Noname manuscript No.
(will be inserted by the editor)

Supplemental Material 1

On the Validity of Retrospective Predictive Performance Evaluation Procedures in Just-In-Time Software Defect Prediction

Liyan Song · Leandro L. Minku* · Xin Yao*

the date of receipt and acceptance should be inserted later

Abstract This document presents additional information of the submitted paper "On the Validity of Retrospective Predictive Performance Evaluation Procedures in Just-In-Time Software Defect Prediction".

To answer R1∼RQ3 of this paper, we report Figures 3∼5 to assist reader better follow the explanations on experimental results of these research questions in Sections 6.1∼6.3, respectively. Their corresponding numerical values are reported in this supplementary material for further reference.

Liyan Song (songly@sustech.edu.cn) · Xin Yao (xiny@sustech.edu.cn)
Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, China and Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology, China.

Leandro L. Minku (L.L.Minku@bham.ac.uk)
School of Computer Science, the University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

* the corresponding authors.

Table 1 RQ1: Impact of waiting time on the amount of label noise in the retrospective performance evaluation scenario. Each value represents the average label noise across different lengths of the data stream (100, 2000, 3000, 4000 and 5000 time steps). Averaging across the data stream length is reasonable because it has no significant impact on the label noise associated to waiting time. The last row reports the median label noise associated to different waiting times across datasets.

Dataset	15	30	60	90
Brackets	0.1696	0.1507	0.1276	0.1051
Broadleaf	0.1875	0.1758	0.1560	0.1476
Camel	0.2071	0.1979	0.1769	0.1578
Fabric	0.3472	0.3352	0.3069	0.3011
jGroups	0.3958	0.3967	0.3972	0.4003
Nova	0.2869	0.2515	0.2124	0.1722
Django	0.3552	0.3413	0.3098	0.2976
Rails	0.3256	0.3072	0.2942	0.2766
Corefx	0.4201	0.3945	0.3722	0.3276
Rust	0.1753	0.1610	0.1411	0.1271
Tensorflow	0.3215	0.2782	0.1979	0.1327
VScode	0.3025	0.2191	0.1742	0.1234
wp-Calypso	0.6550	0.6252	0.4201	0.1964
Median	0.3215	0.2782	0.2124	0.1722

Table 2 RQ2: Impact of the training label noise on the validity of retrospective performance evaluation procedures. Each reported value is the average performance validity across evaluation label noises that are associated to different waiting time of 15, 30, 60 and 90 days. The last row reports the median performance validity with respect to training label noise across datasets. As the evaluation label noise also has significant impact on the performance validity, values in this table may not perfectly depict the effects of training label noise. Despite that, we show this table to provide additional experimental results of RQ2.

Brackets	label noise	0.2498	0.2782	0.3469	0.4101
	validity	0.9930	0.9929	0.9942	0.9944
Broadleaf	label noise	0.3227	0.3342	0.3972	0.4402
	validity	0.9951	0.9932	0.9896	0.9901
Camel	label noise	0.3138	0.3592	0.4264	0.4706
	validity	0.9954	0.9954	0.9946	0.9883
Fabric	label noise	0.4579	0.4940	0.5507	0.5958
	validity	0.9929	0.9872	0.9891	0.9926
jGroups	label noise	0.5736	0.5829	0.6360	0.6570
	validity	0.9879	0.9911	0.9905	0.9876
Nova	label noise	0.2913	0.3586	0.4470	0.5171
	validity	0.9897	0.9880	0.9573	0.9859
Django	label noise	0.5312	0.5862	0.6633	0.7015
	validity	0.9705	0.9620	0.9570	0.9542
Rails	label noise	0.4906	0.5490	0.6187	0.6708
	validity	0.9701	0.9560	0.9587	0.9742
Corefx	label noise	0.4849	0.5512	0.6312	0.7118
	validity	0.9875	0.9858	0.9942	0.9927
Rust	label noise	0.3855	0.4306	0.4750	0.5764
	validity	0.9969	0.9965	0.9722	0.9737
Tensorflow	label noise	0.2419	0.3379	0.4594	0.5606
	validity	0.9914	0.9890	0.9905	0.9859
VScode	label noise	0.2815	0.3121	0.3993	0.5504
	validity	0.9968	0.9921	0.9854	0.9889
wp-Calypso	label noise	0.2015	0.4050	0.6518	0.7475
	validity	0.9884	0.9912	0.9678	0.9806
Median	label noise	0.3400	0.4800	0.6200	0.7600
	validity	0.9929	0.9909	0.9859	0.9806

Table 3 RQ2: Impact of the evaluation label noise on the validity of retrospective performance evaluation procedures. Each value is the average performance validity across training label noises that are associated to different evaluation waiting time of 15, 30, 60 and 90 days. The last row reports the median performance validity with respect to evaluation label noise across datasets. As the training label noise also has significant impact on the performance validity, values in this table may not perfectly depict the effects of evaluation label noise. Despite that, we show this table to provide additional experimental results of RQ2.

Brackets	label noise	0.1281	0.1404	0.1635	0.1839
	validity	0.9943	0.9942	0.9940	0.9933
Broadleaf	label noise	0.1849	0.2054	0.2352	0.2524
	validity	0.9976	0.9970	0.9949	0.9929
Camel	label noise	0.1480	0.1582	0.1668	0.1743
	validity	0.9963	0.9959	0.9956	0.9945
Fabric	label noise	0.0830	0.0836	0.1298	0.1371
	validity	0.9897	0.9897	0.9901	0.9904
jGroups	label noise	0.3226	0.3243	0.3248	0.3254
	validity	0.9895	0.9894	0.9894	0.9894
Nova	label noise	0.2374	0.3002	0.3162	0.3292
	validity	0.9937	0.9912	0.9848	0.9823
Django	label noise	0.2881	0.2923	0.2956	0.3021
	validity	0.9727	0.9689	0.9661	0.9636
Rails	label noise	0.2801	0.2902	0.2920	0.2998
	validity	0.9728	0.9668	0.9646	0.9640
Corefx	label noise	0.3275	0.3322	0.3481	0.3712
	validity	0.9922	0.9887	0.9893	0.9888
Rust	label noise	0.0946	0.1012	0.1375	0.1538
	validity	0.9970	0.9973	0.9933	0.9895
Tensorflow	label noise	0.2419	0.2851	0.3242	0.3538
	validity	0.9910	0.9906	0.9910	0.9899
VScode	label noise	0.2243	0.2465	0.2716	0.2941
	validity	0.9981	0.9970	0.9942	0.9923
wp-Calypso	label noise	0.5047	0.5361	0.5737	0.6000
	validity	0.9768	0.9867	0.9845	0.9844
median	label noise	0.1500	0.3000	0.4500	0.6000
	validity	0.9938	0.9931	0.9894	0.9845

Table 4 RQ3: Impact of the training waiting time on the validity of retrospective performance evaluation procedures. Each value represents the average performance validity across different evaluation waiting times (15, 30, 60 and 90 days). Averaging across the evaluation waiting time is reasonable because the evaluation waiting time has no significant impact on the performance validity. The last row reports the median performance validity with respect to the training waiting time across datasets.

Dataset	15	30	60	90
Brackets	0.9944	0.9942	0.9929	0.9930
Broadleaf	0.9901	0.9896	0.9932	0.9951
Camel	0.9883	0.9946	0.9954	0.9954
Fabric	0.9926	0.9891	0.9872	0.9929
jGroups	0.9876	0.9905	0.9911	0.9879
Nova	0.9859	0.9573	0.9880	0.9897
Django	0.9542	0.9570	0.9620	0.9705
Rails	0.9742	0.9587	0.9560	0.9701
Corefx	0.9927	0.9942	0.9858	0.9875
Rust	0.9737	0.9722	0.9965	0.9969
Tensorflow	0.9859	0.9905	0.9890	0.9914
VScode	0.9889	0.9854	0.9921	0.9968
wp-Calypso	0.9806	0.9678	0.9912	0.9884
Median	0.9876	0.9891	0.9911	0.9914