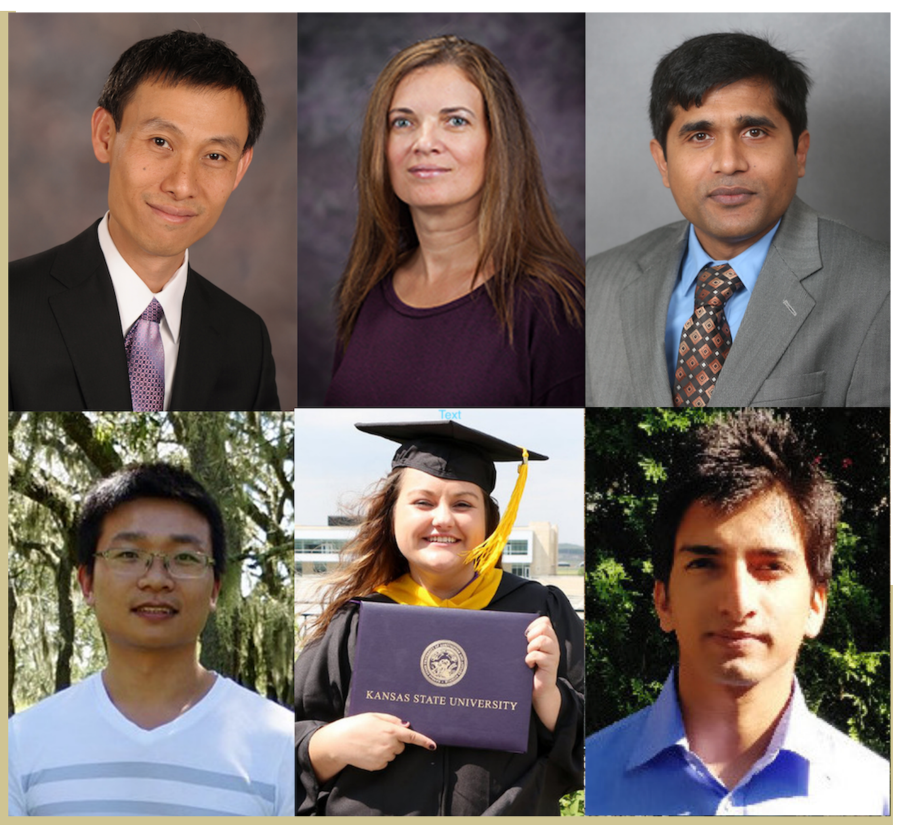
Comparing Traditional Machine Learning and Deep Learning Approaches for Security Vetting of Android Apps

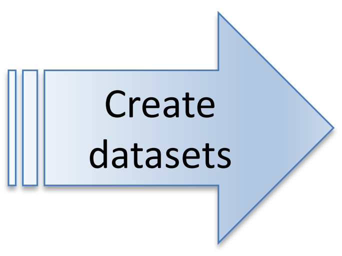


PIs: Xinming Ou (USF), Doina Caragea (K-State), Sankardas Roy (BGSU)
Students: Guojun Liu (USF), Emily Alfs (K-State), Dewan Chaulagain (BGSU)

Award #1717862,#1717871,#1718214 SaTC: CORE: Small: Collaborative: Data-driven Approaches for Large-scale Security Analysis of Mobile Applications. $200K, $200K, $100K, 8/15/2017-7/31/2020.

Datasets – used for both Deep Learning and traditional Machine Learning experiments

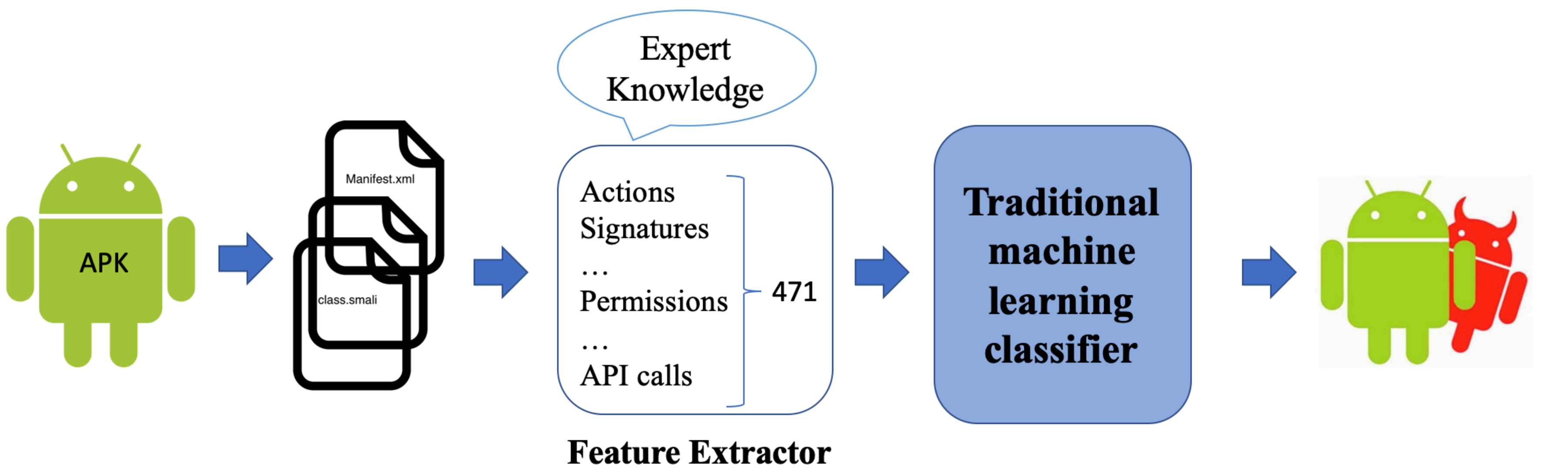
- Generated VirusTotal reports for 1,456,350 apps released between 2016 and 2018
- Generated VirusTotal reports for 339,853 apps released between 2018 and 2019
- App scanning using VirusTotal lasted one year and a half



- AMD malware dataset (2010 – 2016): 24,553
- Newer benign (after 2016): 370,701
- Newer malicious (after 2016): 24,868

Traditional Machine Learning (ML) Based Vetting System

- The ML vetting system in our study uses apk features to classify benign and malicious apps
- Specifically, each app is represented using 471 binary features which represent permissions, intent actions, discriminative APIs, obfuscation signatures, and native code signatures

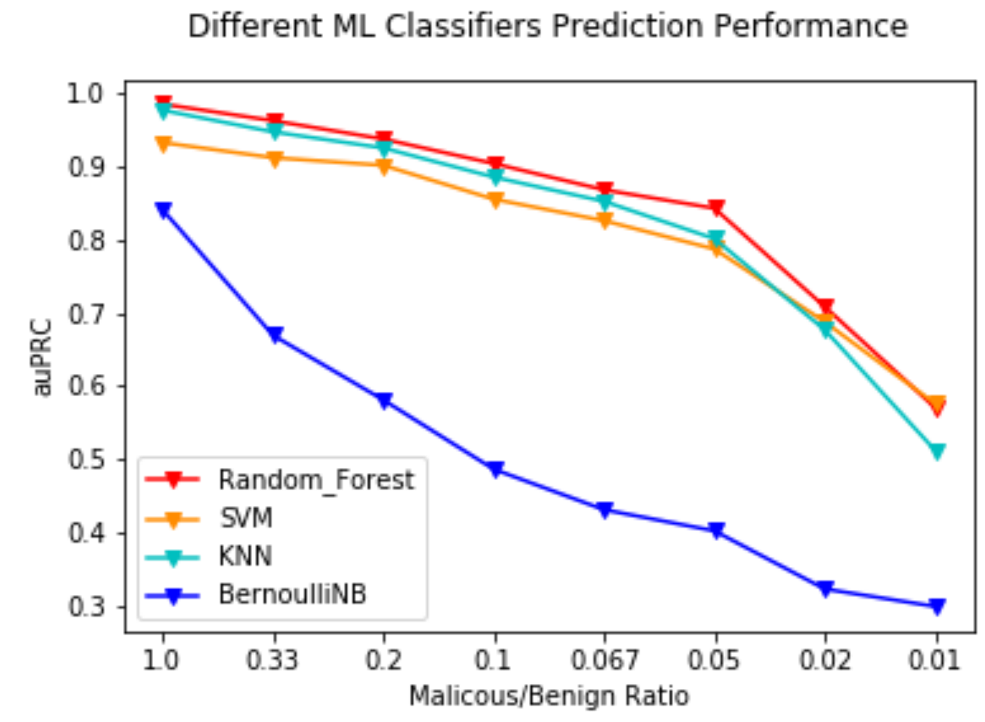


Main Challenges

- Feature engineering has to keep up with evolving app trends
- More training does not always lead to better performance

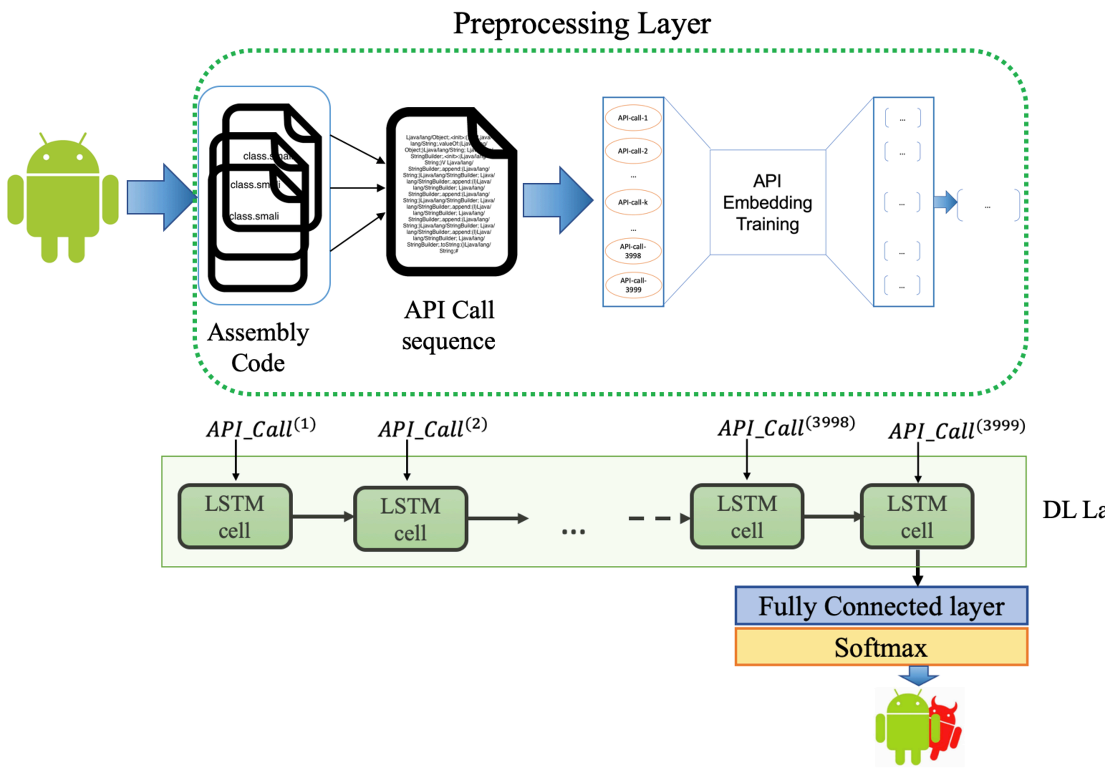
Experiments with Traditional ML Classifiers

- We experiment with datasets that exhibit realistic malicious-to-benign ratios (e.g., smaller than 0.05)
- We use the area under the precision-recall curve (auPRC) to evaluate classifiers' performance
- Experimented with Bernoulli Naïve Bayes, k-Nearest Neighbors, Support Vector Machines, and Random Forest classifiers
- K-Nearest Neighbors and Support Vector Machines take much longer time (**days**)
- **The performance of traditional ML classifiers degrades for highly unbalanced data**



Deep Learning (DL) Based Vetting System

Overview of DL Vetting System

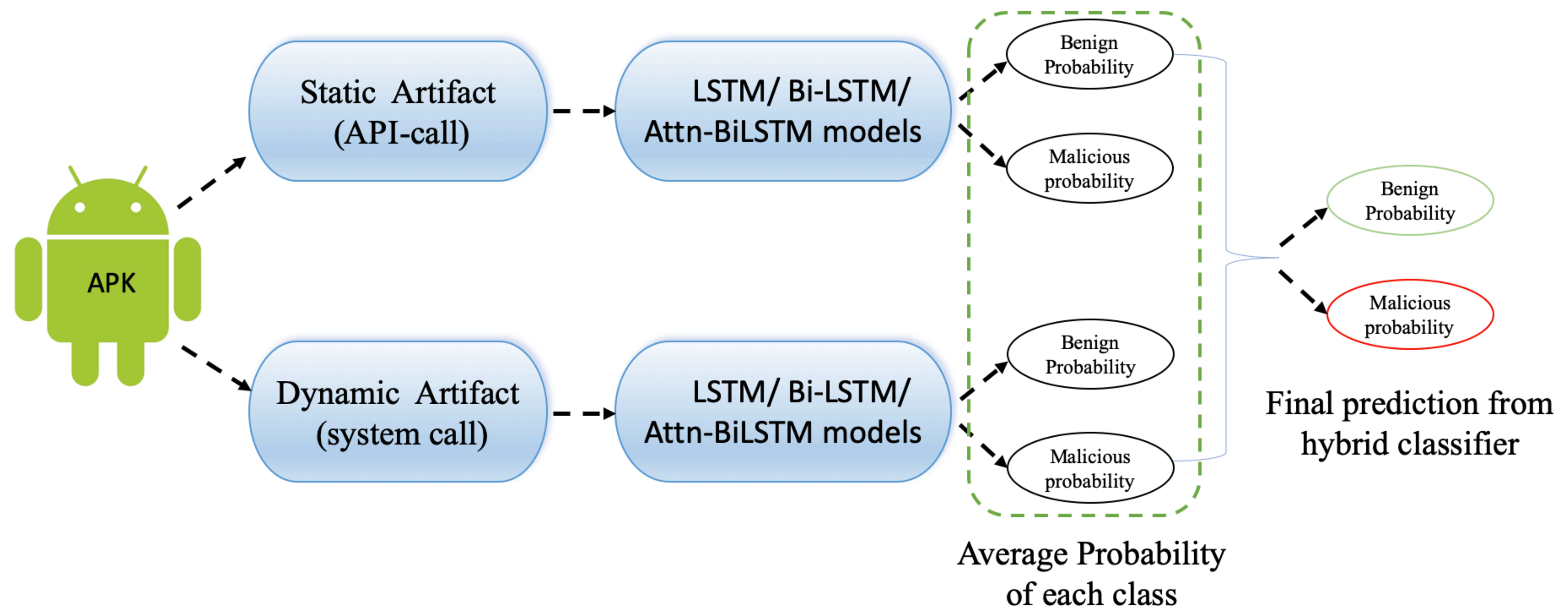


- For each app, it feeds the corresponding raw apks into the preprocessing layer, and generates an API call sequence
- Applies different embedding techniques (e.g.,Word2vec) to generate embeddings for API calls (regarded as "words")
- Each app, represented as a sequence of (max) 4000 API calls using the API call embeddings, is fed into a Long Short-Term Memory network (LSTM)

Benefits & Challenges

- The ability of DL approaches to automatically identify predictive features could benefit mobile app vetting systems
- Efficiently applying DL for large-scale malware detection comes with its own challenges

Predictive Power of Static/Dynamic Artifacts



DL versus Traditional ML Results

- Both traditional ML and DL classification models have good performance on balanced data
- Performance of both traditional ML and DL models decreases on unbalanced data
- DL model has better performance on highly unbalanced data