An Empirical Evaluation of Explainable AI for Vulnerability Detection and Localization

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

[1] Faculty of Engineering of the University of Porto, Portugal

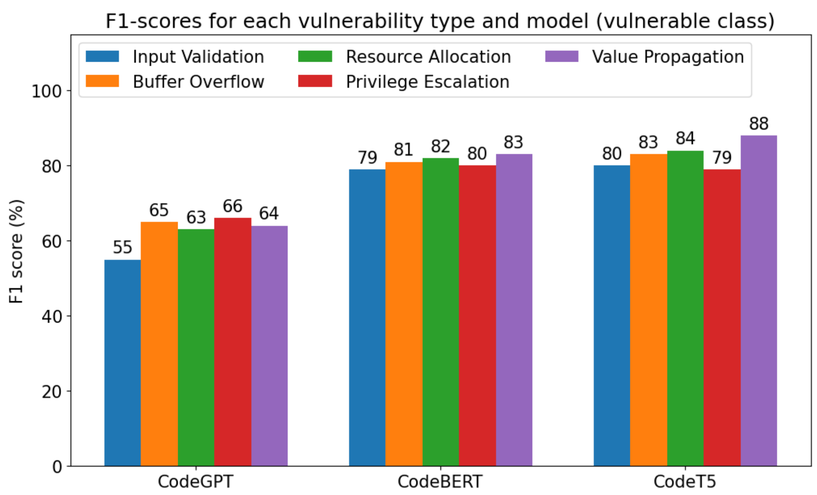
[2] Carnegie Mellon University, USA

[3] LASIGE, Faculty of Sciences of the University of Lisbon, Portugal

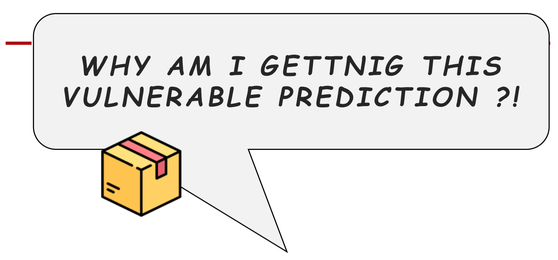
CMU Portugal 2023
Doctoral Symposium

MOTIVATION

BACKGROUND. TRANSFORMER-BASED YIELD STATE-OF-THE-ART RESULTS FOR VULNERABILITY DETECTION.



PROBLEM. THEY ARE "BLACK BOXES" SO DEVELOPERS DON'T USE THEM IN THE INDUSTRY.



Prediction: **VULNERABLE**

```
1 char filename[65];
2 char* temp;
3 temp = argv [1] ? argv[1] : strcpy(filename, temp);
```

EXPLAINABLE AI (xAI) HELPS DEVELOPERS COMPREHEND AND TRUST THE OUTPUTS OF AI MODELS.

DOES IT WORK IN THE CONTEXT OF VULNERABILITY DETECTION?

METHODOLOGY

5 DATASETS

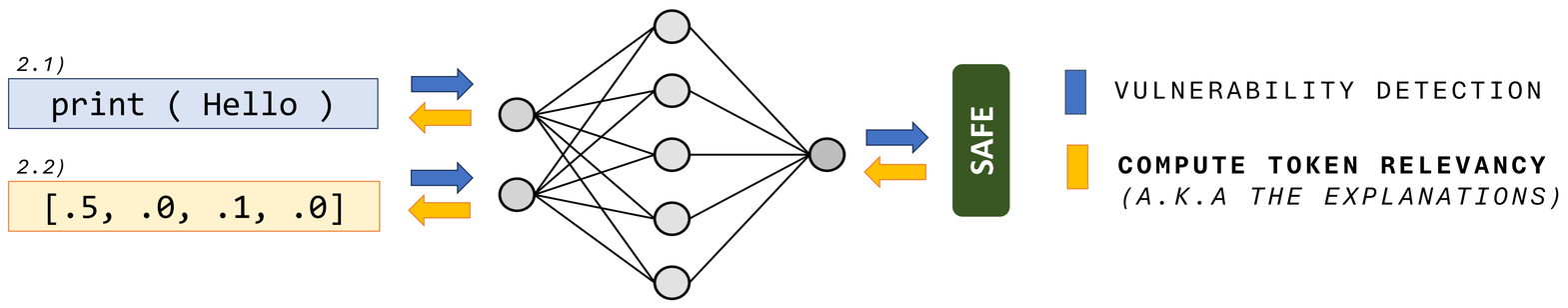
INPUT VALIDATION
BUFFER OVERFLOW
RESOURCE ALLOCATION
PRIVILEGE ESCALATION
VALUE PROPAGATION

3 MODELS

CODEGPT (DECODER-ONLY)
CODEBERT(ENCODER-ONLY)
CODET5 (ENCODER-DECODER)

STEP 1) FINETUNE EACH MODEL FOR BINARY VULNERABILITY DETECTION (W/ TRAINING DATA)

STEP 2) COMPUTE THE EXPLANATIONS USING INTEGRATED GRADIENTS (W/ TEST DATA)



RESULTS

COMPUTATIONAL-BASED EVALUATION METRICS

1. FAITHFULNESS

DO THE EXPLANATIONS TRULY REFLECT MODEL'S DECISIONS?

CODEBERT & CODET5 — CODEGPT

2. ROBUSTNESS

CAN WE TRUST THE MODEL? IS IT IMMUNE TO ATTACKS?

ALL MODELS ARE SENSITIVE TO MINOR INPUT CHANGES

3. COMPLEXITY

HOW INTERPRETABLE ARE THE EXPLANATIONS?

COMPLEXITY VARIES WITH MODEL AND VULNERABILITY TYPE.

BUFFER OVERFLOW-RELATED VULNERABILITIES HAVE THE LEAST COMPLEX EXPLANATIONS FOR BOTH MODELS.

USING xAI TO DETECT & LOCATE A BUFFER OVERFLOW WITH CODET5

Prediction: **VULNERABLE**

```
1 char filename[65];
2 char* temp;
3 temp = argv [1] ? argv[1] : strcpy(filename, temp);
```

CONTRIBUTE **NEGATIVELY** TO "VULNERABLE"

CONTRIBUTE **POSITIVELY** TO "VULNERABLE"

VULNERABILITY LOCALIZATION

IMPLICATIONS

- **TAKEAWAY 1.** LOW-COMPLEXITY EXPLANATIONS ALLOW VULNERABILITY LOCATION AT TOKEN-LEVEL, BOOSTING REPAIR SUCCESS CHANCES.
- **TAKEAWAY 2.** HIGH-STABLE EXPLANATIONS HELP DEVELOPERS TRUST MODEL PREDICTIONS.

Carnegie Mellon Portugal

fct Fundação para a Ciência e a Tecnologia

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

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

S3D Software and Societal Systems Department

GET IN TOUCH