Selina Narain, Neelam Boywah, Zoya Haq
DTSC 870 - Masters Project - Fall 2023
Advisor: Professor Dr. Wenjia Li

Progress Report 2

Timeline: October 4, 2023 - October 18, 2023

Accomplishments: What did you accomplish?

Research Topic Idea:

- Comparing machine learning algorithms for accuracy and efficiency in detecting malware in android applications.

Approach: Using Machine Learning and Deep Learning Techniques

- SVM using feature extraction and selection (FEST). Kernels (Non-linear, linear, and RBF)
- Random Forest
- Naive Bayes
- KNN
- Logistic Regression
- Possibly: Deep Neural Network (DNN) - or Convolutional Neural Network (CNN)

Dataset:

Android Malware Dataset for Machine Learning

- Includes: drebin-215-dataset-5560malware-9476-benign.csv and dataset-features-categories.csv
- This dataset provides relevant features used in detecting malware such as API call signatures and manifest permissions.
- Consists of feature vectors of 215 attributes extracted from 15,036 applications. 5,560 malware apps from Drebin project and 9,476 benign apps.

Android Malware Analysis

- Includes MalGenome - Androme Genome Project Malware
    - The set of malware has a size of 1260 applications, grouped into a total of 49 families.
- Dataset with Drebin aspects
    - Total of 5560 applications consisting of 179 malware families
- AndroZoo
    - 5669661 applications Android from different sources (including Google Play)

- VirusShare
- DroidCollector
    - 8000 benign applications and 5560 malware samples

Tools & Technologies:

- Programming Language: Python
- Google Collaboratory
- Jupyter Notebook
- Visual Studio Code
- Scikit-learn Package
- Visualizations
    - Power BI
    - Microsoft Excel (.csv file for Power BI dashboard report)

Upcoming Plan: What do you plan to do in upcoming weeks?

- Continuing any necessary research
- Exploratory data analysis to better understand datasets
- Setting up our Data Science Environments to start implementation and testing for machine learning models.
- Further research into Adversarial attacks

Obstacles: Were there any obstacles or barriers that prevented you from getting things done?

After going through our findings from our research, we were able to start formulating our topic idea for the research paper as well as determining which machine learning and deep learning algorithms we will be using based on our literature review. Therefore, there were not any significant obstacles.