# Automated Identification of Security Issues from Commit Messages and Bug Reports

[ : ] SourceClear

Yaqin Zhou, Asankhaya Sharma
SourceClear, Inc

## Motivation

Majority of vulnerabilities do not go through public disclosure with CVEs. The unidentified vulnerabilities put developers' products at risk of being hacked. Motivated to find the unidentified vulnerabilities in open source libraries, we design an automatic vulnerability identification system.

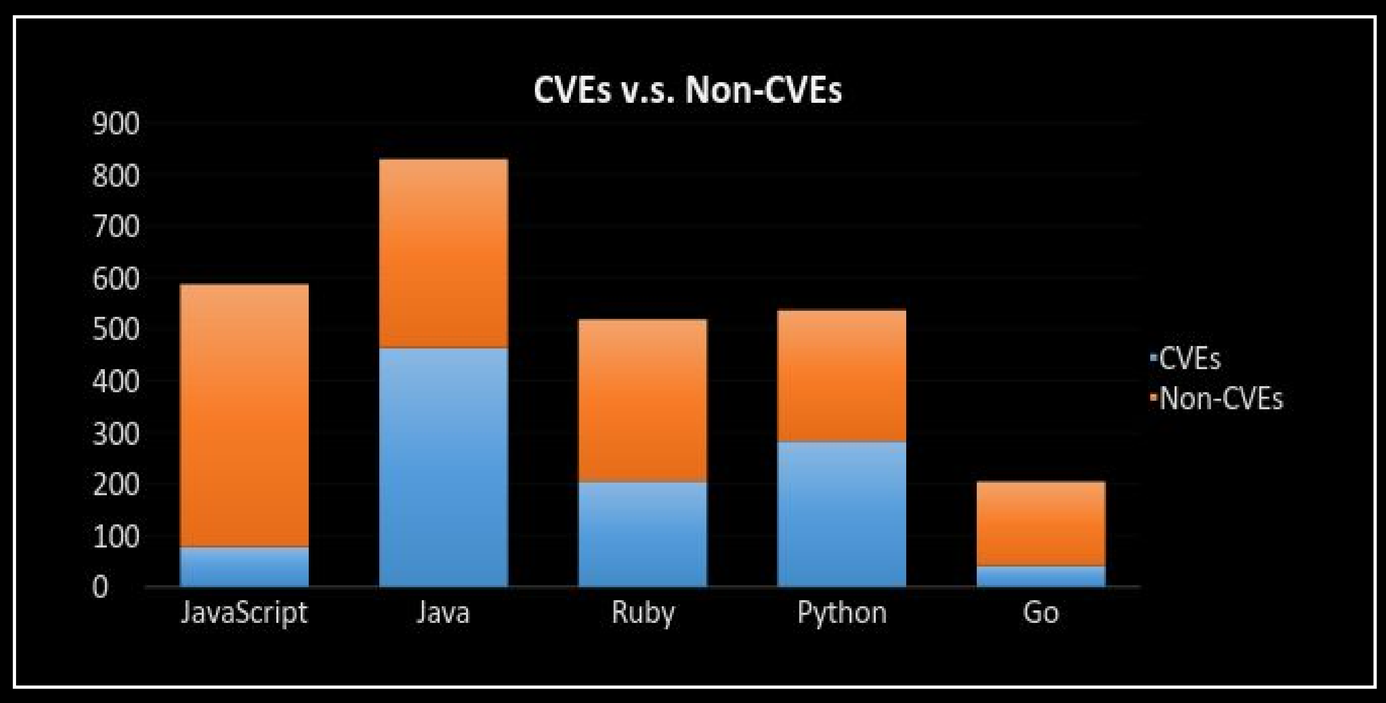


Fig 1: Number of vulnerabilities with CVEs v.s. without CVEs released in SourceClear Registry

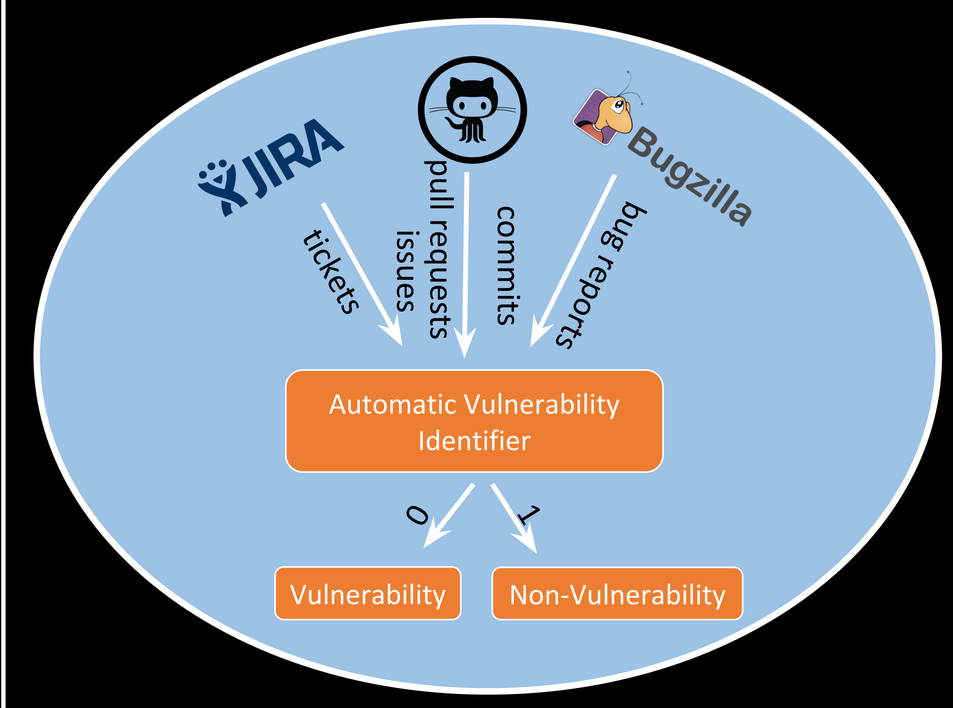


Fig 2: Workflow of automatic vulnerability identifier

## Approach

Built on machine learning and natural language processing techniques, our automatic vulnerability identifier extracts a wide range of security-related information from the commits/bug reports stream in real time, geared towards tracking vulnerabilities among a large-scale of projects at low cost.

## Training

A probability-based K-fold stacking learning algorithm that ensembles multiple individual classifiers to efficiently locate the tiny portion of vulnerabilities among massive data, and flexibly balances between the precision and recall rate.

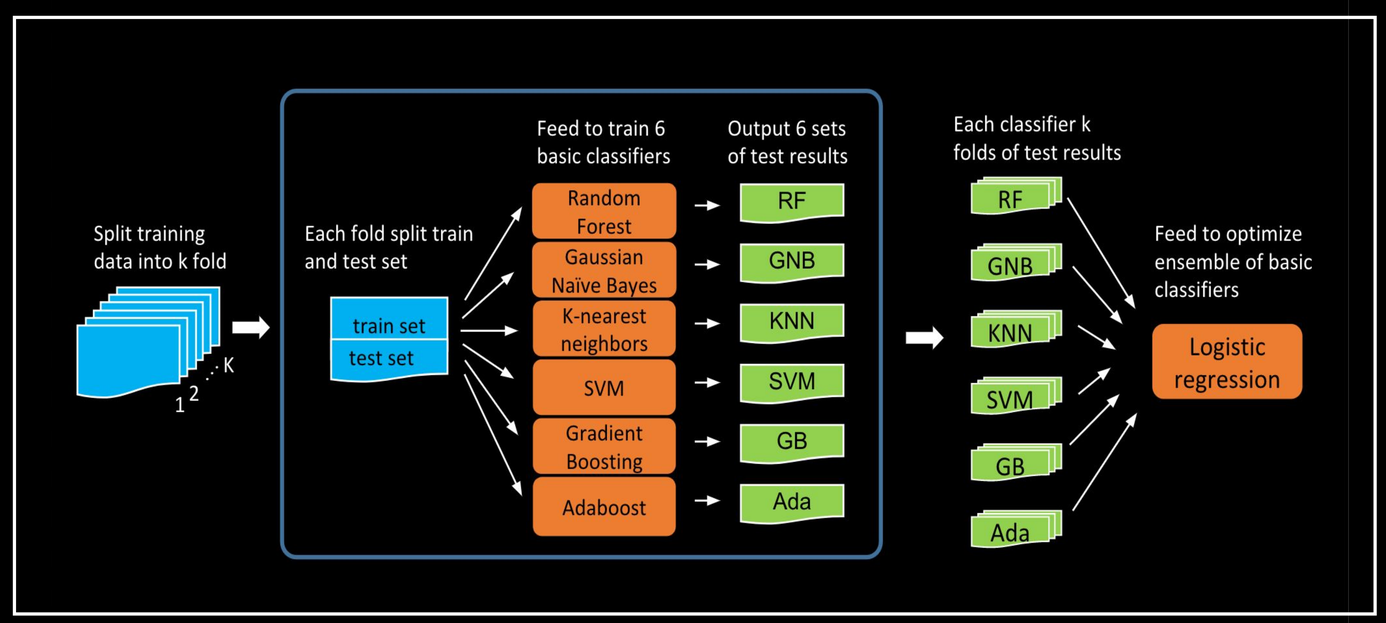


Fig 3: K-fold stacking model

## Production Observation

- 3 months use of commit model in production achieves *precision 0.83* and *recall rate 0.74*
- During the same period, our automatic vulnerability detection system found hidden vulnerabilities *349*, even larger than the total number of CVEs *333*

| Commits (Total) | Commits (Positive) | Commits (Negative) | True positive | False positive |
|---|---|---|---|---|
| 2268 | 215 | 2053 | 160 | 32 |

Table 1: Performance of the trained commit model on production

## Contact

- *Yaqin Zhou*, yaqin@sourceclear.com; *Asankhaya Sharma*, asankhaya@sourceclear.com, https://asankhaya.github.io
- *SourceClear*, https://www.sourceclear.com