CISUC

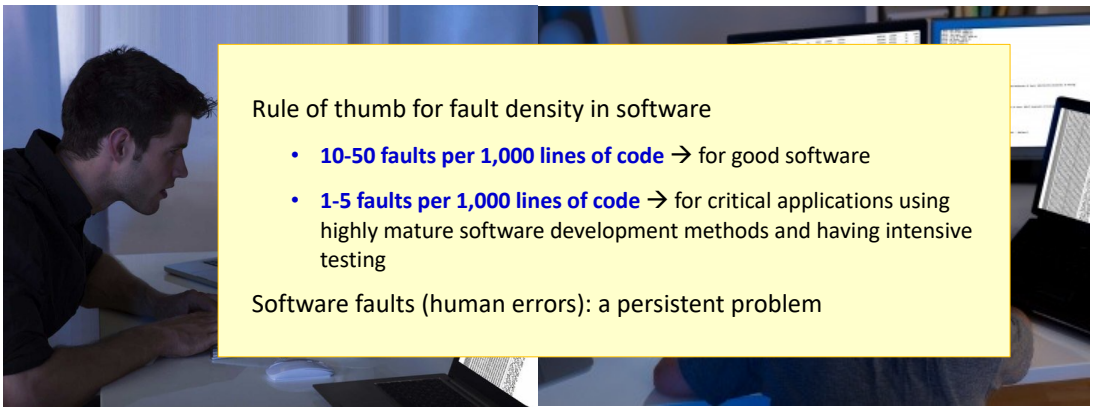
iReview: an Intelligent Code Review Evaluation Tool using Biofeedback

Henrique Madeira
CISUC, University of Coimbra

82nd IFIP WG 10.4 Meeting, June 23-26, 2022 – Old Town Alexandria, VA, USA

1

BASE - Biofeedback Augmented Software Engineering



Rule of thumb for fault density in software

- **10-50 faults per 1,000 lines of code** → for good software
- **1-5 faults per 1,000 lines of code** → for critical applications using highly mature software development methods and having intensive testing

Software faults (human errors): a persistent problem

87% of the severe software defects in deployed code are caused by human cognitive failures

Source: Huang, F., Liu, B., Wang, S., & Li, Q. (2015). The impact of software process consistency on residual defects

2

HUMAN COGNITIVE FAILURES

Distraction

Fatigue

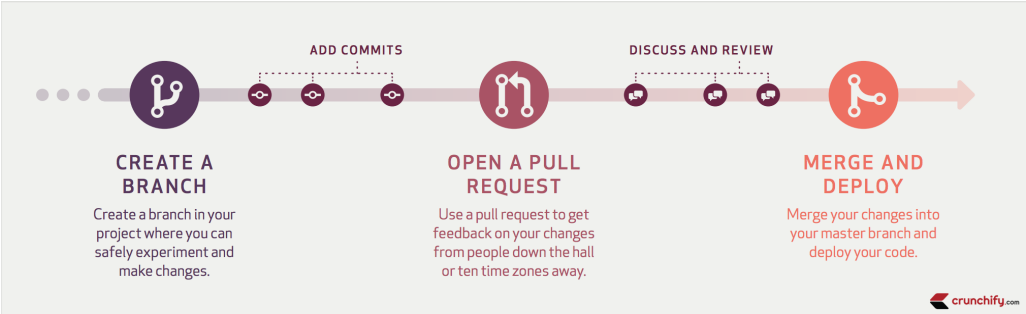
Code Comprehension Difficulty

Stress

3

Modern code review (asynchronous reviews)

Reviews by circulation are often called **modern code reviews** (although they have been proposed long time ago). These reviews are often associated with ***pull requests*** in distributed version-control systems like Git.



4

Modern code review (asynchronous reviews)

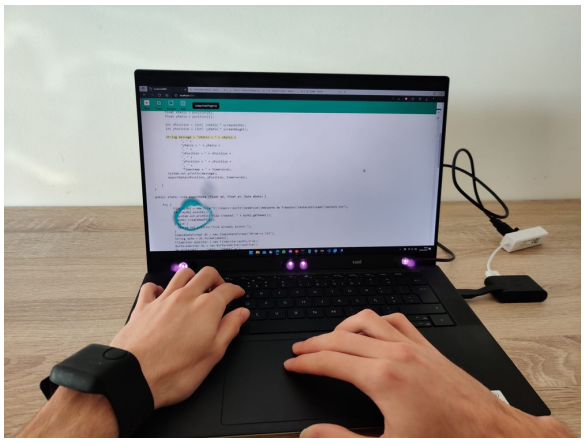
Reviews by circulation are often called **modern code reviews** (although they have been proposed long time ago). These reviews are often associated with ***pull requests*** in distributed version-control systems like Git. They have three important features:

Caitlin Sadowski Emma Söderberg Luke Church Michal Sipko Alberto Bacchelli, "**Modern Code Review: A Case Study at Google**", International Conference on Software Engineering, Software Engineering in Practice track (ICSE SEIP), 2018.

Shaumik Daityari, "**12 Best Code Review Tools for Developers (2022 Edition)**" (https://kinsta.com/blog/code-review-tools/), accessed on June 26, 2022.

5

6

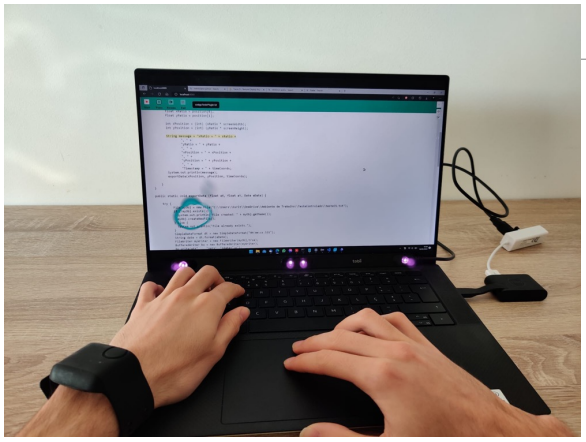


iReview…

Assess	Assess code comprehension difficulty through measuring cognitive load changes using a low-cost smartwatch to obtain Heart Signals.
Indicate	Indicate the code regions that are associated with high cognitive load and classified as “badly reviewed” using a desktop eye-tracker.
Explain	Explain the classification result (why “badly reviewed”?).

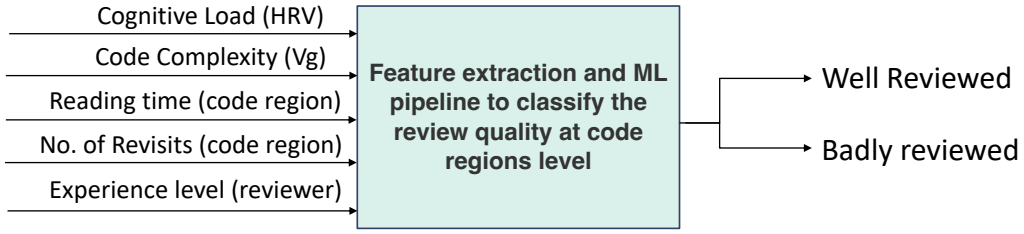
7

iReview…



8

iReview code review quality classification



iREVIEW PERFORMANCE

The capacity to classify the code region as “badly” reviewed when not all the bugs are detected.

The capacity to classify the code region as “well” reviewed when all the bugs are detected.

iReview performance results (logistic regression classifier)

PRGRAM	A	B	C	D
ALL	86.57%±5.65	68.00%±7.73	13.43%±5.65	33.00%±7.79
BSORT	92.31%±4.41	50.00%±2.28	7.69%±4.41	50.00%±7.28
FIBO	100.00%±0.00	90.91%±4.76	0.00%±0.00	9.09%±4.76
HONDT	95.24%±3.53	80.00%±6.63	4.76%±3.53	20.00%±6.63
MATDET	64.71%±7.92	83.87%±6.09	35.29%±7.92	16.13%±6.09

Ideal case: 100% Ideal case: 0%

A: Classified as bad/ not all bugs were detected
B: Classified as good/ all bugs were detected
C: Classified as bad/ all bugs were detected
D: Classified as good/ not all bugs were detected

3/18/24 14

14

iReview performance results (K-Nearest neighbors classifier)

PRGRAM	A	B	C	D
ALL	84.85%±5.94	70.27%±7.57	15.15%±5.94	29.73%±7.57
BSORT	87.10%±5.55	66.67%±7.81	12.90%±5.55	33.33%±7.81
FIBO	100.00%±0.00	90.91%±4.76	0.00%±0.00	9.09%±4.76
HONDT	94.74%±3.7	85.71%±5.8	5.26%±3.70	14.29%±5.80
MATDET	61.54%±8.06	77.14%±6.96	38.46%±8.06	22.86%±6.96

Ideal case: 100% Ideal case: 0%

A: Classified as bad/ not all bugs were detected
B: Classified as good/ all bugs were detected
C: Classified as bad/ all bugs were detected
D: Classified as good/ not all bugs were detected

3/18/24 15

15

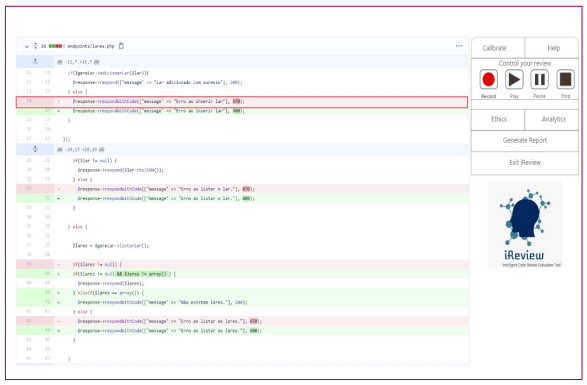
iReview tool looks

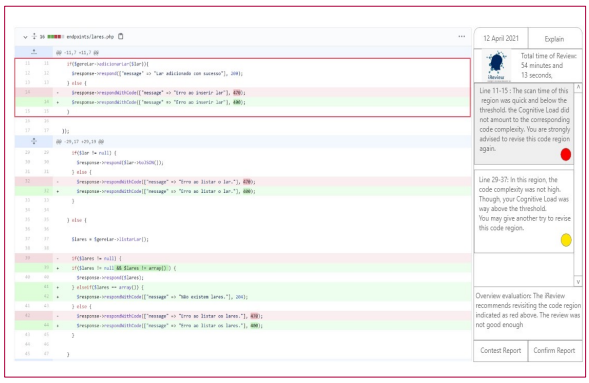
Main Interface

a) Start Review (record)
b) Pause and take a break
c) Stop and Generate the report

Evaluation Report

a) Overall evaluation
b) Code regions identification
c) Reasoning (XAI)





16

iReview Researchers & Recent Publications:

Patent Pending, Method and System for Scoring a Software Review Process, PCT/IB2021/059461

Hijazi, Haytham, Joao Duraes, Ricardo Couceiro, Joao Castelhano, Raul Barbosa, Júlio Medeiros, Miguel Castelo-Branco, Paulo De Carvalho, and Henrique Madeira. "Quality Evaluation of Modern Code Reviews Through Intelligent Biometric Program Comprehension." **IEEE Transactions on Software Engineering** 01 (2022): 1-1.

H. Hijazi, J. Cruz, J. Castelhano, R. Couceiro, M. Castelo-Branco, P. d. Carvalho and H. Madeira, iReview: An Intelligent Code Review Evaluation Tool using Biofeedback, in The 32nd International Symposium on Software Reliability Engineering (**ISSRE 2021**), 2021.

17

Acknowledgements

iReview was partially supported by Fundação para a Ciência e a Tecnologia, I.P./MCTES through PIDDAC and ECSEL Joint Undertaking (JU), under contract No 876852, project "ECSEL/0017/2019 | 876852-ECSEL-RIA-VALU3S", "Verification and Validation of Automated Systems' Safety and Security".

iReview was partially supported by the BASE (Biofeedback Augmented Software Engineering) project, Fundação para a Ciência e a Tecnologia, contract No IT057-18-7327.

FCT Fundação para a Ciência e a Tecnologia
REPÚBLICA PORTUGUESA
ECSEL Joint Undertaking
European Commission

CISUC

18