iJRASET

International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37673

Significant Permission Identification for Machine Learning Based Android Malware Detection

Acquiline[1], Abhijith A[2], Athul Babu[3], Amrutha Muralidharan Nair[4]
[1, 2, 3]Dual Degree MCA, Department of Computer Science, De Paul Institute of Science & Technology, Angamaly, MG University
[4]Asst. Professor, Department of Computer Science, De Paul Institute of Science & Technology, Angamaly, MG University

Abstract: The dreadful rate of growth of malicious apps has become a significant issue that sets back the prosperous mobile scheme. A recent report indicates that a brand new malicious app for golem is introduced each ten seconds. To combat this serious malware campaign, we'd like a scalable malware detection approach that may effectively and expeditiously determine malware apps. varied malware detection tools are developed, together with system-level and network-level approaches. However, scaling the detection for an outsized bundle of apps remains a difficult task. during this paper, we tend to introduce SIGPID, a malware detection system supported permission usage analysis to address the speedy increase within the range of golem malware. rather than extracting and analyzing all golem permissions, we tend to develop 3-levels of pruning by mining the permission information to spot the foremost important permissions that may be effective in identifying between benign and malicious apps. SIGPID then utilizes machine-learning based mostly classification ways to classify totally different families of malware and benign apps. Our analysis finds that solely twenty two permissions square measure important. we tend to then compare the performance of our approach, victimisation solely twenty two permissions, against a baseline approach that analyzes all permissions. The results indicate that once Support Vector Machine (SVM) is employed because the classifier, we are able to bring home the bacon over ninetieth of preciseness, recall, accuracy, and F-measure, that square measure concerning constant as those created by the baseline approach whereas acquisition the analysis times that square measure four to thirty two times but those of victimisation all permissions. Compared against alternative progressive approaches, SIGPID is more practical by sleuthing ninety three.62% of malware within the information set, and 91.4% unknown/new malware samples.

Keywords: SIGPID (Significant Permission Identification), SVM(Support Vector Machine), Android, Malware, Benign, Data pruning.

I. INTRODUCTION

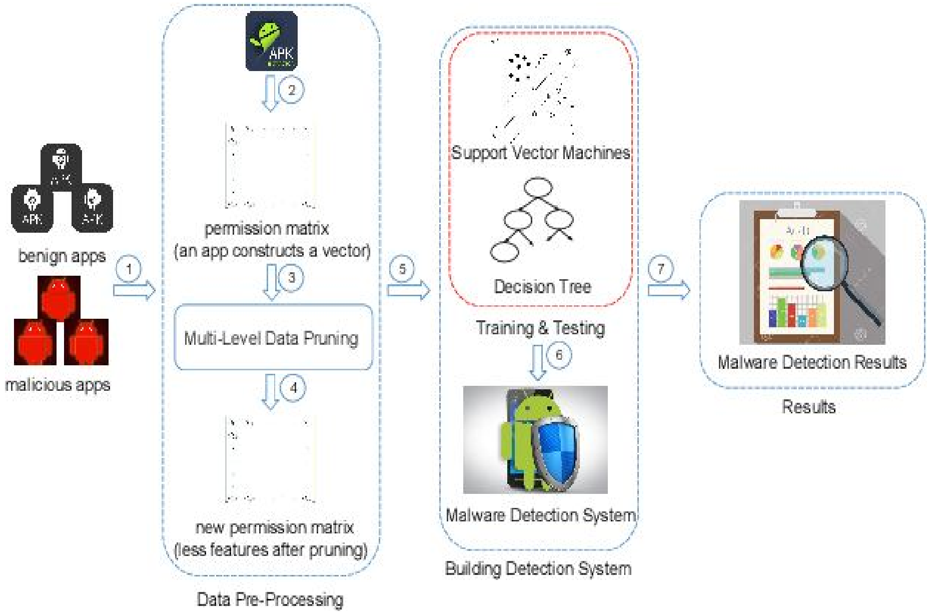
Android is presently the foremost used smart-mobile device platform across the world, occupying eighty fifth of market share. Now, there area unit nearly three million apps accessible for downloading from Google Play and over sixty five billion downloads up to presently. the recognition of automaton as well spurs interests from cyber-criminals United Nations agency prove malicious apps which is able to steal sensitive data and compromise mobile systems. The malware infection issue has been therefore serious that a recent report indicates that cardinal of all mobile malware target automaton devices in 2016 alone, over 3.25 million new malicious automaton apps unit of measurement uncovered. This roughly interprets to honor introduction of a replacement malicious automaton app each 10s. These malicious apps area unit created to perform differing types of attacks at intervals the sort of trojans, worms, exploits, and viruses. Some unrespectable malicious apps have over fifty variants, that produces it very hard to sight all. So, we've Associate in Nursing inclination to propose a malware detection system supported permission usage analysis to handle the short increase at intervals the vary of automaton malware. rather than extracting and analyzing all automaton permissions, we've Associate in Nursing inclination to develop pruning by mining the permission data to spot the foremost necessary permissions which is able to be effective in distinctive between benign and malicious apps. during this paper, we tend to tend to gift SIGPID, degree approach that extracts very important permissions from apps, and uses the extracted data to effectively sight malware pattern supervised learning algorithms. the planning objective of SIGPID is to sight malware efficiently and accurately. Our approach analyzes permissions then identifies alone people who unit of measurement very important in characteristic between malicious and benign apps.

II. EXISTING FRAMEWORK

1) Mobile malware makes an attempt to evade detection throughout app analysis by mimicking security-sensitive behaviors of benign apps that offer similar practicality (eSMS messages), and suppressing their payload to cut back the possibility of being discovered (e.g., corporal punishment solely its payload at night). Since current approaches focus their analyses on the kinds of securitysensitive resources being accessed (e.g., network), these evasive techniques in malware build differentiating between malicious and benign app behaviors a tough task throughout app analysis. The malicious and benign behaviors among apps may be differentiated supported the contexts that trigger security sensitive behaviors, i.e., the events and conditions that cause the security-sensitive behaviors to occur. during this work, we tend to introduce AppContext, associate approach of static program analysis that extracts the contexts of security-sensitive behaviors to help app analysis in differentiating between malicious and benign behaviors. we tend to implement a model of AppContext and measure AppContext on 202 malicious apps from numerous malware datasets, and 633 benign apps from the Google Play Store. The spite of a security-sensitive behavior is a lot of closely associated with the intention of the behavior (reflected via contexts) than the sort of the security-sensitive resources that the behavior accesses.
2) The wide-spreading mobile malware has become a dreadful issue within the progressively common mobile networks. Most of the mobile malware depends on network interface to coordinate operations, steal users' non-public info, and launch attack activities. TextDroid, an efficient and automatic malware detection methodology combining tongue process and machine learning. TextDroid will extract distinguishable options (n-gram sequences) to characterize malware samples. A malware findion model is then developed to detect mobile malware employing a Support Vector Machine (SVM) classifier. The trained SVM model presents a superior performance on 2 completely different information sets, with the malware detection rate reaching ninety six.36% within the check set and seventy six.99% in associate app set captured within the wild, severally. additionally, we tend to conjointly style a flow header mental image methodology to see the highlighted texts generated throughout the apps' network interactions, that assists security researchers in understanding the apps' advanced network activities.
3) AntMonitor uses the humanoid VPN service API to intercept traffic on humanoid devices and perform traffic analysis. A system that may mechanically establish mobile apps by regularly learning the apps' distinguishable options via HTTP traffic observations. many studies utilize text analysis for the aim of distinctive malicious behaviors.
4) Asdroid detects sneaky behaviors in humanoid apps by UI matter linguistics and program behavior contradiction. However, it solely uses a couple of keywords to hide sensitive operations like "send sms","call phone". WHYPER uses informatics techniques to spot sentences that describe the necessity for a given permission within the app description.
5) A framework referred to as UIPicker for distinctive users' personal info on an oversized scale that is predicated on a completely unique combination of informatics, machine learning and program analysis techniques. As for the traffic analysis, associate Ngram model in informatics has been utilized in associate automatic network protocol identification system. The projected system initial extracts data point message format by clump the N-grams with identical linguistics, so the data point format is employed to classify the raw traffic information. Note that none of the on top of work focuses on mobile malware detection exploitation network flows. we tend to utilize associate N-gram model in informatics and therefore the linguistics correlation in HTTP flow header for mobile malware identification. Among all the options generated by the N-gram sequencing methodology, a feature choice algorithmic program is applied to mechanically choose options with high correlations to malware, which needs no previous information of the HTTP flows.

III.PROPOSED FRAMEWORK

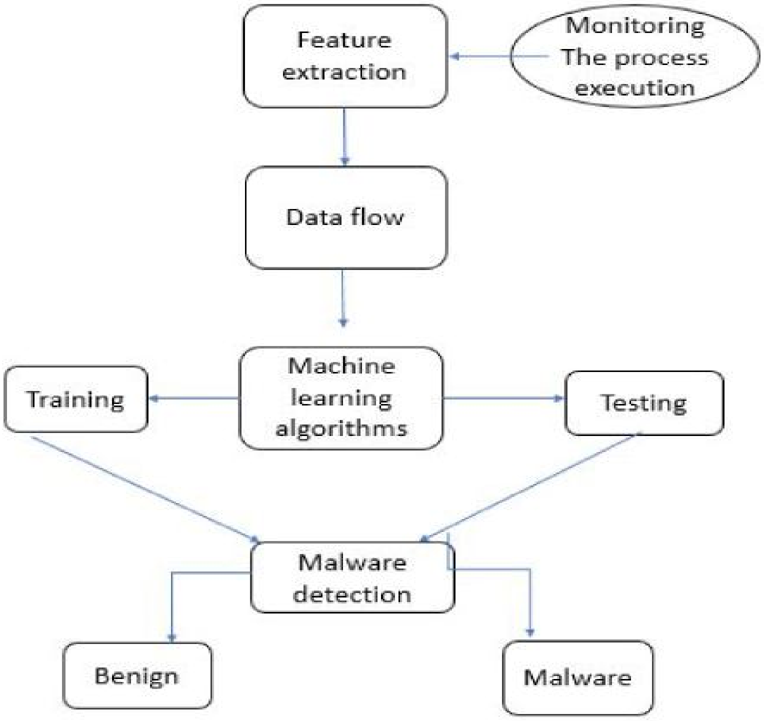
Recently, the threat of humanoid malware is spreading chop-chop, particularly those repackaged humanoid malware. though understanding humanoid malware exploitation dynamic analysis will give a comprehensive read, it's still subjected to high price in surroundings readying and manual efforts in investigation. This study propose a malware detection system supported permission usage analysis to deal with the speedy increase within the variety of humanoid malware. rather than extracting and analyzing all humanoid permissions, we tend to develop pruning by mining the permission knowledge to spot the foremost significpermissions which will be effective in identifying between benign and apps.In this system we tend to principally uses 3 ways to extract permissions from files. The 3 ways ar (i) Permission Ranking with Negative Rate,(ii) Support based mostly Permission Ranking ,(iii) Permission Mining With Association Rules (PMAR).



Architecture of proposed system.

IV. FRAME WORK DEMONSTRATION.

The goal of great Permission IDentification (SigPID) system is to realize high malware detection accuracy and potency whereas analyzing the littlest range of permissions. To do so, our system extracts permission lists from application packages however rather than that specialize in all permissions, SigPID primarily focuses on permissions that may improve the malware detection rate. This, in effect, eliminates the necessity to research permissions that have very little or no important influence on malware detection effectiveness. during a shell, SigPID prunes permissions that have low impacts on detection effectiveness victimization multi-level knowledge pruning, that consists of 3 major components: (i) permission ranking with negative rate; (ii) permission mining with association rules; and (iii) support primarily based permission ranking. once pruning, SigPID employs supervised machine learning classification strategies to spot potential humanoid malware. Finally, SigPID reports malware detection outline to the analysts.



V. FOCAL POINTS

A. Data Collection

The first step in the malware detection is the collection of apk files and these files are used for extraction.

B. Extract Permissions

Out of 135 permissions, 22 permissions are significant and are extracted based on pruning which is summerised as below :

1) Permission Ranking with Negative Rate : Typically, the number of benign apps tends to be much large than the number o f malicious apps therefore, the size of Bengin is much larger than the size of malware.

We can find rank using the below equation

$$R(P_j) = \frac{\sum_i M_{ij} - S_B(P_j)}{\sum_i M_{ij} + S_B(P_j)}$$

$$S_B(P_j) = \frac{\sum_i B_{ij}}{size(B_j)} * size(M_j)$$

Bij represents whether or not the jth permission is requested by the ith benign app sample.
Mij represents whether or not the jth permission is requested by the ith malware sample.
SB(Pj) represents the support of jth permission in matrix B
The result of R (Pj) has a value ranging between [-1, 1]. If
R(Pj) = 1, this means that permission Pj is only used in malicious dataset, which is a high risk permission.
If R(Pj) = -1, this means that permission Pj is only used in benign dataset.

2) Support based mostly Permission Ranking : some permissions square measure found solely in benign apps that can not be found in malware apps and vice-versa. as an example, we discover the permission BIND TEXT SERVICE solely in benign apps. As a result, we have a tendency to might contemplate that any app that uses BIND TEXT SERVICE is benign.

3) Permission Mining with Association Rules (PMAR): when pruning one hundred ten of a hundred thirty five permissions by mistreatment PRNR and SPR with PIS, we would like to more explore approaches which will take away non-influential permissions even additional. By inspecting the reduced permission list that contains twenty five important permissions, we discover 3 pairs of permissions that invariably seem along in Associate in Nursing app. as an example, permission WRITE SMS and permission browse SMS square measure invariably used along. They additionally each belong to the Google's "dangerous" permission list. Yet, it's reserve to think about each permissions, collectively of them is enough to characterize sure behaviors. As a result, we are able to associate one, that encompasses a higher support, to its partner.

C. Classification

Classify the options on the idea of targets like malware or benign.

D. Evaluation

The ultimate step is that the analysis method here accuracy,precision,recall,f1 score square measure computed from the classifier.

VI. CONCLUSIONS

In this paper, we've shown that it's potential to scale back the amount of permissions to be analyzed for mobile malware detection, whereas maintaining high effectiveness and accuracy. SIGPID has been designed to extract solely vital permissions through a scientific, 3-level pruning approach. supported our dataset, which incorporates over a pair of,000 malware, we tend to solely have to be compelled to contemplate twenty two out of one hundred thirty five permissions to enhance the runtime performance by eighty five.6% whereas achieving over ninetieth detection accuracy. The extracted vital permissions can even be utilized by different unremarkably used supervised learning algorithms to yield the F-measure of a minimum of eighty fifth in fifty five out of sixty seven tested algorithms. SIGPID is extremely effective, in comparison to the progressive malware detection approaches likewise as existing virus scanners. It will sight ninety three.62% of malware within the information set, and 91.4% unknown/new malware.

REFERENCES

[1] IDC, "Smartphone os market share, 2017 q1." [Online]. Available: https://www.idc.com/promo/smartphone-market-share/os

[2] Statista, "Cumulative number of apps downloaded from the google play as of may 2016." [Online]. Available: https://www.statista.com/ statistics/281106/number-of-android-app-downloads-from-google-play/

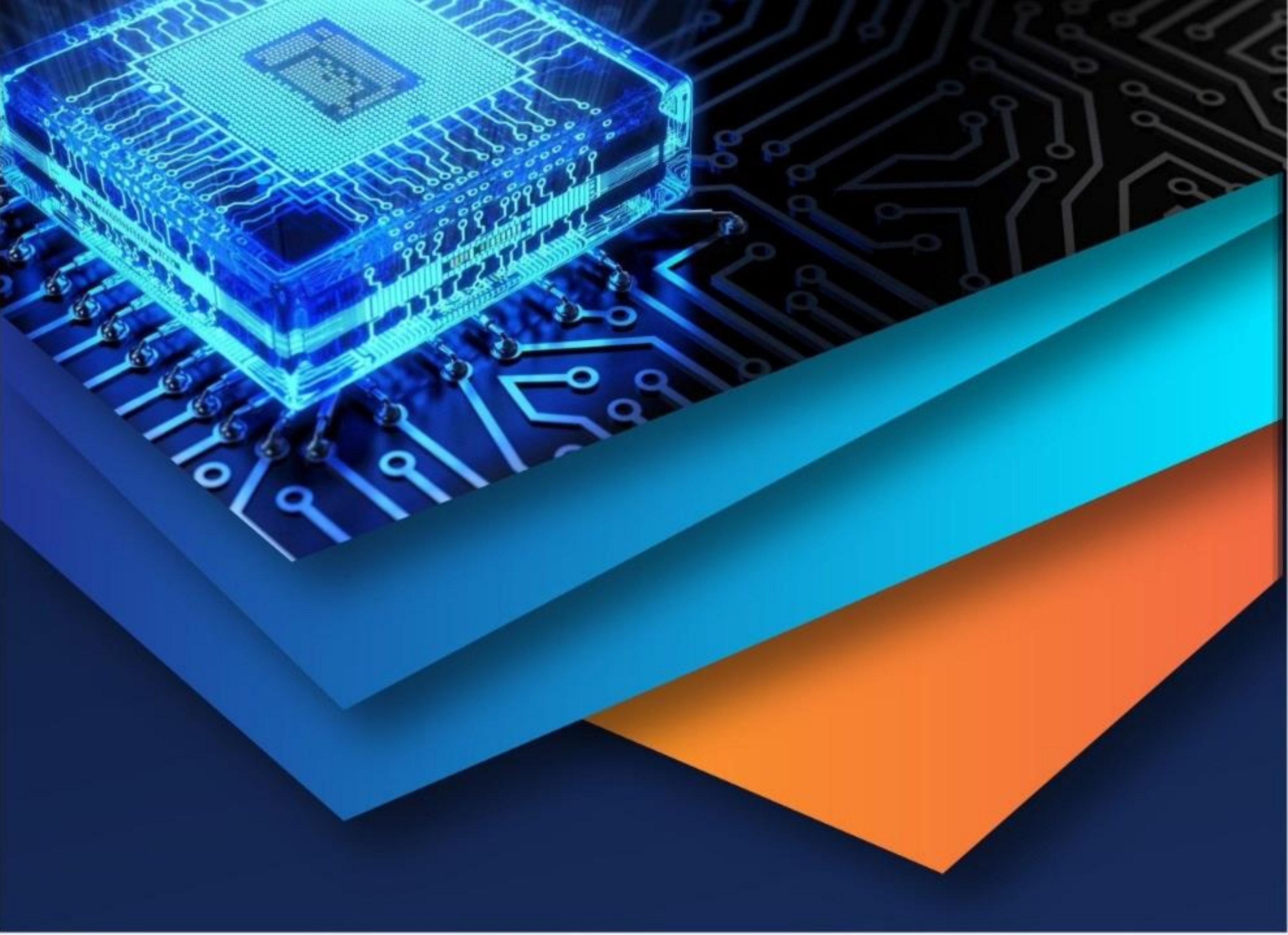
[3] G. Kelly, "Report: 97% of mobile malware is on android. this is the easy way you stay safe," in Forbes Tech, 2014.

[4] G. DATA, "8,400 new android malware samples every day." [Online]. Available: https://www.gdatasoftware.com/blog/2017/ 04/29712-8-400-new-android-malware-samples-every-day

[5] Symantec, "Latest intelligence for march 2016," in Symantec Official Blog, 2016.

[6] M. Grace, Y. Zhou, Q. Zhang, S. Zou, and X. Jiang, "Riskranker: scalable and accurate zero-day android malware detection," in Proceedings of the 10th international conference on Mobile systems, applications, and services. ACM, 2012, pp. 281–294.

[7] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner, "Android permissions demystified," in Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011, pp. 627–638.



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY