*Machine Learning Classifier for Mobile Malware Detection*

Dissertation Chapter # 3: Methodology

Student Name: Harjinder

Table of Contents

[Chapter # 3 Methodology 3](#_Toc95568732)

[Introduction 3](#_Toc95568733)

[Research Philosophy 3](#_Toc95568734)

[Data Interpretation 4](#_Toc95568735)

[Research Strategy 5](#_Toc95568736)

[Naïve Bays Classifier 5](#_Toc95568737)

[Support Vector Machine & Random Forest Classifier 6](#_Toc95568738)

[Decision Tree Classifier 7](#_Toc95568739)

[Data collection Method 8](#_Toc95568740)

[Primary Dataset: 8](#_Toc95568741)

[Secondary Dataset: 8](#_Toc95568742)

[Data Analysis Method 8](#_Toc95568743)

[Research Consequences 9](#_Toc95568744)

[Research Sources 9](#_Toc95568745)

[Ethical Concern 10](#_Toc95568746)

[References 11](#_Toc95568747)

# Chapter # 3 Methodology

## Introduction

Android Mobile phone malware detection and prevention necessary required to secure the user personal asset & records. Various mobile users uses various application from Google play store, even the play store is not secure due to unreliable apps, android user increasing daily , Operating system of android phone is very good & flexible & inexpensive so every type of user is in range to buy them. Google developed in Java programming beside this java kernel written in c language, android emulator is main player of interface in which application exist to be tested before deploying. Latest android 11, and 12 version is much secured and reliable with stable security version, so there is need to secure the android phone from malware threat, adware is common malware which was used as playing card with interactive advertisements & promotions & gifts discounts links. Trojan is very dangerous malware which capture adware, worm and other malware from the mobile operating system. This research is going to develop machine learning classifier by using the dataset to detect and prevent the malware observation using the python programing. The common objectives of malware to destroy cyber security confidentiality and integrity by theft of data, commonly attacked denial of service, mobile hijacking, draining the power flooding the device & exploited from known source. Malware might be damage mobile SMS, Bluetooth attack, phone jail breaking install application without permissions and premium rate attack deliver valuable content to mobile device and spyware attack.

## Research Philosophy

T mobile G1 was first android phone, key based event might be interrupt the user for doing any event on the phone. Android 1.0 released in 2008, the operating system including Google map, HTML browser with Gmail service. Google play store is called Google market in which various Google application stores. In 2009 android 1.1 arrived with details of business reviews including text messages, allow to speaker the phone, show and hide the dial pad. Beside this in 2009 android release 1.5 cupcake first version of Google search screen on display on galaxy phone the ability to upload video on YouTube, automatic screen rotation, third party keyboard support, copy & paste features, web browser and the ability to check the phone also in 2009 Google release android 1.6 Donut version support to use CDMA based network support different screen sizes it included search box to quick switching function from camera to other event including power control WIFI, GPA, Bluetooth functionality , beside this Google releases android 2.0 version Éclair which support text to speech function including multiple accounts, Google maps navigation minor changes in API. Android release Froyo in 2010 which android 2.2 version several new feature, mobile hot spot WIFI, push notification and GINGER Bread in 2010 Nexus phone co developed by Google and Samsung updated UI design extra-large size and video resolution support multi input touch, in 2011 android release honeycomb virtual button has been introduced and Android OREO in 2017 & android 9.0 in 2018 PIE version improved battery life, improve navigation operating security & buttons. Android 11 released in 2020 conversation central in the messages with front of notification including chat bubbles long pressing the power button it proceed to new interface introducing new security features to secure user data , user will allows to access app only once, next time the permission required to access the app from the phone. First malware attack found in android phone Ransomware attack in 2010 which flash player name in android phone. In 2011 uninstalling apps automatically in phone due to malware attack happened due to system update automatically installed update without user permission. In 2020 latest android phone Google introduced built in machine learning classification technique to update the security features in android 11 version.

## Data Interpretation

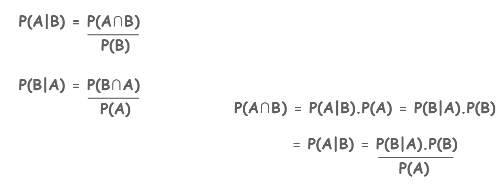
Mobile malware is malicious software that target mobile phone or wireless enabled personal digital assistance by causing the collapse on the operating system and loss or leakage of confidential information, beside this wireless phone PDA network are more complex, safety and security are commonly used to protect the malware attack. Android malware detection if there is presence of malware by using the attributes extracted from android applications as feature. Since the dataset values belong to either 0 or 1 only label encoding of last column will be enough, by training and testing the feature of dataset including train feature and label feature build the Keras sequential model, and predict the model after training and testing the dataset and build the machine learning models to get the F1 score and precision recall value, confusion matrix build to inform the predicted label value and true label value from the dataset. At kaggle competition dataset attribute name transact, on service, bind service, attach interface, service connection, android operating system binder, SENS SMS etc. and dataset at Github the attribute selection android permission access all downloads, android permission access Bluetooth share, android permission access on cache file system, android permission access on cache properties. The neural network model in static model has a bias towards bengin. The dataset name is android malware detection by analyzing permission on android phone. Various applications needs permission before launching, it includes jupyter notebook to prepared the dataset and create the model, dataset created using androgurad. Retrained model using the genetic algorithm techniques.

## Research Strategy

Research strategies of this research used with advance machine learning technique to predict the android phone permission dataset. Classification of android apps done by using the pseudo dynamic analysis of system API calls sequences, developed on deep learning & convolutional neural network classifier including LSTM model deep neural network classify of naïve bays classifier method. Machine learning classification uses the supervised machine learning models.

### Naïve Bays Classifier

Genetic global dictionary for storing mapping all distinct API calls to numbers in the dataset and pickles the dictionary, extract all features vectors for 8 test cases in dataset and pickles them, extract all features vectors for 600 apps in dataset and various features, and build the naïve bays classifier method. Naïve bays classifier of probability of A predict the value of probability of B value extract the featured values on dataset based on x1, x2 and x3 based value. Given hypothesis H and evidence E bays theorem states that relationship between the probability of the hypothesis before getting the evidence P(H) and the probability of the hypothesis after getting the evidence of P(H|E) is under estimation.



*Figure 1: Probability distribution*

In machine learning classification technique the frequency measures on likelihood values based on the bays theorem value. The fancy of naïve bays classify the target variable attributes based on the classification techniques of algorithm which founded on interesting attribute of dataset value build the machine learning model after feature selection from the given dataset and training and testing the dataset and builds the machine learning to check the precision call and F1 score of model. Naïve bays classification method for text classification implemented through the NaiveBaysText python class.

### Support Vector Machine & Random Forest Classifier

Machine learning classifier has become the great tool to classify the malware from the android phone, in these days everyone is part of android phone and thousands of application has been customized and installed on android phone daily, beside this the rapid growth of malware designed for mobile devices. Android malware detection has increasingly important in the cyber security domain, using the standard machine learning classifier to classify and detect malware by using the dataset from kaggle and Github repository. ML classifier automatic detect the malware after feature selection and distinguish between the malware families and automate the detection and prevention process after training and testing the feature.

Exploring the malware dataset in python notebook and perform the data cleaning process and splitting the dataset into test and train and after feature selection building the model by using the random forest classifier and support vector machine classifier importing the python libraries, by using sklearn python library. Random forest evaluation on given dataset importing the confusion matrix and target train and test variable selection and predict the models. Building the confusion matrix with F1 score to validate the normalization process. Support vector machine classifier predict the values on given feature engineering selection in python notebook to predict the security permission on android phone attributes values given on building the dataset values on demands values , the model predict the highest accuracy rates with good precision call value on the machine learning prediction and classification technique method. Support vector generate the decision boundary. The flow method of machine learning as follows in the given diagram as below:

*Figure 2: Flow Diagram*

### Decision Tree Classifier

Artificial intelligence method seen everywhere in our daily lives, the technology is become more powerful and interactive it is possible to simulate AI based cyber security attacks which performed using advanced malware incorporate to advanced feature evasion technique to avoid the security parameter in order to detect the extreme data security permission on android phone the objectives of malware detection and prevention system is to classify and obtain the correlation based feature selection by using the artificial neural network which compared with existing model and improved version to get the highest prediction rate, data processing and feature engineering selection in python to classify the dataset attribute values and perform data cleaning process to get the missing value and present the data feature selection to build the model in training and testing the feature selection of the model to validate the dataset values and build the models, finding the sensitive permission and taken to identify the malicious one. Malware is designed on particular malicious software to distort and interrupt the mobile or computer applications. Google play store is largest application store in the world, with commanding over 88% Android mobile shares in the market. Random forest is the core idea to generate multiple decision tree based on the feature selection of dataset.

## Data collection Method

### Primary Dataset:

Primary dataset has been collected from Github and Kaggle repository which is used to predict the malware security permission attribute values on android phone. Dataset attribute selection has been made from both kaggle and Github repository which means to validate the feature selection attribute.

### Secondary Dataset:

Secondary dataset has been collected from Google scholars, IEEE Journals, IEEE Xplore, IEEE conference paper, ACM library and MDPI research including peer review journals article with good impact factor. Dataset was part of other research which does not means to copy from other sources, it get the idea and generate the research plot.

## Data Analysis Method

Static and dynamic malware analysis (Chapaneri, R., 2020) using the machine learning classification technique as mentioned. Malware is section of programing written in intention of harming the mobile device and capture data and locations. Data analytics techniques based on identify the patter in dataset and demonstrate the static and dynamic analysis which learns the machine learning model various machine learning algorithm used to classify the dataset and obtain the F1 score and precision recall score with highest accuracy rate, various machine learning method used such as random forest decision tree classifier, support vector machine and naïve bays classifier including neural network modeling to get the highest accuracies about 98% implemented the dynamic analysis. (Ali, W. Abdulghafor, R., 2020) mobile malware detection & prevention based on supervised machine learning modeling after training and testing the feature engineering dataset value, android zero app cannot be detected which formulated on signature based analysis due to learn the existing feature and build the new feature model to detect the malware values from the dataset by using random forest decision tree and SVM, including naïve bays classifier which are common to detect. In previous releases of android operating system which does not support security features, now the 11 version of android is stable and secured from third party hacking & malware activity.

## Research Consequences

Research result formulated on the machine learning classification technique detection dataset columns values which based on android permission resulted in column, certain columns in dataset based on same value for APK files and analyzed information predicted on cleaning the value, removing null values building feature engineering, removing highly correlated feature engineering and testing and train the dataset to select the target variable from the dataset and build the machine learning model. Distribution of malicious and bengin file in the dataset present the graphical view of the permission attribute value the dataset contains the balanced samples of malicious attribute value achieve through the random sampling method. Since the android applications represented from Google application play store the android API version improved on latest releases of android operating system, SVM classifier good classifier of machine learning tool beside this naïve bays classifier analyze them on text feature selection and takes decision on the selection attribute values which founded on binary classification which mean 0 and 1 technique. Model validation and tuning towards the end of the model development, the research used F1 score to generate matrix which helps the research to evaluate the model performance. The validation model selection constructed on by training and testing the dataset target variable values & feature engineering obtain the data cleaning process and build the machine learning models.

## Research Sources

Research sources as follows

1. Github
2. Kaggle
3. IEEE
4. Google Scholars
5. ACM Digital library
6. MDPI Research
7. Google Scholars
8. Anaconda Python
9. Jupyter notebook
10. Machine learning & deep learning python libraries.

## Ethical Concern

This research studies of universal values based on mobile phone which is essential equality of both men and women which is natural rights to use the mobile device for both gender to solves their daily lives problem and business & education concern. The right behavior of this research to used machine learning modeling to avoid malware and remove them from the phone device and wrong behavior of doing the malware activities such as download software’s from illegal sites, Google play stores is reliable platform for downloading the software and applications from the Google server. Ethical behavior confirms to generally accept social norms, ethical codes difficult in some situation to enable the mobile malware without any reason for deploying the anti-malware tool in mobile device. This research ethics contribute to society and mankind avoid harm to others be honest and trustworthy be fair and don’t discriminate honor property rights including mobile data sharing rights Bluetooth sharing rights and application sharing permission is mandatory before using the other user mobile device. This research conduct in good faith including social moral values, it meets the scientific & artificial intelligence standards the results describes the guidance to society members and culture to avoid unreliable source and data sharing activity, because mobile phone contains sensitive data & banking information. Mobile subject treat well, reduce coercive and reward response, honestly I saying not lying stealing or cheating the other research works not hiding the truth for someone and showing the good and truthful character.

# References

Raghuraman, C., Suresh, S., Shivshankar, S. and Chapaneri, R., 2020. Static and dynamic malware analysis using machine learning. In *First International Conference on Sustainable Technologies for Computational Intelligence* (pp. 793-806). Springer, Singapore.

Abdullah, T.A., Ali, W. and Abdulghafor, R., 2020. Empirical study on intelligent Android malware detection based on supervised machine learning. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, *11*(4).

Rana, M.S., Gudla, C. and Sung, A.H., 2018, December. Evaluating machine learning models for Android malware detection: A comparison study. In *Proceedings of the 2018 VII International Conference on Network, Communication and Computing* (pp. 17-21).

Kuruvila, A.P., Kundu, S. and Basu, K., 2020, July. Analyzing the efficiency of machine learning classifiers in hardware-based malware detectors. In *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (pp. 452-457). IEEE.

Gyunka, B.A., Abikoye, O.C. and Adekunle, A.S., 2020, November. Anomaly Android Malware Detection: A Comparative Analysis of Six Classifiers. In *International Conference on Information and Communication Technology and Applications* (pp. 145-157). Springer, Cham.

Osman, M.Z., Abidin, A.F.Z., Romli, R.N. and Darmawan, M.F., 2021, August. Pixel-based Feature for Android Malware Family Classification using Machine Learning Algorithms. In *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)* (pp. 552-555). IEEE.

Anshori, M., Mar'i, F. and Bachtiar, F.A., 2019, September. Comparison of machine learning methods for android malicious software classification based on system call. In *2019 International Conference on Sustainable Information Engineering and Technology (SIET)* (pp. 343-348). IEEE.

https://www.kaggle.com/vishnu0399/android-malware

<https://github.com/rozer821/Android_Malware_classifier>

<https://github.com/anoopmsivadas/android-malware-detection>

## 