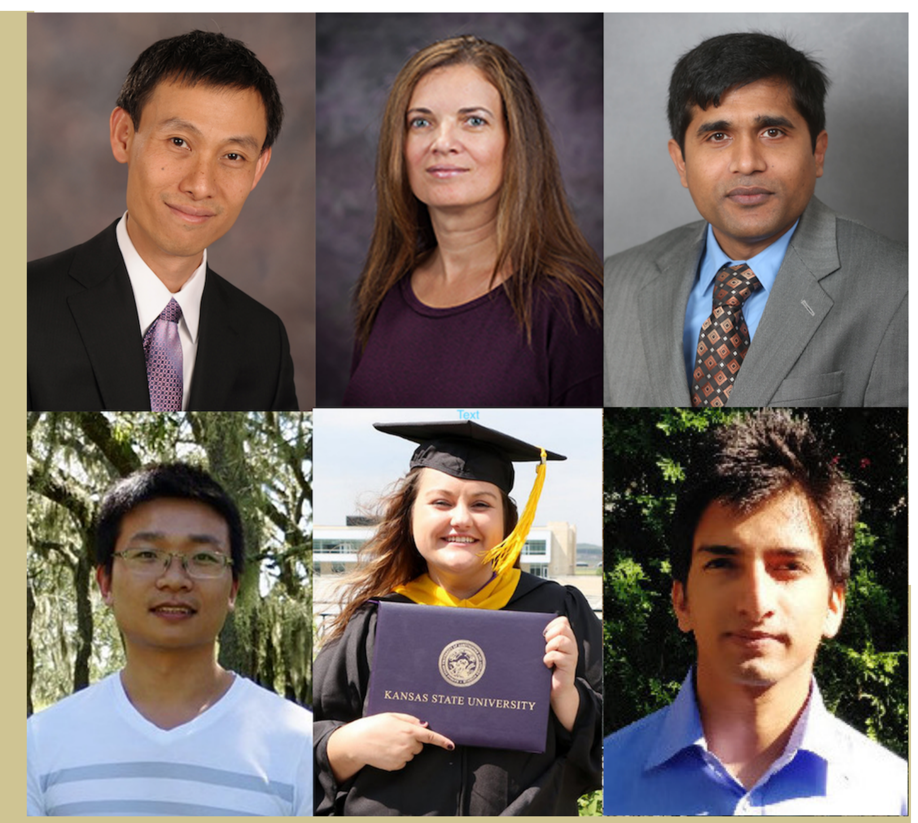
A Systematic Study of Comparing Traditional Machine Learning and Deep Learning for Security Vetting of Android Apps



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 – for both DL and classical ML experiments

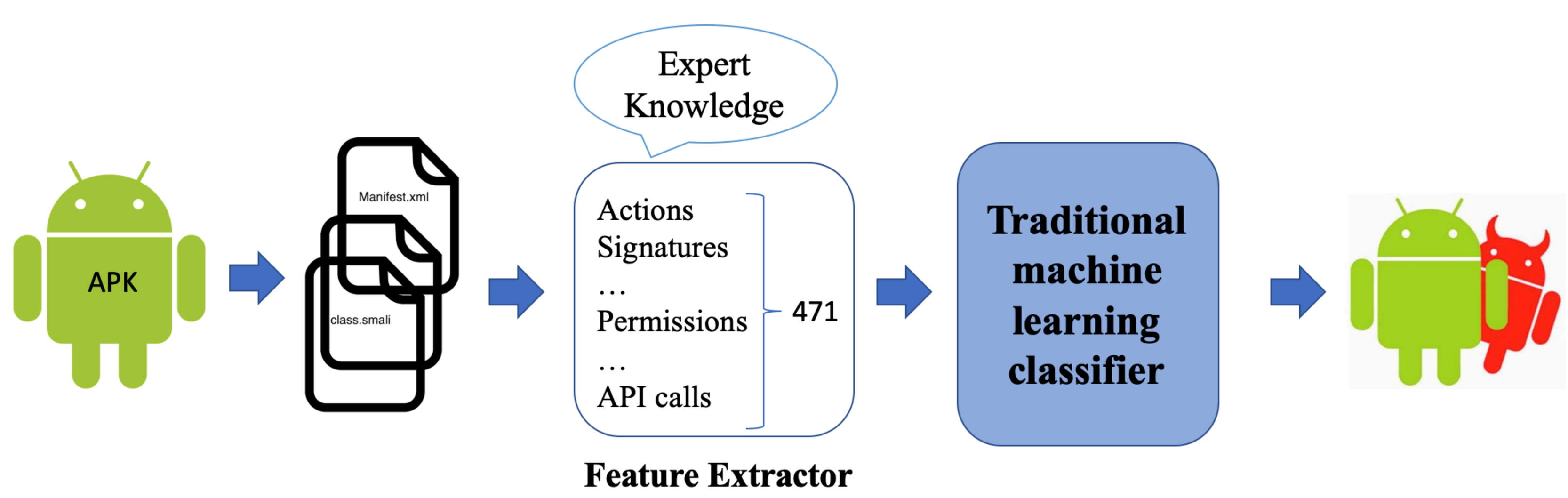
- Data collection lasted one and half years
- Labeled 1,456,350 apps released between 2016 and 2018
- Labeled 339,853 apps between 2018 and 2019

Create datasets

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

Traditional ML Based Vetting System

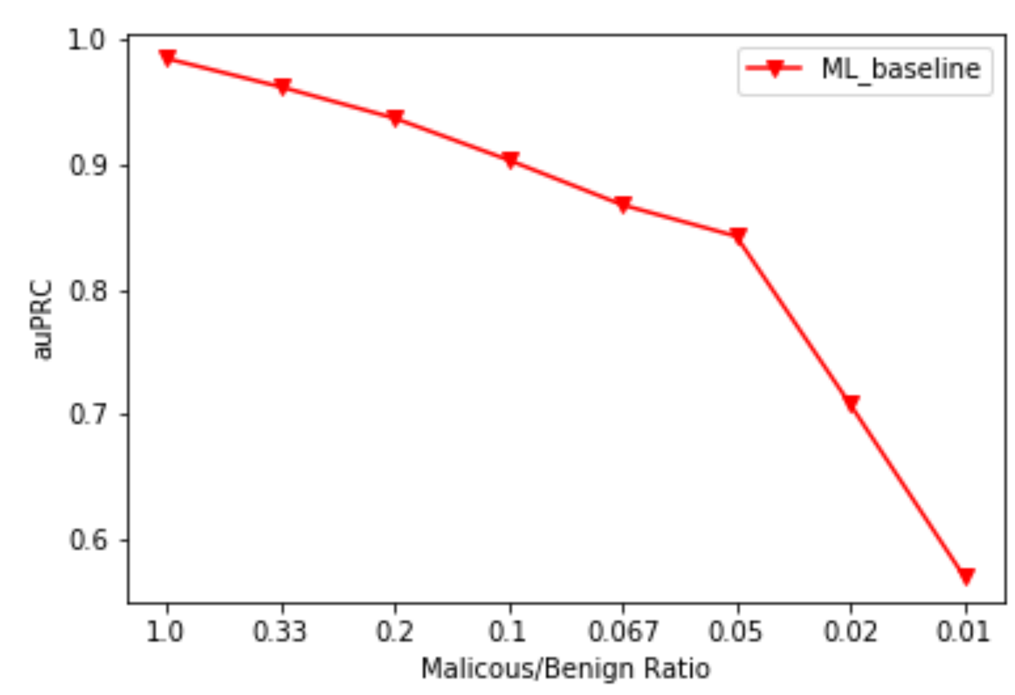
- Uses specific apk features to classify benign and malicious apps
- The ML system used in our experiment is based on 471 features extracted from permissions, intent actions, discriminative APIs, obfuscation signatures, and native code signatures



Main Challenges

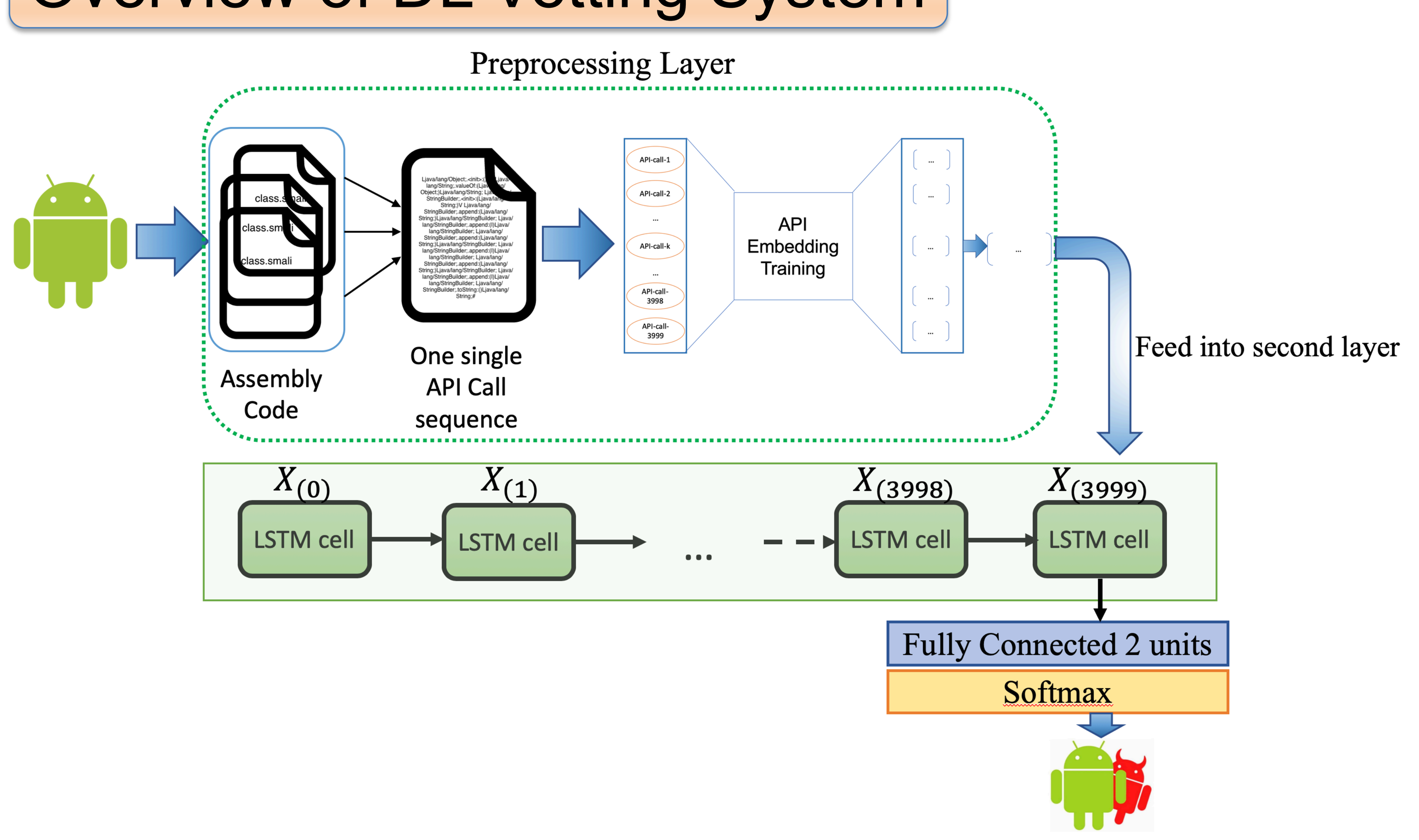
- Feature engineering has to keep up with evolving app trends
- Feature extractor has to keep up with changing app format

- We built our datasets with real-life malicious:benign ratio (less than 0.05)
- We use the area under the precision-recall curve (auPRC) to evaluate the classifier's performance for real-world application
- Experimented with Bernoulli Naïve Bayes, k-nearest neighbors, support vector machines, and random forest classifiers
- Traditional ML model meets challenges on highly unbalanced dataset



DL Shows Advantage over Traditional ML for Highly Unbalanced Data

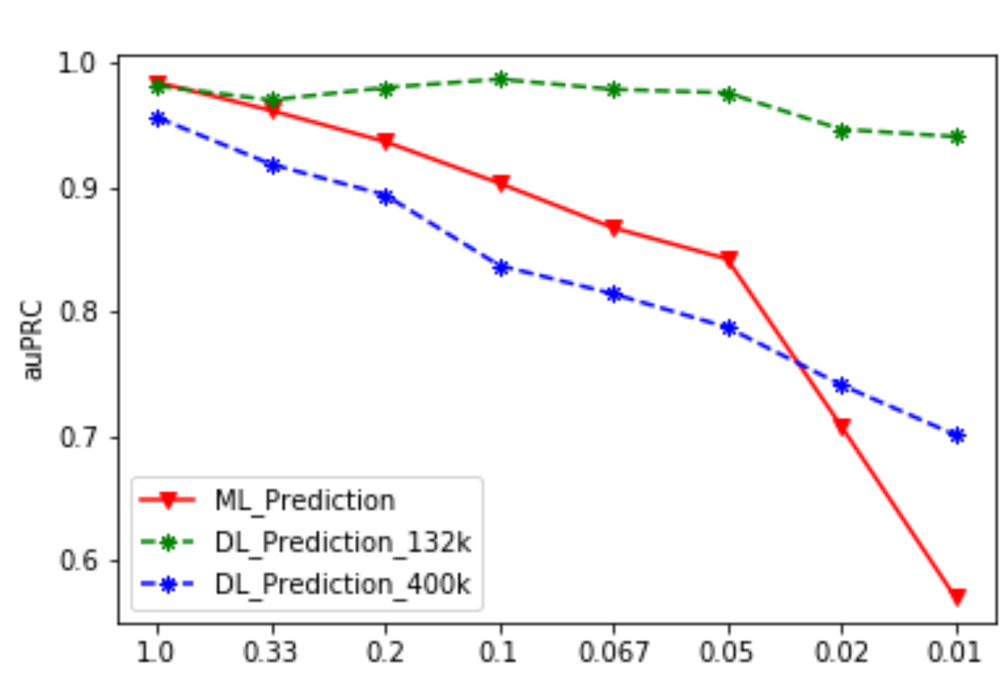
Overview of DL Vetting System



- Feeds raw apks into preprocessing layer; then generates API call sequence
- Treats each API call as a word; it uses the first 4000 API calls for each app
- Applies different embedding techniques such as Word2vec, GloVe, ELMo and BERT
- Each app, represented as a vector, is fed into an LSTM neural network layer with 4000 neurons

DL vs. Traditional ML Results

- Both traditional ML and DL models have good performance on balanced data
- Both models' performance decreases on unbalanced data
- DL model has better performance on highly unbalanced data



Benefits & Challenges

- Automated feature capability of DL could benefit mobile app vetting systems
- Efficiently applying DL for large-scale malware detection comes with significant challenges

UNIVERSITY OF SOUTH FLORIDA