IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657

Volume 12, Issue 3, 2024

Malware Detection A Framework for Reverse Engineered Android Applications through Machine Learning Algorithms

[1]A NAGARAJU, [2]P.DIVAKAR

[1](Assistant Professor), MSC, DNR college (A) PG courses Bhimavaram

[2]MSC, scholar, DNR college (A) PG courses Bhimavaram

ABSTRACT

Today, Android is one of the most used operating systems in smartphone technology. This is the main reason, Android has become the favorite target for hackers and attackers. Malicious codes are being embedded in Android applications in such a sophisticated manner that detecting and identifying an application as a malware has become the toughest job for security providers. In terms of ingenuity and cognition, Android malware has progressed to the point where they're more impervious to conventional detection techniques. Approaches based on machine learning have emerged as a much more effective way to tackle the intricacy and originality of developing Android threats. They function by first identifying current patterns of malware activity and then using this information to distinguish between identified threats and unidentified threats with unknown behavior. This research paper uses Reverse Engineered Android applications' features and Machine Learning algorithms to find vulnerabilities present in Smartphone applications. Our contribution is twofold. Firstly, we propose a model that incorporates more innovative static feature sets with the largest current datasets of malware samples than conventional methods. Secondly, we have used ensemble learning with machine learning algorithms such as AdaBoost, SVM, etc. to improve our model's performance. Our experimental results and findings exhibit 96.24% accuracy to detect extracted malware from Android applications, with a 0.3 False Positive Rate (FPR). The proposed model incorporates ignored detrimental features such as permissions, intents, API calls, and so on, trained by feeding a solitary arbitrary feature, extracted by reverse engineering as an input to the machine.

1. INTRODUCTION

To this degree, it is guaranteed that mobile devices are an integral part of most people's daily lives. Furthermore, Android now controls the vast majority of mobile devices, with Android devices accounting for an average of 80% of the global market share over the past years. With the ongoing plan of Android to a growing range of smart phones and consumers around the world, malware targeting Android devices has increased as well. Since it is an open-source operating system, the level of danger it poses, with malware authors and programmers implementing unwanted permissions, features and application components in Android apps. The option to expand its capabilities with third-party software is also appealing, but this capability comes with the risk of malicious device attacks. When the number of smart phone apps increases, so does the security problem with unnecessary access to different personal resources. As a result, the applications are becoming more insecure, and they are stealing personal information, SMS frauds, ransom ware, etc.

In contrast to static analysis methods such as a manual assessment of AndroidManifest.xml, source files and Dalvik Byte Code and the complex analysis of a managed environment to study the way it treats a program, Machine Learning includes learning the fundamental rules and habits of the positive and malicious settings of apps and then data-venabling. The static attributes derived from an application are extensively used in machine learning methodologies and the tedious task of this can be relieved if the static features of reverse-engineered Android Applications are extracted and use machine learning SVM algorithm, logistic progression, ensemble learning and other algorithms to help train the model for prediction of these malware applications

2. EXISTINGSYSTEM

The methods proposed in this related work contribute to key aspects and a higher predictive rate for malware detection. Certain research has focused on increasing accuracy, while others have focused on providing a larger dataset, some have been implemented by employing various feature sets, and many studies have combined all of these to improve detection rate efficiency. In [21] the authors offer a system for detecting Android malware apps to aid in the organization of the Android Market. The proposed framework aims to provide a machine learning-based malware detection system for Android to detect malware apps and improve phone users' safety and privacy. This system monitors different permission-based characteristics and events acquired from Android apps and examines these features employing machine learning classifiers to determine if the program is goodware or malicious.

The paper uses two datasets with collectively 700 malware samples and 160 features. Both datasets achieved approximately 91% accuracy with Random Forest (RF) Algorithm. [22] Examines 5,560 malware samples, detecting 94 % of the malware with minimal false alarms, where the reasons supplied for each detection disclose key features of the identified malware. Another technique [23] exceeds both static and dynamic methods that rely on system calls in terms of resilience. Researchers demonstrated the consistency of the model in attaining maximum classification performance and better accuracy compared to two state-of-the-art peer methods that represent both static and dynamic methodologies over for nine years through three interrelated assessments with satisfactory malware samples from different sources. Model continuously achieved 97% F1- measure accuracy for identifying applications or categorizing malware.

Disadvantages

- The system is not implementedmachine learning algorithm and ensemble learning.
- The system is not implementedReverse Engineered Applications characteristics.

3. PROPOSED SYSTEM

We present a novel subset of features for static detection of Android malware, which consists of seven additional selected feature sets that are using around 56000 features from these categories. On a collection of more than 500k benign and malicious Android applications and the highest malware sample set than any state-of-the-art approach, we assess their stability. The results obtain a detection increase in accuracy to 96.24 % with 0.3% false-positives.

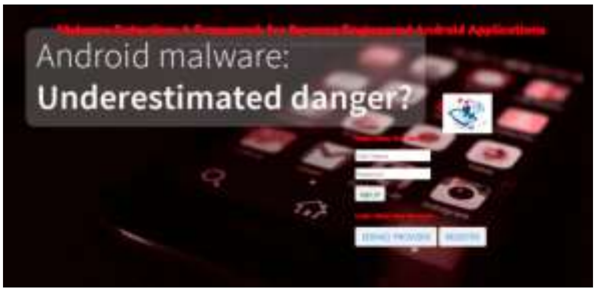
With the additional features, we have trained six classifier models or machine learning algorithms and also implemented a Boosting ensemble learning approach (AdaBoost) with a Decision Tree based on the binary classification to enhance our prediction rate. Our model is trained on the latest and large time aware samples of malware collected within recent years including the latest Android API level than state-ofthe-art approaches.

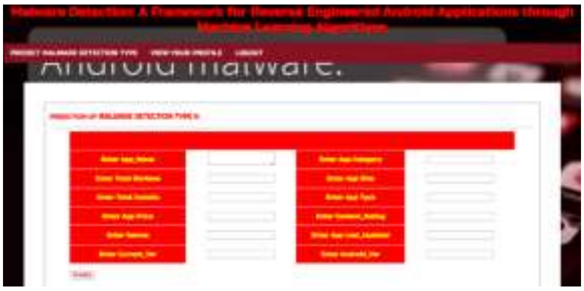
Advantages

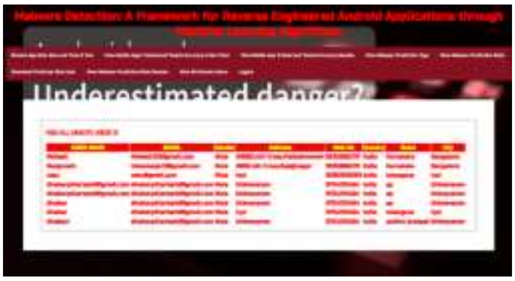
- The proposed system chooses the characteristics based on their capability to display all data sets. Enhanced efficiency by reducing the dataset size and the hours wasted on the classification process introduces an effective function selection process.

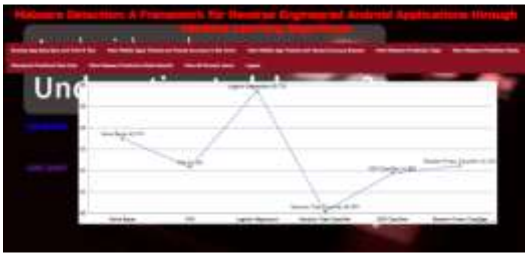
- The system used in this study also incorporates larger feature sets for classification. Although this problem arises in machine learning quite often to some extent choosing thetype of model for detection or classification can highly impact the high dimensionality of the data being used.

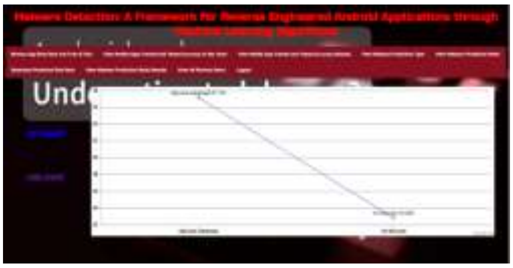
4. OUTOUT SCREENS

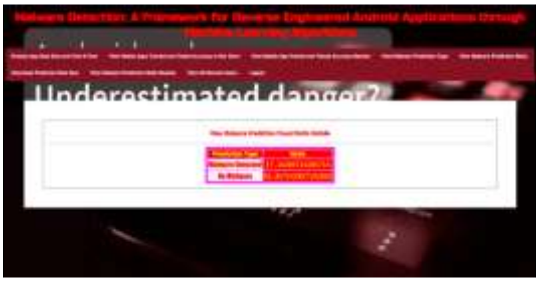


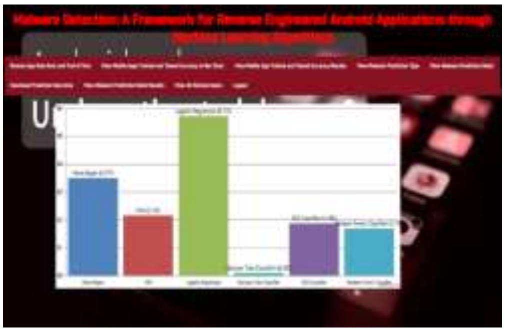









Malware Detection: A Framework for Reverse Engineered Android Applications through Machine Learning Algorithms


Malware Detection: A Framework for Reverse Engineered Android Applications through Machine Learning Algorithms

5. CONCLUSION

In this research, we devised a framework that can detect malicious Android applications. The proposed technique takes into account various elements of machine learning and achieves a 96.24% in identifying malicious Android applications. We first define and pick functions to capture and analyze Android apps' behavior, leveraging reverse application engineering and AndroGuard to extract features into binary vectors and then use python build modules and split shuffle functions to train the model with benign and malicious datasets. Our experimental findings show that our suggested model has a false positive rate of 0.3 with 96% accuracy in the given environment with an enhanced and larger feature and sample sets. The study also discovered that when dealing with classifications and high-dimensional data, ensemble and strong learner algorithms perform comparatively better. The suggested approach is restricted in terms of static analysis, lacks sustainability concerns, and fails to address a key multi collinearity barrier. In the future, we'll consider model resilience in terms of enhanced and dynamic features. The issue of dependent variables or high inter correlation between machine algorithms before employing them is also a promising field.

6. REFERENCES

[1] A. O. Christiana, B. A. Gyunka, and A. Noah, “Android Malware Detection through Machine Learning Techniques: A Review,” Int. J. Online Biomed. Eng. IJOE, vol. 16, no. 02, p. 14, Feb. 2020, doi: 10.3991/ijoe.v16i02.11549. [2] D. Ghimire and J. Lee, “Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines,” Sensors, vol. 13, no. 6, pp. 7714–7734, Jun. 2013, doi: 10.3390/s130607714. [3] R. Wang, “AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review,” Phys. Procedia, vol. 25, pp. 800–807, 2012, doi:

10.1016/j.phpro.2012.03.160. [4] J. Sun, H. Fujita, P. Chen, and H. Li, “Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble,” Knowl.-Based Syst., vol. 120, pp. 4–14, Mar. 2017, doi: 10.1016/j.knosys.2016.12.019. [5] A. Garg and K. Tai, “Comparison of statistical and machine learning methods in modelling of data with multicollinearity,” Int. J. Model. Identif. Control, vol. 18, no. 4, p. 295, 2013, doi: 10.1504/IJMIC.2013.053535.