EXAMENSARBETE Vulnerability Detection Using Ensemble Learning
Automatisk detektering av sårbarheter i öppen källkod med hjälp av maskininlärning
STUDENT Rasmus Lindqvist, Viktor Bard
HANDLEDARE Pierre Nugues (LTH), Emil Wåreus (Debricked)
EXAMINATOR Jacek Malek (LTH)

Automatisk detektering av sårbarheter i öppen källkod med hjälp av maskininlärning

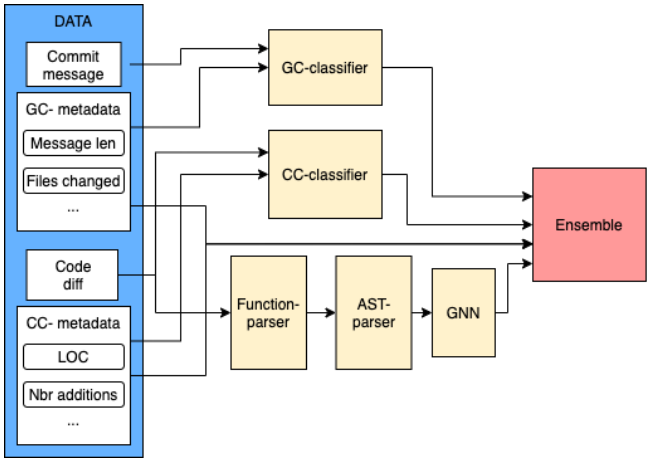
POPULÄRVETENSKAPLIG SAMMANFATTNING **Rasmus Lindqvist, Viktor Bard**

Användadet av öppen källkod ökar ständigt då det bidrar till ett billigt och snabbt sätt att utveckla mjukvara. Samtidigt blir är det svårare att hantera säkerhetsrisker i denna kod då koden är tillgänglig för utomstående parter att hitta delar som kan exploateras. Detta arbete undersöker potentialen i att kombinera flera olika maskininlärningsmodeller för att automatiskt hitta sårbarheter i kod.

Enligt en undersökning av Synopsis innehöll 99 % av analyserade kodbaser öppen källkod. Manuell detektering är dyrt och opraktiskt när mängden kod växer. Automatisk detektering av sårbarheter i öppen källkod har potentialen att vara billigare och bidra till större säkerhet.

Vi använder oss av olika data från GitHub och flera olika maskininlärningsmodeller i en så kallad ensemble. Ensemblen bygger på två enskilda modeller, tidigare utvecklade på Debricked. Båda är språkbehandlingsmodeller där den första bygger på meddelanden i commits och den andra på kodändringen. Utöver detta använder vi oss av tillhörande metadata till commitsen så som antal ändrade filer och rader kod. Slutligen testar vi inverkan av att lägga till en ytterligare enskild modell som klassificerar en kodändring som säkerhetsrelaterad eller inte, baserat på syntaxträd.

I den slutgiltiga arkitekturen, visat i figuren, använder vi oss av de tre olika modellerna och metadata, som input till ensembeln. Vi experimenterar med olika maskininlärningsalgoritmer till ensembel-modellen och logistisk regression visar sig vara den sammantaget bästa modellen.



Resultaten visar att modellen kan hitta 85.2 % av alla sårbarheter med en precision på 83.6 %. Detta är en ökning med 5.3 procentenheter fler hittade sårbarheter och 4.7 procentenheter bättre precision, jämfört med den bästa tidigare modellen. Tillägget av en kod-syntax gör att vi hittar fler sårbarheter men med sämre precision. Med hjälp av en säkerhetsexpert analyserar vi resultaten manuellt och hittar tidigare orapporterade sårbarheter.