NEW YORK INSTITUTE OF TECHNOLOGY

DTSC 870 - M01

Project Proposal

Professor Wenjia Li

Team Members

Zoya Haq	1222440	zhaq01@nyit.edu
Neelam Boywah	1226855	nboywah@nyit.edu
Selina Narain	1261565	snarai01@nyit.edu

Table of Contents

Abstract

Android has recently risen to become one of the most popular mobile operating systems, as it is able to support a vast amount of mobile applications. However, harmful Android malicious apps downloaded have and potentially can endanger users' security and privacy. The majority of them go unnoticed because there are not any reliable or accurate malware detection methods. In this research study, we propose to examine a malware detection method for the Android platform using data science methods and a base model using existing data. This study aims to achieve numerous objectives such as developing an accurate model and analyzing the detection model with hopes of resulting in malware identification and detection quickly and accurately within malicious Android applications.

Project Title: Malicious Mobile Applications

Problem Statement

In the modern day, there has been an increase in malware across mobile applications, posing a threat to the security and privacy of mobile users. This research study aims to develop a technique for malware detection and analysis across mobile applications, with the goal of mitigating posed risks and protecting mobile users' data.

Research Scope

The scope of this research project will cover the identification, detection, and analysis of malware in mobile Android applications. Based on the current dataset found, this includes analysis of adware, banking malware, SMS malware, riskware, and benign.

Hypothesis

This research study seeks to test the hypothesis that machine learning algorithms, trained on a large Android malware dataset containing both legitimate and malicious mobile applications, can detect malicious applications across mobile devices.

Objectives:

- Develop an accurate model based on data science algorithms that can effectively identify and classify malicious mobile applications. The models should be able to attain a high accuracy and precision within training data to predict onto testing data in order to detect benign or malicious apps.
- Identify and analyze patterns and characteristics of malicious mobile applications to understand their potential impact on user privacy and security. This analysis will provide insight into the methodology and trends of attackers' behaviors.
- Evaluate the results of the model to detect whether the accuracy is enough to be applied in real-world scenarios of malware detection.

Significance of the Study

Motivation

The widespread adoption of mobile applications have attracted those with malicious intent that are seeking to exploit vulnerabilities within these enterprises to ultimately compromise user data and privacy. As the number of mobile users continues to rise, the threat to malicious cyber acts increases with it.

Principal Literatures

To delve into a deeper understanding of Android, our team analyzed a few journal articles that discuss detection of Android malware on mobile devices. *DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket* (Arp et al., 2014), explains the importance of mobile malware, especially in the modern landscape. It also discusses some features that are notable in detecting Android's malware. The article mentions that every Android application developed, requires a manifest (AndroidManifest.xml), providing data for installation and execution. Based on the execution, there are a few features that would be useful in detecting malware such as hardware components, requested permissions, application components (activities, services, content providers and broadcast receivers), and filtered intents (allows information regarding events to be shared between different components and applications). This is crucial when looking at overall datasets in order to understand what behaviors and features to identify prior to developing the detection model.

Furthermore, *DroidEnemy: Battling adversarial example attacks for Android malware detection* (Bala et al., 2020), focuses further on the increasing threat of mobile malware targeting Android-based devices and therefore needing to have effective detection mechanisms to combat these attacks. In addition, it was useful to note that many attackers evade detection, which involves tampering with mobile applications, similar to what was found from the previous publication. One interesting note stems from the data poisoning attack, which is used to weaken malware detectors by providing the adversary with access to the learning algorithm's training dataset. This attack involves employing four strategies: label modification, data modification, logic corruption, and data injection, enabling the attacker to manipulate the training data and compromise the detection system's effectiveness. There are additional attacks where combined

with reverse engineering can provide attackers with a surplus of sensitive or restricted data. From the research, experimental results show that the performance of Android malware detection significantly deteriorates when faced with hostile malicious attacks.

Importance

This research is valuable as it would be crucial to users who are utilizing technology in their day to day lives so they are cautious and careful when they use certain applications. These malicious applications can potentially steal a users' personal information and identity. With this study, we can detect and ensure that malicious applications are being caught and flagged to avoid users being harmed in the future.

Methodology

Research Design

This research will be comprised of the following activities necessary to achieve the objectives:

1. Detection Practices: Research existing detection and reactive methods used by companies such as Google or Apple to identify malware in mobile applications.
2. Data Collection: Identification of datasets that are most appropriate to be used for this study, including data on the various attributes of mobile malware attacks.
3. Data Analysis: Identifying trends within the dataset to provide insight into the methodology and trends of attackers' behaviors.
4. Preprocessing Data: Cleaning and preparing the data and performing feature extraction to ensure all data is relevant to the research scope.
5. Feature Selection: Identifying and selecting the most appropriate features in the dataset to be utilized by the machine learning algorithm.
6. Machine Learning Model Development: Identifying and applying the most appropriate machine learning algorithm to be used and running the model against the data.
7. Model Evaluation: Analyzing the model using metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) to assess the effectiveness of the model.
8. Analysis: Analyze the findings in malware detection and interpret the results to draw conclusions.
9. Mitigation Measures: Evaluating the best practices to ensure prevention of installation and execution of malicious apps.
10. Conclusions: Develop overall conclusions on the effectiveness and outcomes of the research study, identifying the most prominent findings as well as strengths and weaknesses.

Tools and Technologies

Various tools and technologies will be used to conduct our research and test malicious applications. These include Visual Studio Code as the IDE. There will also be use of Jupyter Notebook for programming language Python, to support data science methods. This includes packages such as Pandas, NumPy, Scikit-learn, TensorFlow, and Keras which support execution

of data science algorithms and analytics. Through the use of these tools, we can analyze and parse through large datasets obtained throughout the research. In addition, we will also be using visualization tools such as the Matplotlib and Seaborn packages for Python.

Important Concepts

Adware: type of malware program attack that utilizes advertisements to gather and steal user's sensitive and personal information.

Deep Learning: a branch of machine learning that trains neural networks to perform complex tasks. These networks can recognize patterns and gestures in data resulting in accurate predictions.

Malware Detection: the process of identifying malware (malicious software) on a computer network or system.

Machine Learning: a type of artificial intelligence tool that allows computers to recognize patterns and improve performance without the use of humans.

Malicious Applications or Intrusive Software: programs designed to harm and exploit computer networks and infrastructure. Some examples include viruses, spyware, Trojans, ransomware.

Ransomware / Spyware: types of malware attacks designed to gather information about a user and send that data to cybercriminals where they can then use that information to harm and steal from users.

Viruses (type of malware attack): designed to infect and replicate over computer systems which leads to harming a systems overall functionality.

Required Resources

In order to successfully complete this study, datasets from Kaggle will be utilized. From a preliminary review, we are able to identify two datasets we are interested in using:

1. Android Malware Dataset for Machine Learning: This dataset provides relevant features used in detecting malware such as API call signatures and Manifest Permissions.
2. Investigation of the Android Malware (CIC-InvesAndMal2019): This dataset includes samples, identified as malware and benign. This is further classified into adware, ransomware, scareware, SMS Malware.

From these datasets, we will be able to use the tools and technologies previously discussed to apply machine learning algorithms to create a model for malware detection.

Citations

Arp, Daniel, et al. *DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket*, Feb. 2014, www.researchgate.net/publication/264785935_DREBIN_Effective_and_Explainable_Detection_of_Android_Malware_in_Your_Pocket.

Bala, Neha, et al. "DroidEnemy: Battling adversarial example attacks for Android malware detection." *sciencedirect.com*, ScienceDirect, 21 September 2023, https://doi.org/10.1016/j.dcan.2021.11.001.

"Investigation of the Android Malware (CIC-InvesAndMal2019)." *University of New Brunswick Est.1785*, www.unb.ca/cic/datasets/invesandmal2019.html.

Tiwari, Shashwat. "Android Malware Dataset for Machine Learning." *Kaggle*, 13 Mar. 2021, www.kaggle.com/datasets/shashwatwork/android-malware-dataset-for-machine-learning.