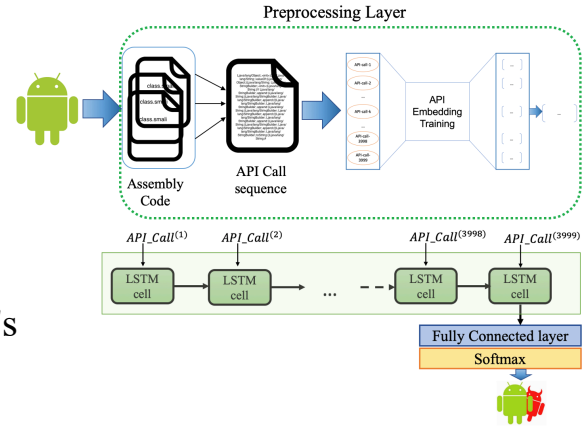
Data-driven Approaches for Large-scale Security Analysis of Mobile Applications

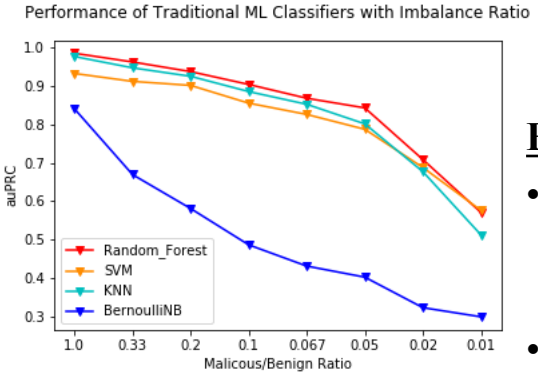
Challenge:

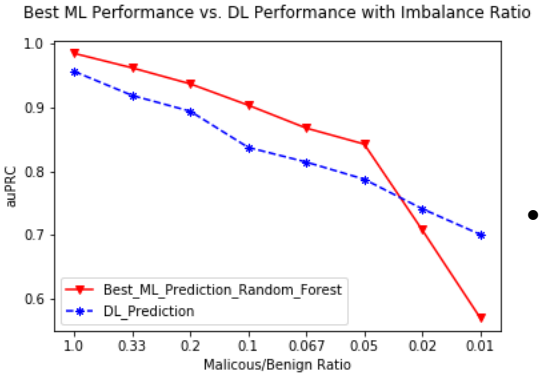
- Huge number of apps, highly unbalanced data -- only a small fraction of apps are malicious
- Evolving app features make distinguishing malware a moving target
- Noisy ground truth regarding an app's maliciousness status

Solution:

- Build a large dataset that reflects the real word app stores like Google Play
- Label apps by extensively leveraging off-the-shelf resources like VirusTotal
- Explore deep learning approaches to address the evolving feature challenge
- Compare traditional machine learning and deep learning approaches with respect to their ability to handle imbalanced data and noisy ground truth







Scientific Impact:

- Our experiments show that on highly-imbalanced data deep learning classifiers work better than traditional machine learning.
- Hybrid deep learning models that combine static and dynamic artifacts work better than models trained with either static or dynamic artifacts alone.

Broader Impact:

- This research helps improve the technology for scaling the security vetting process of mobile apps.
- Three graduate students (two PhD students and one MS student) have been supported by this grant, and other MS students have been involved in the project.
- Data and approaches designed in this project have been used in classes taught by the PIs at their respective institutions.

- **Awards:** #1717862,#1717871,#1718214, 8/15/2017-7/31/2020.
- **PIs:** Xinming Ou (USF), xou@usf.edu
Doina Caragea (K-State), dcaragea@ksu.edu
Sankardas Roy (BGSU) , sanroy@bgsu.edu