INTERPRETING DEEP LEARNING MODELS FINE-TUNED FOR DETECTING VULNERABILITIES RELATED TO MISSING CODE

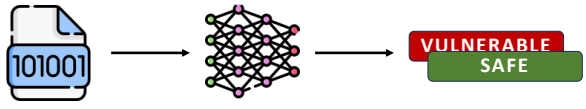
Claudia Mamede[1,2], Claire Le Goues[2], José Campos[1,3], Rui Abreu[1,4]

[1] Faculty of Engineering, University of Porto, Portugal
[2] Carnegie Mellon University, USA
[3] LASIGE, Faculty of Sciences, University of Lisbon, Portugal
[4] INESC-ID, Instituto Superior Técnico, University of Lisbon, Portugal

CIÊNCIA 2024

MOTIVATION

BACKGROUND. LEARNING-BASED VULNERABILITY DETECTION IS TYPICALLY TREATED AS A BINARY CLASSIFICATION PROBLEM.

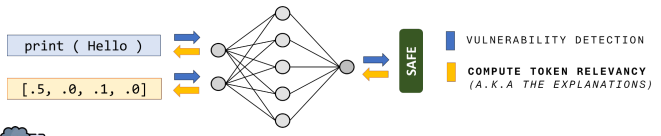


Function, code slice or file — Neural network — Binary label

PROBLEM 1. MODEL'S HIGH ACCURACY DOES NOT ENSURE REAL-WORLD EFFECTIVENESS

*IF THE 101001 HAS > 500 LoC, **HOW CAN WE FIND AND FIX THE VULNERABILITY?***

EXPLAINABLE AI (xAI) HELPS DEVELOPERS INTERPRET MODEL OUTPUTS BY ASSIGNING RELEVANCY SCORES TO INPUT TOKENS.

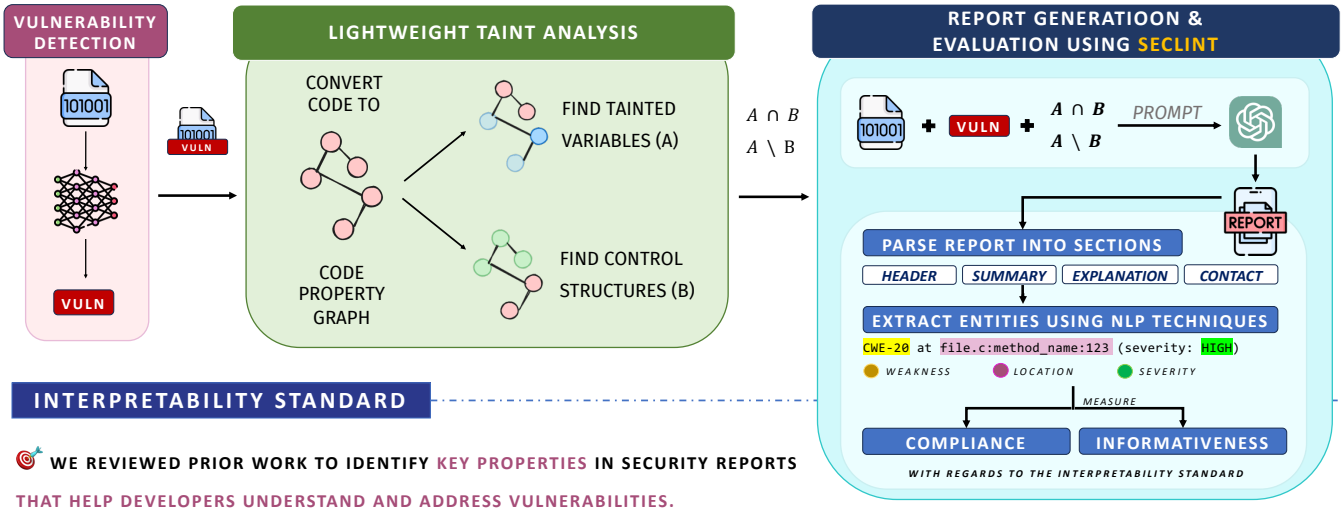


PROBLEM 2. EXISTING xAI METHODS ONLY EXPLAIN TOKENS PRESENT IN THE INPUT

HOW CAN WE EXPLAIN VULNERABILITIES ARISING FROM THE ABSENCE OF CODE?

IN VULNERABILITY DETECTION, xAI MAY HELP US LOCATE VULNERABILITIES.

METHODOLOGY



INTERPRETABILITY STANDARD

WE REVIEWED PRIOR WORK TO IDENTIFY KEY PROPERTIES IN SECURITY REPORTS THAT HELP DEVELOPERS UNDERSTAND AND ADDRESS VULNERABILITIES.

```
1. !VULN-DETECT: <WEAKNESS NAME/ID> AT <LOCATION> (SEVERITY: <LEVEL>)
2. !WHAT:        DESCRIBE THE WEAKNESS/PROBLEM
3. !WHY:         DESCRIBE ITS IMPACT
4. !HOW:         DESCRIBE HOW THE WEAKNESS CAN BE TRIGGERED
5. !WHEN:        DESCRIBE WHEN THE PROBLEM WAS FOUND
6. !WHERE:       DESCRIBE WHERE THE PROBLEM IS LOCATED
(…)
```

! – MANDATORY FIELDS ? – OPTIONAL FIELDS

FUTURE WORK

WE ARE STILL REFINING THE INTERPRETABILITY STANDARD

CHECK OUT THE QR CODE BELOW FOR THE UPDATED VERSION & GIVE US FEEDBACK !

Acknowledgments: This work was partially funded by Fundação para a Ciência e a Tecnologia under the scope of the CMU Portugal Program

Carnegie Mellon Portugal — fct Fundação para a Ciência e a Tecnologia — U.PORTO FEUP Faculdade de Engenharia Universidade do Porto — S3D Software and Societal Systems Department

LEARN MORE ABOUT THIS RESEARCH

INTERPRETING DEEP LEARNING MODELS FINE-TUNED FOR DETECTING VULNERABILITIES RELATED TO MISSING CODE

Title Example 1

Title

Claudia Mamede[1,2], Claire Le Goues[2], José Campos[1,3], Rui Abreu[1,4]

Name

[1] Faculty of Engineering, University of Porto, Portugal
[2] Carnegie Mellon University, USA
[3] LASIGE, Faculty of Sciences, University of Lisbon, Portugal
[4] INESC-ID, Instituto Superior Técnico, University of Lisbon, Portugal

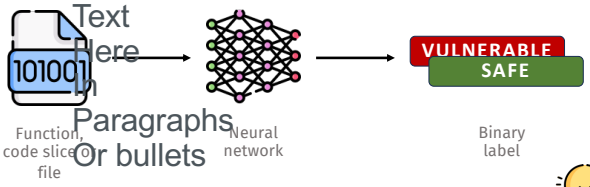
Affiliation

CIÊNCIA 2024

MOTIVATION

Introduction

BACKGROUND. LEARNING-BASED VULNERABILITY DETECTION IS TYPICALLY TREATED AS A BINARY CLASSIFICATION PROBLEM.



Methods

PROBLEM 1. MODEL'S HIGH ACCURACY DOES NOT ENSURE REAL-WORLD EFFECTIVENESS

IF THE HAS > 500 LoC, HOW CAN WE FIND AND FIX THE VULNERABILITY?

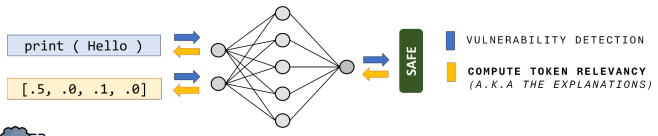
Text Here In Paragraphs Or bullets

EXPLAINABLE AI (xAI) HELPS DEVELOPERS INTERPRET MODEL OUTPUTS BY ASSIGNING RELEVANCY SCORES TO INPUT TOKENS.

Results

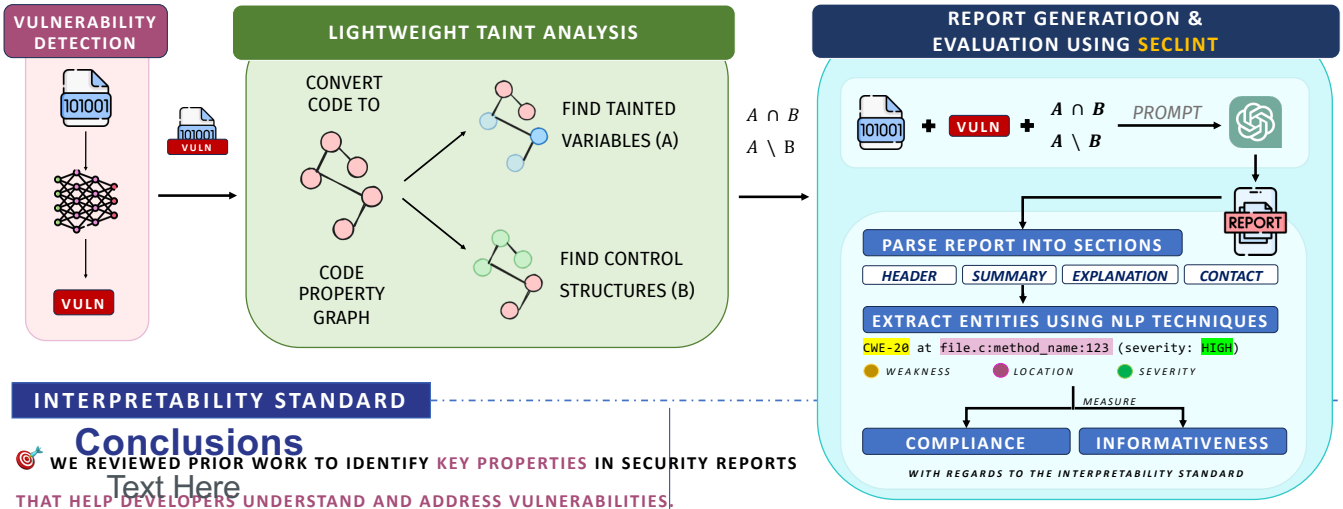
PROBLEM 2. EXISTING xAI METHODS ONLY EXPLAIN CODE PRESENT IN THE INPUT

HOW CAN WE EXPLAIN VULNERABILITIES ARISING FROM THE ABSENCE OF CODE?



IN VULNERABILITY DETECTION, xAI MAY HELP US LOCATE VULNERABILITIES.

METHODOLOGY



INTERPRETABILITY STANDARD

Conclusions

WE REVIEWED PRIOR WORK TO IDENTIFY KEY PROPERTIES IN SECURITY REPORTS THAT HELP DEVELOPERS UNDERSTAND AND ADDRESS VULNERABILITIES.

Text Here In Paragraphs Or bullets

```
1. !VULNDETECT: <WEAKNESS NAME/ID> AT <LOCATION> (SEVERITY: <LEVEL>)
2. !WHAT:       DESCRIBE THE WEAKNESS/PROBLEM
3. !WHY:        DESCRIBE ITS IMPACT
4. !HOW:        DESCRIBE HOW THE WEAKNESS CAN BE TRIGGERED
5. !WHEN:       DESCRIBE WHEN THE PROBLEM WAS FOUND
6. !WHERE:      DESCRIBE WHERE THE PROBLEM IS LOCATED
(…)
```

! – MANDATORY FIELDS ? – OPTIONAL FIELDS

Acknowledgements

This work was partially funded by Fundação para a Ciência e a Tecnologia under the scope of the CMU Portugal (Ref. FCT scholarship or Project's referece/name)

FUTURE WORK

CHECK OUT THE QR CODE BELOW FOR THE UPDATED VERSION & GIVE US FEEDBACK !

"Insert Logos from your Host Institution/Company"

LEARN MORE ABOUT THIS RESEARCH