Malware Detection in Android Smartphones

REU Fellow: Neha Bala
REU Mentor: Professor Wenjia Li, Ph.D.
Email: nbala02@nyit.edu, wli20@nyit.edu

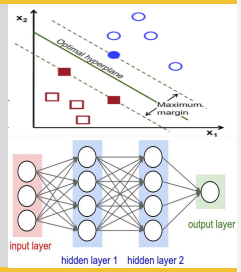
ABSTRACT

Due to their popularity, Android smartphone are vulnerable to mobile malware. Mobile malware gains access to a device by tricking users into installing malicious applications. Researchers have taken different machine learning based approaches to detect malware. In this project, we build an existing system, however, instead of using a traditional machine learning algorithm, we implement a deep learning based model to get results with a larger dataset.

BACKGROUND

SigPID System:
- Performs three levels of data pruning
- Uses Support Vector Machine & Decision Tree

Deep Learning Model:
- Popular subset of machine learning
- Performs best with a larger dataset

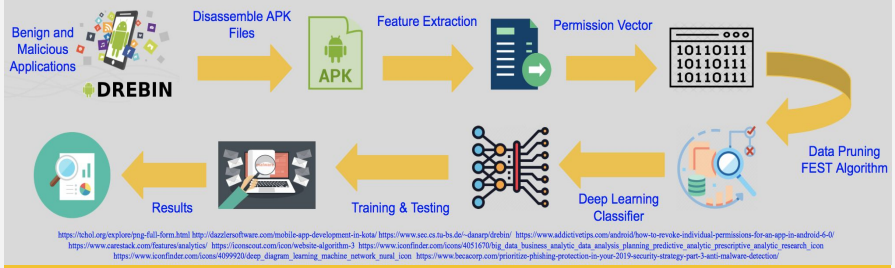


OBJECTIVES

Overall Objective:
Research malware detection in smartphones specifically in Android devices

Specific Objective:
Perform data analysis using deep learning based approach to determine malicious applications.

ARCHITECTURAL OVERVIEW



RESULTS

Fig. 1 Original Data Set Results

Correctly Classified Instances	95.3216%
Incorrectly Classified Instances	4.6784%
Kappa statistic	0.9064
Mean absolute error	0.0703
Root mean squared error	0.2076
Relative absolute error	14.0517%
Root relative squared error	41.5108%

Fig. 2. Significant Features By Data Pruning

Feature	Feature Number
android.permission.read_phone_state	2
android.permission.access_network_state	3
android.permission.send_sms	5
android.permission.receive_boot_completed	6
android.permission.wake_lock	8
android.permission.access_coarse_location	11
android.permission.vibrate	13
com.android.browser.permission.read_history_bookmarks	18
com.google.android.c2dm.permission.receive	44
android.permission.read_external_storage	46

Fig. 3 Post Data Pruning Results

Correctly Classified Instances	93.6047%
Incorrectly Classified Instances	6.3953%
Kappa statistic	0.8721
Mean absolute error	0.0903
Root mean squared error	0.2373
Relative absolute error	18.0556%
Root relative squared error	47.4591%

Fig 4. Correlation Between Features

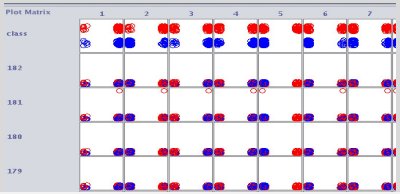
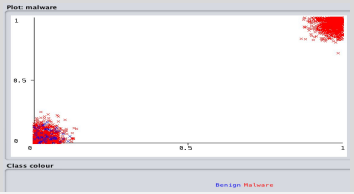


Fig 5. Wake_Lock Permission



DISCUSSION & CONCLUSION

- The following results were based off of permission features. Due to a delay in the data training process, other features such as API calls and URL features were not used to their full extent. If the dataset of features was larger, the deep learning based results could have been improved.
- Implementing the SigPID with a deep learning based classifier, the results indicate that this approach is effective in detecting malware with a larger data set. Without data pruning the classifier detected 95.3216% of malware applications with 4.6784% of false detection. With data pruning the classifier detected 93.6047% of malware applications and 6.3953% of false detection.

REFERENCES

[1] Sun, L., Li, Z., Yan, Q., Srisa-An, W., & Pan, Y. (2016). SigPID: Significant permission identification for android malware detection. *2016 11th International Conference on Malicious and Unwanted Software (MALWARE)*.
[2] Sun, L., Li, Z., Yan, Q., Srisa-An, W., & Pan, Y. (2018). SigPID: Significant permission identification for Machine-Learning-Based Android Malware Detection. Retrieved from https://ieeexplore.ieee.org/abstract/document/8255798
[3] "Support Vector Machine - Introduction to Machine Learning Algorithms." *Medium*, Towards Data Science, towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.
[4] "Challenges in Deep Learning." *By*, hackernoon.com/challenges-in-deep-learning-57bbf6e73bb.

ACKNOWLEDGEMENT

This project is funded by National Science Foundation Grant No. 1852316 and by New York Institute of Technology.