Title

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

Abstract—The abstract goes here.

Index Terms—Computer Society, IEEEtran, journal, LaTeX, paper, template.

1 Introduction

(RQ1) What is RQ1?
RQ1 results
(RQ2) What is RQ1?
RQ2 results

Paper organization. Section 2 situates this paper with respect to the related work. Section 3 discusses the design of our case study, while Section 4 presents the results with respect to our two research questions. Section 5 discloses the threats to the validity of our study. Finally, Section 6 draws conclusions.

2 Related Work & Research Questions

3 Case Study Design

In this section, we describe the design of our case study experiment that we perform in order to address our research questions. Figure **??** provides an overview of the approach that we apply to each studied system. The crux of our approach is that we calculate a ground truth performance such that the performance estimates derived from model validation techniques can be compared against it. We describe each step in the approach below.

3.1 Studied Systems

In selecting the studied systems, we identified two important criteria that needed to be satisfied:

- **Criterion 1 — Sufficient EPV**: Since we would like to study cases where EPV is low-risk (i.e, $\geq$ 10) and high-risk ($<$ 10), the systems that we select for analysis should begin with a low-risk EPV. Our rationale is that we prefer under-sampling to over-sampling when producing our sample dataset. For example, if we were to select systems with an initial EPV of 5, we would need to over-sample the defective class in order to raise the EPV to 10. However, the defective class of a system with an initial EPV of 15 can be under-sampled in order to lower the EPV to 10.
- **Criterion 2 — Sane defect data**: Since it is unlikely that more software modules have defects than are free of defects, we choose to study systems that have a rate of defective modules below 50%.

We began our study using the 101 publicly-available defect datasets described in Section 2. To satisfy criterion 1, we exclude the 78 datasets that we found to have an EPV value lower than 10 in Section 2. To satisfy criterion 2, we exclude an additional 5 datasets because they have a defective ratio above 50%.

Table **??** provides an overview of the 18 systems that satisfy our criteria for analysis. To combat potential bias in our conclusions, the studied systems include proprietary and open source systems, with varying size, domain, and defective ratio.

4 Case Study Results

In this section, we discuss our selection criteria for the studied systems and then present the results of our case study with respect to our two research questions.

(RQ1) What is RQ1?

(RQ2) What is RQ1?

5 Threats to Validity

5.1 Construct Validity

5.2 External Validity

6 Conclusions

Conclusions

- itemize

[1]

Acknowledgments

References

[1] C. Tantithamthavorn, S. McIntosh, A. E. Hassan, A. Ihara, and K. Matsumoto, "The Impact of Mislabelling on the Performance and Interpretation of Defect Prediction Models," in *Proceedings of the International Conference on Software Engineering*, 2015, p. To appear.

- *M. Shell is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: see http://www.michaelshell.org/contact.html*
- *J. Doe and J. Doe are with Anonymous University.*

Manuscript received April 19, 2005; revised September 17, 2014.