Scientific and Practical Cyber Security Journal (SPCSJ) 7(2): 34-44 ISSN 2587- 4667
Scientific Cyber Security Association (SCSA)

MALICIOUS ANDROID APPS DETECTION USING MACHINE LEARNING

Bilal Ahmad Kamal, Adeela Muhammad Askari, Salman Tariq
Air University Islamabad Pakistan

ABSTRACT: The use of smartphones has wisely evolved in the 20th century. Many people all over the world can connect to their smartphones in a variety of ways. Some invaders are leveraging the power of the rapidly growing smartphone usage to Developing rogue Android applications to steal handsets' sensitive data To address these grave issues, a malicious program that is both effective and efficient is required. Numerous malware detection programs have historically been built, but some of them are not able to identify recently developed malware programs or programs contaminated with different Trojan horses, worms, and spyware. The software for detecting fraudulent programs can be enhanced especially thanks to ML algorithms.

The system uses ML classification algorithms like Support Vector Machine (SVM) to improve the malware application detection in the proposed system. The proposed ML classification and fusion algorithms will improve performance metrics like the accuracy of malware application exposure and decrease the complexity of the detection process. The suggested approach integrates detecting software with a training application that users can install on Android cellphones to flag hazardous activities when they are accessible.

KEYWORDS: *support vector machine, ML, Android Apps, malicious*

1. INTRODUCTION

Smartphone usage has been gradually increasing in recent years, along with the growth of Android app users. Given the rise in Android app users, hackers are developing malicious Android apps as a tool to steal sensitive data and commit identity theft/fraud on mobile wallets and banks. Tools and software for detecting malicious applications are widely accessible. However, to deal with and handle increasingly sophisticated harmful apps developed by intruders or hackers, malicious application detection systems must be effective and efficient.

In this project, we created a strategy for detecting anomalous Android apps using machine learning methods.

We must first generate a dataset of previously infected applications as a training set. Then, using the Supervised Learning method, we must compare the training dataset with the trained information to forecast the malware Android apps up to 95.2% of the time.

1.1 Machine Learning

Machine learning is a study that involves feeding data and knowledge in the form of observations and in-person interactions into computers so they can learn better over time and autonomously. The goal of machine learning is demonstrated by the aforementioned sentence. The main thought behind machine learning is to give computers knowledge through data so they can interact and observe the outside world. Machine learning is a subset of machine intelligence. It will let the computer learn and prepare itself to carry out numerous tasks much like a person. In machine learning, there are so many different types of algorithms that can be distinguished based on how they learn or by similarities in their design or functionality. The following variables are present in all possible combinations of machine learning algorithms. This is also the fundamental idea to comprehend the field. The variables that aid in comprehending machine learning principles are representation, evaluation, and optimization. To help a computer learn, representation is a collection of classifiers or machine language. Finding a scoring function is done through the process of evaluation. The final search optimization technique is frequently referred to as the classifier with the greatest score. According to the aforementioned metrics, the primary goal of algorithms is to generalize beyond the training samples themselves in order to obtain high-quality analysis data that has never been used. Although there are numerous distinct and various ways to train a machine, from employing

We must choose the best learning algorithm that may improve performance and provide us with the correct accuracy of the clustering to decision tree, two layers of ANN (Artificial Neural Network).

outcome. Because of their computing capability, reading machines are frequently helpful to humans. where computing forces can quickly spot trends and even highlight the key

features in a massive data set that humans would have otherwise overlooked. Humans frequently utilize machine learning to enhance their problem-solving skills, and many systems employ it to develop educated advisors on a range of issues.

1.2 Mobile Malware

Mobile malware is malicious software that targets mobile devices, especially wireless smartphones and Personal Digital Assistants (PDA). With the rise in popularity and complexity of PDA networks and wireless mobile devices, it has become more challenging to maintain protection and security from electronic attacks like viruses, worms, and other malware. Malicious software, commonly referred to as malware, is intended to attack a mobile device, such as a smartphone or tablet, in order to harm or interrupt it. The majority of mobile malware is made to compromise smartphones, allow criminal users to remotely manage them, and steal the user's personal data that is kept on the device.

All of our contact information as well as many other entries on our smartphone could be accessed by or deleted by mobile viruses. It could send an infected SMS to each and every phone in your contact list.

call directory and grow on your side over the network. Fraudulent phone bills, obtaining unsuitable content, and losing highly essential data that is saved on the smartphone are the top three issues that worry mobile consumers. Mobile devices have historically been utilised primarily for the programs that are built into them, or more recently, for malware that has been put into smartphones. As the demand for apps grows, hackers today insert several malwares inside applications. Once the programme is downloaded and installed on a smartphone, the malware runs and fixes itself to the device and may transfer a lot of personal data about the user without their knowledge or consent. Malware in applications has grown to be one of the biggest issues facing people today.

2. APPROACHES RELATED TO MALWARE DETECTION

The earlier research studies examine malware detection in smartphone harmful Android applications. The many methods for detecting malware are covered in this section.

A simple yet effective method for malware detection that specifies the Android APIs sub-set as classification functions was proposed by Jaemin Jung et al. in "Detecting Malicious Android Apps using API's Popularity and Relationships" [1]. The number of APIs utilised in an app is its feature because it depends on the use of a number of Android APIs to achieve its main goal. Their methodology creates two rating Android API lists: one for safe APIs and one for dangerous APIs. The most often used APIs by good applications are included in the benign API list, whereas the most frequently used APIs are included in the harmful API list. If the number of inverse values based on benign apps is more than the amount based on hazardous devices, they conclude that the device being offered is fine and determine whether the gadget is benign or harmful by comparing the two numbers. They will then presume that the presented system is malevolent. For the detection of Android malware, the suggested technique obtains an accuracy of 87.35 percent to 89.93 percent. However, it has a low detection accuracy and cannot quickly detect freshly developed malware. Android Malware Detection Using Parallel Machine Learning Classifiers, Suleiman et al., [2]. Droid Fusion can be used with both ensemble learners and traditional method students. They recommend using algorithms based on four rankings to combine the learners. The methods are utilized to provide a finished, better classification model for Android malware detection. A fusion classifier strategy that frequently focuses on multilevel architecture, the results of a Droid Fusion performance comparison with layered generalization. This is utilized to enhance the algorithm as well as to employ an ensemble of the algorithm and random forest learning.

It is challenging to identify which harmful or recently updated software is present. Their algorithms are becoming more complicated. "Pin droid: A revolutionary Android malware detection system employing ensemble learning approaches," Idrees et al., [3] They employed features like permissions and tries to enhance performance categorization and train machine learning models. They started their tests on a variety of device samples by contrasting the results of different algorithms like decision tables, decision trees, and random forests. The decision table, MLP, and decision tree classifiers were then combined using two different approaches. The algorithm is improved using this application, and efficiency is also improved using ensemble learning algorithms like the random forest algorithm. It is challenging to determine whether a programmer is recently updated or maliciously produced. This is not being used in a real-time environment. "Feature Selection and ensemble of classifiers for Android malware detection," Coronado et al., [4]. For the committee, they proposed and investigated a hybrid classifier strategy based on random forest and random group classifiers. In their method, a task that creates a model Meta ensemble incorporates random forest. For efficiency, learning methods like random forest algorithms are used. less training data sets were used. This is not carried out in real-time settings. Malicious programs that have just been updated or created are difficult to identify. "Machine learning helped Android malware classification," Milosevic et al., [5]. The tested classification fusion methodology based on Android permissions and source code-based analysis with static method is mentioned in the study. They employed classifiers like JRip, JRandom Forests, and linear regression. However, the experiments only employed a small sample of data.

The JRip method is applied to classification fusion in order to better both the algorithm itself and ensemble learning algorithms like random forests. This system does not operate in authentic. With a machine learning approach like SVM and L2 linear classifiers, Lindorfer et al., [6] "MARVIN: Efficient and comprehensive mobile app classification through static and dynamic analysis." This research examines a thorough and effective classification of mobile applications. The system rates the likelihood of malicious activity in a scale from 0 to 10 for various untested Android applications. Despite employing an efficient algorithm, they were unable to produce an exact result.

3. CURRENT MALWARE IDENTIFICATION APPROACHES

Many Android malware detection technologies have historically been created, but some of these solutions are not able to identify newly created malicious applications or malware applications that have been infected with different Trojan, worms, spyware, etc. The detection of several harmful applications among the millions of Android applications is still a difficult process when done the old-fashioned method. Additionally, in the current system, non-machine learning techniques for identifying malicious applications based on their traits, traits, and behaviours have been created, although the accuracy of diagnosis is still low.

3.1 Issues in the Current System

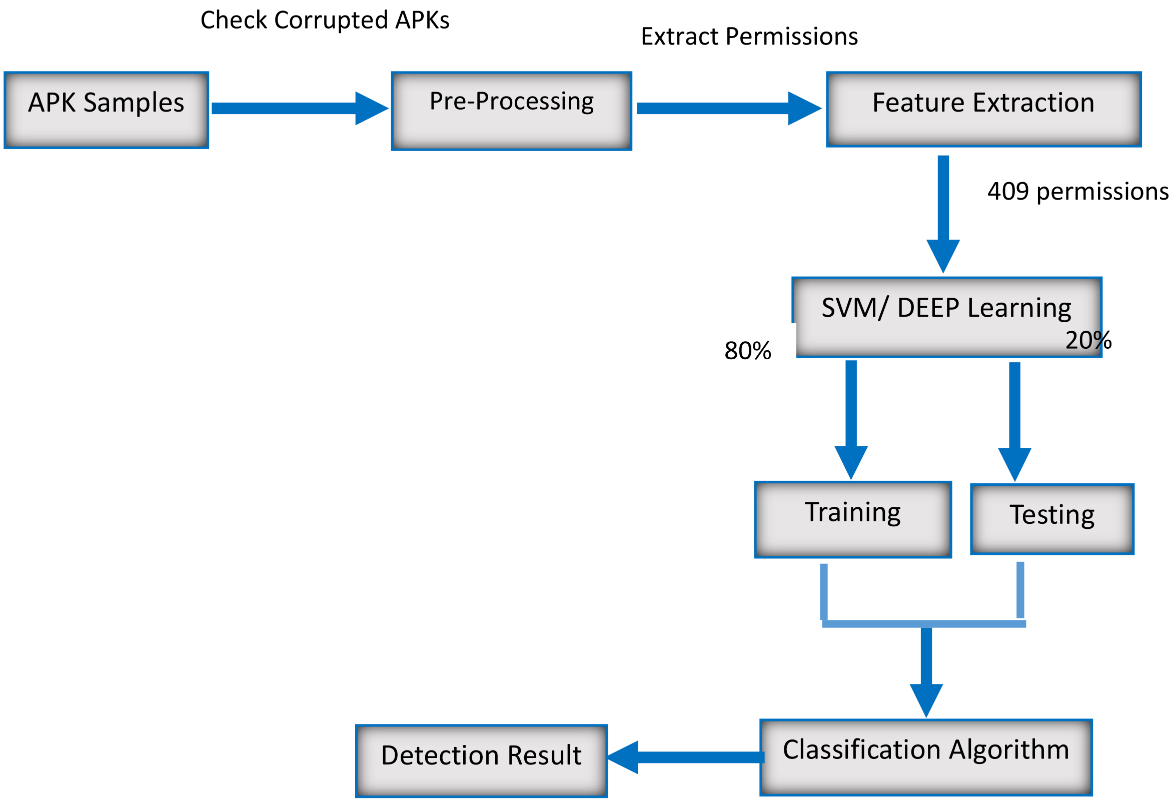
The detection of several harmful applications among the millions of Android applications is still a difficult process when done the old-fashioned method. This demonstrates that the technique for Android malware detection [12] using popular ML algorithms like random forest, K-means, does not accurately detect any other new malware. However, only detection was unable to show us how detection functions in a real-time setting. As we all know, the real-time view is crucial for giving us an authentic perspective of the software that has developed over the course of detection with the provided datasets, but when it comes to a real-time mobile device, it should be able to gather the datasets and identify malicious applications on its own.

4. MACHINE LEARNING CLASSIFIERS FOR MALWARE APP DETECTION

Classification techniques, the Support Vector Machine (SVM) system, is used in the proposed system to improve malware detection. The suggested ML classification and fusion techniques will effectively enhance the detection of malware applications and will improve performance metrics like the accuracy of revealing the malware applications and reducing detection time complexity. The suggested approach integrates the detection software into a trained application that can be installed on an Android smartphone and can identify dangerous applications in a smartphone user's accessibility from the beginning. The benefits of the suggested system include its reliability, ease of identifying newly updated or created malicious Android applications, effective and efficient detection, an increase in accuracy for exposing Android malware applications from 95% to 99%, and a decrease in time complexity for detection. Additionally, the system covers more program privileges so that we may find malware even in unexpected places. The original view of the software's method of machine learning is provided via real-time application.

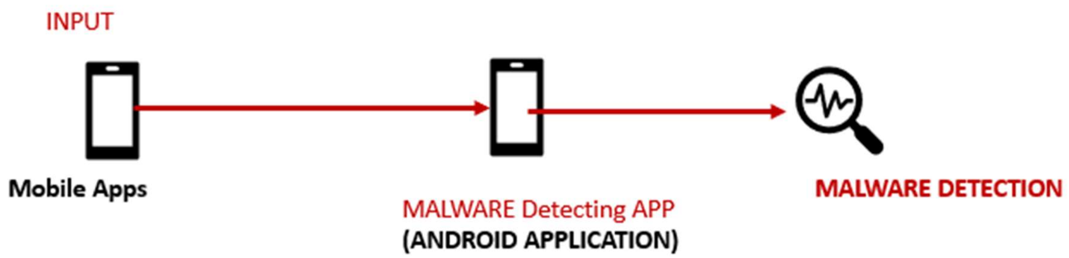
Implementation Diagram

At first stage, we provided apk samples to check corrupted apks and comes pre-processing to Extract Permissions. After extracting Permissions, we began to find 409 permissions whereas SVM/DEEP learning is used. After that Training and Testing was performed. Then it goes with Classification Algorithm and Detection results as drawn below.



SVM System.

It depicts the data classification and fusion system procedure. Fusion of classifications occurs in this module. To improve prediction accuracy, two or more classification algorithms are combined in a process called classification fusion. Most frequently, very accurate classified data are used. Due to the fact that we are utilizing algorithms—SVM. Therefore, by using this approach, we are enhancing the accuracy of the detection of malware from 95% to 99%, which can efficiently and effectively detect the malware in an application. The SVM method is a method in that fusion, where here each classified result from the algorithm is taken and by means of the SVM () the classification algorithm with high accuracy and consistency is voted and a further percentage of accuracy is drawn. In many existing works they have considered ranking method whereas in our case we are considering the SVM method. By means of this SVM the detection of malware in real time is also increased.



Malware Detection

The malware detection module procedure is depicted in, the real-time view is crucial for giving us a fresh perspective on the software that has learnbeened through dataset-based detection, but when it comes to a real-time mobile device, it should be able to gather the datasets and identify harmful applications on its own. Therefore, in our suggested method, we are integrating the detection software into a training application to install into an Android smartphone and identify the malicious application in an original view for the accessibility of the end users. The apps display the malicious and innocent scores for the app; if the malicious score is higher than the innocent score, "POTENTIAL THREAT" is displayed; otherwise, "SAFE" is displayed. Here, the malevolent score is determined by averaging Type 1 permissions, while the innocence score is determined by averaging Type 0 privileges.

Performance Analysis

Performance Evaluation

The efficiency of several categorization modules is shown in the aforementioned. These modules are improved and studied to produce effective results by comparing with the current and suggested techniques in the detecting attacks, such as:

Malicious App Identification Efficiency

Malware detection speed allows for quick and accurate identification of malware applications. With the suggested algorithms, such as SVM, detection is accomplished successfully and efficiently. For quick and accurate malware app detection, it covers additional app permissions.

Effectiveness of Malicious App Screening

Malicious detection accuracy demonstrates how each everything change's accuracy affects our ability to recognize malware apps. Methods based on machine learning are more trustworthy. The data fusion concept also improves the accuracy of revealing the infected application.

Malicious Software Detector Period

The length of time it took for an application to identify a smartphone's malware app is provided by the malware detection time. There is less time complexity for detection. The classification fusion technique improves app detection performance and speeds up the process.

5. EXPERIMENTAL EVALUATION

The suggested solution integrates the detection software into a training program by installing it into an Android smartphone using the Android Studio IDE and detecting the malicious application in a first-person perspective of the end user accessibility in smartphones. The malicious score and innocence score of the app are displayed in a safe real-time application, which displays "SAFE" if the malicious score is lower than the innocent score and "POTENTIAL THREAT" if the malicious score is higher than the innocence score. In this case, the malevolent score is determined by averaging Type 1 permissions, while the innocence score is determined by averaging Type 0 rights. By computing the entire permission score, it will also display the app's overall goodness %. The application is now prepared to find any malware that may be installed on smartphones.

Android Malicious Application Detection SVM, Version

The number of Malware Apps Found Compared to the Number of Mobile Apps Figure above compares the quantity of legitimate mobile apps to the quantity of malware-identified apps. It also demonstrates how SVM, improves the accuracy of malware program identification. The SVM has the highest accuracy in the Neural Network. The analysis shows clearly that the number of malware apps identified has been in increasing numbers with the SVM algorithm.

		Actual	
		Positive	Negative
Predicted	Positive	1480	8
	Negative	7	860

Figure. No. of Malware APPs identified Vs No. of Mobile Apps

Data Accuracy (%) vs. Number of Malware Apps Identified

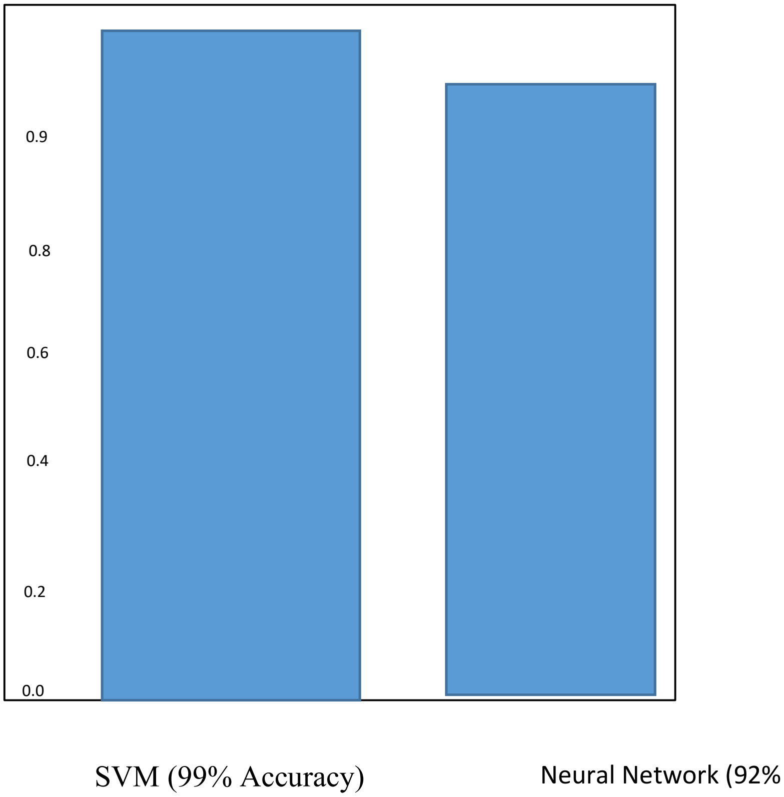
The amount of malware programs found about data accuracy is shown above. This graph displays the SVM algorithm's data accuracy for identifying all malicious apps. The Positive is regarded to be the data accuracy in percentage and the Negative is considered to be the number of malware applications found for all three methods, giving us an accurate result for detecting the proper amount of malware apps with high type 1 permissions.

Data Accuracy

The accuracy above displays the performance of each classification technique, including, SVM.

SVM vs Neural Network

Figure. Data Accuracy



CONCLUSION AND FUTURE WORK

One of a user's primary sources of income is now a smartphone. A person's complete details and data sources are now available on his or her smartphone for everyday use as smartphone usage grows. In that situation, an outside intruder could learn a lot about a user just installing the malicious application in the smartphone, making smartphone security one of the day's biggest issues of the day. As for as Future work is concerned, we can increase Neural Network accuracy from 92% which would be a great addition.

BIBLIOGRAPHY

1. Vinod, P., Zemmari, A., & Conti, M. (2019). A machine learning based approach to detect malicious Android apps using discriminant system calls. Future Generation Computer Systems, 94, 333-350.
2. Xiao, J. X., Lu, Z. C., & Xu, Q. H. (2018, December). A new Android malicious application detection method using feature importance score. In Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence (pp. 145-150).
3. Kambar, M. E. Z. N., Esmaeilzadeh, A., Kim, Y., & Taghva, K. (2022, January). A survey on mobile malware detection methods using machine learning. In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0215-0221). IEEE.
4. Arslan, R. S. (2021). AndroAnalyzer: Android malicious software detection based on deep learning. PeerJ Computer Science, 7, e533.
5. Jiang, X., Mao, B., Guan, J., & Huang, X. (2020). Android malware detection using fine-grained features. Scientific Programming, 2020.
6. de la Puerta, J. G., Pastor-López, I., Porto, I., Sanz, B., & Bringas, P. G. (2021). Detecting malicious Android applications based on the network packets generated. Neurocomputing, 456, 629-636.
7. Mohamed, S. E., Ashaf, M., Ehab, A., Shereef, O., Metwaie, H., & Amer, E. (2021, May). Detecting Malicious Android Applications Based On API calls and Permissions Using Machine learning Algorithms. In 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 1-6). IEEE.
8. Jung, J., Lim, K., Kim, B., Cho, S. J., Han, S., & Suh, K. (2019, June). Detecting malicious Android apps using the popularity and relations of APIs. In 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) (pp. 309-312). IEEE.
9. Razgallah, A., Khoury, R., Hallé, S., & Khanmohammadi, K. (2021). A survey of malware detection in Android apps: Recommendations and perspectives for future research. Computer Science Review, 39, 100358.
10. OS, J. N. (2021). Detection of malicious Android applications using Ontology-based intelligent model in mobile cloud environment. Journal of Information Security and Applications, 58, 102751.

11. Liu, L., Ren, W., Xie, F., Yi, S., Yi, J., & Jia, P. (2021). Learning-Based Detection for Malicious Android Application Using Code Vectorization. Security and Communication Networks, 2021.

12. Sharma, T., & Rattan, D. (2021). Malicious application detection in Android—a systematic literature review. Computer Science Review, 40, 100373.

13. Song, Y., Geng, Y., Wang, J., Gao, S., & Shi, W. (2021). Permission Sensitivity-Based Malicious Application Detection for Android. Security and Communication Networks, 2021.

14. Chen, X. R., Shi, S. S., Xie, C. L., Yang, Z., Guo, Y. J., Fang, Y., & Wen, W. P. (2021, February). SUIP: An Android malware detection method based on data flow features. In Journal of Physics: Conference Series (Vol. 1812, No. 1, p. 012010). IOP Publishing.

15. Sembera, V., Paquet-Clouston, M., Garcia, S., & Erquiaga, M. J. UNCOVERING AUTOMATIC OBFUSCATION-AS-A-SERVICE FOR MALICIOUS ANDROID APPLICATIONS.