrossRef]

[9]. Stephen Feldman ,DillonStadther ,Bing Wang, "Manilyzer: Automated Android Malware Detectionthrough Manifest Analysis ," in 2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems.

[10]. William Enck, Peter Gilbert, Seungyeop Han,Vasant Tendulkar, Byung Gon Chun, Landon P. Cox, Jaeyeon Jung, Patrick Mcdaniel, Anmol N. Sheth, "TaintDroid: An Information-Flow Tracking System for Real time Privacy Monitoring on Smartphones," ACM Transactions on Computer Systems (TOCS) TOCS Homepage archive Volume 32 Issue 2, June 2014

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

[11]. G. Y. Wang, After Access: Inclusion, Development, and a More Mobile Internet, International Journal of Communication, 2017, 11:323-326

[12]. Q. Shi, X. Ding, J. Zuo and G. Zillante, Mobile Internet based construction supply chain management: A critical review, Automation in Construction, 2016, 72: 143-154

[13]. Liu, K.; Xu, S.; Xu, G.; Zhang, M.; Sun, D.; Liu, H. A Review of Android Malware Detection Approaches Based on Machine Learning. IEEE Access 2020, 8, 124579–124607. [Google Scholar] [CrossRef]

[14]. Gilski, P.; Stefanski, J. Android os: A review. Tem J. 2015, 4, 116. Available online: https://www.temjournal.com/content/41/14/temjournal4114.pdf (accessed on 19 May 2021).