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

An Investigation of Online and Offline Learning Models for Online Just-in-Time Software Defect Prediction – Supplementary Material

George G. Cabral · Leandro L. Minku · Adriano L. I. Oliveira · Dinaldo A. Pessoa · Sadia Tabassum*

This supplementary material complements the paper entitled "An Investigation of Online and Offline Learning Models for Online Just-in-Time Software Defect Prediction".

1 Hyper-parameter Configurations

* Authors in alphabetical order.

George G. Cabral
Department of Computing, Federal Rural University of Pernambuco, 52.171-900, Recife, BR
E-mail: george.gcabral@ufrpe.br

L.L. Minku (corresponding author) and S. Tabassum
School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
E-mail: sxt901@student.bham.ac.uk and E-mail: l.l.minku@bham.ac.uk

Adriano L. I. Oliveira
Center for Informatics, Federal University of Pernambuco, Recife, 50.740-560, BR
E-mail: alio@cin.ufpe.br

Dinaldo A. Pessoa
Banco Central do Brasil, R. da Aurora, 1259, Santo Amaro, Recife, 50040-090, BR
E-mail: dinaldoap@gmail.com

Table 1: List of hyper-parameter values explored in hyper-parameter tuning using random search

Classifier/Base-Learner	**Investigated Hyper-parameters values and probability distribution**
ORB	orb-waiting-time [90,180] uniform
	orb-ma-window-size [50,200] uniform
	orb-th [0.3,0.5] uniform
	orb-l0 [1.0,20.0] loguniform
	orb-l1 [1.0,20.0] loguniform
	orb-m [1.1,2.718281828459045] uniform
	orb-decay-factor [0.9,0.999] loguniform
	orb-n [3,7] uniform
BORB	borb-waiting-time [90,180] uniform
	borb-ma-window-size [50,200] uniform
	borb-th [0.3,0.5] uniform
	borb-l0 [1.0,20.0] loguniform
	borb-l1 [1.0,20.0] loguniform
	borb-m [1.1,2.718281828459045] uniform
	borb-pull-request-size [50,500] uniform
	borb-sample-size [1000,4000] uniform
Base Learners	oht-n-estimators [10,40] uniform
	oht-grace-period [100,500] uniform
	oht-split-criterion {'gini','info_gain', 'hellinger'} uniform
	oht-split-confidence [1e-07,0.5] loguniform
	oht-tie-threshold [0.05,0.5] uniform
	oht-no-preprune {0,1} uniform
	oht-leaf-prediction {'nba', 'nb', 'mc'} uniform
	ihf-n-estimators [10,30] uniform
	ihf-grace-period [100,500] uniform
	ihf-split-criterion {'gini','info_gain', 'hellinger'} uniform
	ihf-split-confidence [1e-07,0.5] loguniform
	ihf-tie-threshold [0.05,0.5] uniform
	ihf-no-preprune {0,1} uniform
	ihf-leaf-prediction {'nba', 'nb', 'mc'} uniform
	lr-alpha [0.01,1.0] loguniform
	lr-n-epochs [10,80] uniform
	lr-batch-size [128,512] loguniform
	lr-log-transformation {0,1} uniform
	mlp-learning-rate [0.0001,0.01] loguniform
	mlp-n-hidden-layers [1,3] uniform
	mlp-hidden-layers-size [5,15] uniform
	mlp-dropout-input-layer [0.1,0.3] uniform
	mlp-dropout-hidden-layer [0.3,0.5] uniform
	mlp-n-epochs [10,80] uniform
	mlp-batch-size [128,512] loguniform
	mlp-log-transformation {0,1} uniform
	nb-n-updates [10,80] uniform
	irf-n-estimators [20,100] uniform
	irf-criterion {'gini',entropy} uniform
	irf-min-samples-leaf [100,300] uniform
	irf-max-features [3,7] uniform

2 Performance Metrics

The paper corresponding to this supplementary material reported the results of the metric G-Mean. This section reports the value of the following additional performance metrics for the approaches investigated in the study for reference:

- True Positive Rate (TPR), a.k.a., Recall, Recall1, Sensitivity.
- True Negative Rate (TNR), a.k.a., Recall0, Specificity.
- False Positive Rate (FPR).
- False Negative Rate (FNR).
- Precision.
- Matthews Correlation Coefficient (MCC).

Among these, we recommend readers to use the TPR, TNR, FPR and FNR. These metrics are ratios associated to the counters of the four quadrants of the confusion matrix, fully representing the performance of the models in an unbiased manner. The G-Mean combines them into a single metric in an unbiased way [2] that incorporates all four quadrants, given that $\textit{G-Mean} = \sqrt{TPR \times TNR}$, $TPR = 1 - FNR$ and $TPR = 1 - FNR$. Other performance metrics can be computed by combining values from different quadrants of the confusion matrix, but depending on the metric it may become biased as is the case of precision and MCC [2].

We give an example to illustrate such bias below. Consider a classifier that is always able to correctly predict 70% of the clean (negative) examples, and 70% of the defect-inducing (positive) examples. This is a reasonably good classifier. Now, consider the following two test sets that could be used to evaluate it:

- Test set 1 (balanced): 100 clean and 100 defect-inducing examples.
- Test set 2 (imbalanced): 160 clean and 40 defect-inducing examples. This dataset has an imbalance ratio of 20%, which is similar to the median imbalance ratio of the datasets used in this study.

The precision ($\frac{TP}{TP+FP}$) of this classifier is the following when evaluated on each of these test sets:

- Precision on test set 1: $70/(70 + 30) = 70\%$.
- Precision on test set 2: $28/(28 + 48) \approx 37\%$.

As shown above, even though the classifier has exactly the same quality, when it is tested on these two different sets, it receives a better evaluation in test set 1 than in test set 2. Moreover, even though this is a reasonably good classifier, it would appear to be poor if one relied on test set 2. This happens because precision is biased when the data are class imbalanced. In particular, because the number of positive examples is relatively small in an imbalanced test set, no matter how good the classifier is, the numerator of the precision equation will be relatively small. Conversely, because the number of negative examples is relatively large, even if the false positive rate is fairly low, the denominator of the equation tends to become relatively large. Therefore,

the precision tends to get low when tested on class imbalanced data even for classifiers of quite good quality. The more class imbalanced the test set, the worse the precision of a given fixed classifier.

The MCC values ($\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$) computed for this example are also shown below:

- MCC on test set 1:
  $(70 \times 70 - 30 \times 30) / \sqrt{(70+30) \times (70+30) \times (70+30) \times (70+30)} = 40\%$
- MCC on test set 2:
  $(28 \times 112 - 48 \times 12) / \sqrt{(28+48) \times (28+12) \times (112+48) \times (112+12)} = \approx 33\%$

As we can see, the MCC metric also results in a classifier of the same quality having different evaluations depending on the imbalance ratio of the test set, though to a less drastic extent than precision.

It is also worth noting that the adoption of biased metrics such as precision and MCC can be particularly problematic in problems with class imbalance evolution such as JIT-SDP [1], as a classifier whose quality is constant over time would appear to have its quality varying over time due to changes in the imbalance ratio of the problem.

Table 2: Average TPR for BORB-WP

Dataset	BORB-NB-WP	BORB-MLP-WP	BORB-LR-WP	BORB-IRF-WP	BORB-IHF-WP
Neutron	0.851 (0.001)	0.910 (0.003)	0.925 (0.001)	0.911 (0.010)	0.899 (0.003)
Fabric8	0.836 (0.003)	0.725 (0.009)	0.672 (0.001)	0.757 (0.003)	0.717 (0.009)
Nova	0.930 (0.003)	0.887 (0.001)	0.900 (0.000)	0.860 (0.002)	0.893 (0.002)
Tomcat	0.922 (0.005)	0.718 (0.009)	0.687 (0.001)	0.741 (0.006)	0.672 (0.006)
Spring-Integration	0.669 (0.006)	0.675 (0.005)	0.707 (0.002)	0.653 (0.005)	0.744 (0.011)
Brackets	0.889 (0.001)	0.874 (0.002)	0.854 (0.001)	0.875 (0.002)	0.893 (0.006)
Camel	0.670 (0.002)	0.726 (0.003)	0.746 (0.001)	0.723 (0.003)	0.623 (0.003)
NPM	0.734 (0.004)	0.739 (0.009)	0.707 (0.002)	0.679 (0.022)	0.687 (0.021)
JGroups	0.703 (0.006)	0.699 (0.008)	0.634 (0.002)	0.597 (0.009)	0.643 (0.006)
Broadleaf	0.880 (0.002)	0.756 (0.003)	0.715 (0.002)	0.720 (0.005)	0.734 (0.008)

Table 3: Average TPR for BORB-CP

Dataset	BORB-NB-CP	BORB-MLP-CP	BORB-LR-CP	BORB-IRF-CP	BORB-IHF-CP
Neutron	0.979 (0.003)	0.901 (0.001)	0.914 (0.001)	0.910 (0.001)	0.887 (0.004)
Fabric8	0.697 (0.022)	0.723 (0.003)	0.654 (0.001)	0.764 (0.002)	0.714 (0.004)
Nova	0.939 (0.004)	0.865 (0.001)	0.873 (0.001)	0.861 (0.001)	0.930 (0.001)
Tomcat	0.571 (0.016)	0.666 (0.002)	0.671 (0.002)	0.672 (0.002)	0.663 (0.004)
Spring-Integration	0.693 (0.019)	0.770 (0.004)	0.807 (0.001)	0.704 (0.003)	0.747 (0.005)
Brackets	0.958 (0.005)	0.813 (0.002)	0.829 (0.001)	0.837 (0.002)	0.792 (0.003)
Camel	0.735 (0.004)	0.758 (0.002)	0.791 (0.001)	0.797 (0.001)	0.767 (0.002)
NPM	0.643 (0.037)	0.703 (0.015)	0.711 (0.001)	0.704 (0.003)	0.611 (0.005)
JGroups	0.770 (0.006)	0.700 (0.008)	0.637 (0.001)	0.598 (0.008)	0.634 (0.008)
Broadleaf	0.693 (0.022)	0.768 (0.003)	0.796 (0.001)	0.696 (0.002)	0.692 (0.006)

Table 4: Average TPR for ORB-WP

Dataset	ORB-OHT-WP	ORB-NB-WP	ORB-MLP-WP	ORB-LR-WP
Neutron	0.765 (0.011)	0.851 (0.021)	0.479 (0.157)	0.519 (0.016)
Fabric8	0.608 (0.009)	0.768 (0.063)	0.471 (0.207)	0.359 (0.015)
Nova	0.858 (0.004)	0.863 (0.025)	0.667 (0.267)	0.475 (0.006)
Tomcat	0.653 (0.006)	0.706 (0.166)	0.536 (0.068)	0.615 (0.012)
Spring-Integration	0.366 (0.012)	0.573 (0.032)	0.359 (0.108)	0.558 (0.023)
Brackets	0.834 (0.006)	0.868 (0.049)	0.501 (0.185)	0.560 (0.013)
Camel	0.442 (0.010)	0.693 (0.062)	0.473 (0.060)	0.520 (0.007)
NPM	0.618 (0.025)	0.736 (0.021)	0.531 (0.070)	0.508 (0.022)
JGroups	0.627 (0.009)	0.617 (0.049)	0.459 (0.076)	0.468 (0.011)
Broadleaf	0.651 (0.011)	0.622 (0.109)	0.472 (0.083)	0.611 (0.011)

Table 5: Average TPR for ORB-CP

Dataset	ORB-OHT-CP	ORB-NB-CP	ORB-MLP-CP	ORB-LR-CP
Neutron	0.822 (0.007)	0.951 (0.020)	0.453 (0.102)	0.448 (0.006)
Fabric8	0.672 (0.007)	0.546 (0.075)	0.432 (0.044)	0.473 (0.020)
Nova	0.823 (0.004)	0.931 (0.022)	0.463 (0.134)	0.528 (0.005)
Tomcat	0.649 (0.004)	0.529 (0.043)	0.453 (0.062)	0.514 (0.010)
Spring-Integration	0.736 (0.007)	0.472 (0.048)	0.434 (0.055)	0.485 (0.014)
Brackets	0.750 (0.006)	0.810 (0.045)	0.481 (0.093)	0.502 (0.010)
Camel	0.734 (0.003)	0.534 (0.040)	0.446 (0.081)	0.569 (0.006)
NPM	0.695 (0.010)	0.587 (0.175)	0.445 (0.059)	0.471 (0.026)
JGroups	0.635 (0.010)	0.583 (0.054)	0.463 (0.075)	0.487 (0.011)
Broadleaf	0.728 (0.005)	0.488 (0.113)	0.438 (0.054)	0.395 (0.011)

Standard deviations are shown in brackets

Table 6: Average TNR for BORB-WP

Dataset	BORB-NB-WP	BORB-MLP-WP	BORB-LR-WP	BORB-IRF-WP	BORB-IHF-WP
Neutron	0.216 (0.002)	0.768 (0.002)	0.732 (0.000)	0.736 (0.003)	0.758 (0.002)
Fabric8	0.455 (0.003)	0.672 (0.009)	0.725 (0.001)	0.685 (0.005)	0.691 (0.007)
Nova	0.611 (0.006)	0.756 (0.001)	0.734 (0.000)	0.768 (0.001)	0.729 (0.002)
Tomcat	0.118 (0.006)	0.687 (0.006)	0.719 (0.001)	0.664 (0.006)	0.674 (0.005)
Spring-Integration	0.479 (0.010)	0.746 (0.004)	0.727 (0.002)	0.768 (0.008)	0.547 (0.036)
Brackets	0.536 (0.001)	0.744 (0.002)	0.746 (0.001)	0.733 (0.002)	0.624 (0.009)
Camel	0.490 (0.003)	0.705 (0.001)	0.699 (0.000)	0.706 (0.001)	0.770 (0.002)
NPM	0.382 (0.005)	0.593 (0.006)	0.642 (0.002)	0.634 (0.013)	0.526 (0.017)
JGroups	0.457 (0.007)	0.660 (0.004)	0.674 (0.001)	0.739 (0.005)	0.651 (0.007)
Broadleaf	0.420 (0.004)	0.727 (0.002)	0.749 (0.001)	0.737 (0.007)	0.664 (0.005)

Table 7: Average TNR for BORB-CP

Dataset	BORB-NB-CP	BORB-MLP-CP	BORB-LR-CP	BORB-IRF-CP	BORB-IHF-CP
Neutron	0.538 (0.023)	0.760 (0.001)	0.777 (0.000)	0.767 (0.001)	0.736 (0.003)
Fabric8	0.403 (0.018)	0.652 (0.002)	0.691 (0.001)	0.622 (0.001)	0.604 (0.005)
Nova	0.366 (0.012)	0.755 (0.001)	0.757 (0.000)	0.754 (0.000)	0.661 (0.001)
Tomcat	0.488 (0.016)	0.710 (0.002)	0.729 (0.002)	0.709 (0.001)	0.646 (0.005)
Spring-Integration	0.355 (0.026)	0.694 (0.002)	0.666 (0.001)	0.758 (0.001)	0.669 (0.007)
Brackets	0.352 (0.013)	0.742 (0.001)	0.740 (0.001)	0.722 (0.001)	0.664 (0.003)
Camel	0.425 (0.009)	0.640 (0.001)	0.611 (0.000)	0.591 (0.001)	0.572 (0.003)
NPM	0.413 (0.032)	0.652 (0.011)	0.647 (0.001)	0.641 (0.002)	0.646 (0.003)
JGroups	0.334 (0.010)	0.656 (0.004)	0.677 (0.001)	0.739 (0.004)	0.665 (0.007)
Broadleaf	0.407 (0.019)	0.660 (0.002)	0.652 (0.001)	0.729 (0.001)	0.734 (0.003)

Table 8: Average TNR for ORB-WP

Dataset	ORB-OHT-WP	ORB-NB-WP	ORB-MLP-WP	ORB-LR-WP
Neutron	0.821 (0.007)	0.259 (0.060)	0.559 (0.055)	0.509 (0.008)
Fabric8	0.773 (0.009)	0.528 (0.058)	0.592 (0.180)	0.688 (0.007)
Nova	0.765 (0.003)	0.736 (0.024)	0.427 (0.199)	0.579 (0.006)
Tomcat	0.660 (0.004)	0.452 (0.193)	0.478 (0.037)	0.385 (0.009)
Spring-Integration	0.776 (0.014)	0.648 (0.046)	0.665 (0.102)	0.519 (0.034)
Brackets	0.719 (0.007)	0.566 (0.057)	0.607 (0.131)	0.533 (0.011)
Camel	0.891 (0.004)	0.456 (0.064)	0.568 (0.018)	0.515 (0.005)
NPM	0.661 (0.009)	0.427 (0.042)	0.503 (0.040)	0.548 (0.012)
JGroups	0.654 (0.005)	0.577 (0.059)	0.541 (0.031)	0.577 (0.007)
Broadleaf	0.726 (0.010)	0.601 (0.076)	0.548 (0.033)	0.429 (0.010)

Table 9: Average TNR for ORB-CP

Dataset	ORB-OHT-CP	ORB-NB-CP	ORB-MLP-CP	ORB-LR-CP
Neutron	0.837 (0.004)	0.678 (0.045)	0.605 (0.039)	0.560 (0.005)
Fabric8	0.706 (0.004)	0.625 (0.055)	0.582 (0.017)	0.549 (0.008)
Nova	0.775 (0.002)	0.622 (0.040)	0.543 (0.074)	0.492 (0.005)
Tomcat	0.681 (0.002)	0.604 (0.040)	0.528 (0.025)	0.489 (0.010)
Spring-Integration	0.675 (0.004)	0.667 (0.035)	0.575 (0.032)	0.529 (0.013)
Brackets	0.772 (0.005)	0.557 (0.055)	0.526 (0.058)	0.507 (0.007)
Camel	0.631 (0.002)	0.686 (0.034)	0.525 (0.026)	0.438 (0.004)
NPM	0.619 (0.004)	0.547 (0.152)	0.535 (0.026)	0.528 (0.007)
JGroups	0.648 (0.005)	0.584 (0.068)	0.540 (0.030)	0.541 (0.008)
Broadleaf	0.704 (0.004)	0.708 (0.079)	0.599 (0.016)	0.630 (0.006)

Standard deviations are shown in brackets

Table 10: Average FPR for BORB-WP

Dataset	BORB-NB-WP	BORB-MLP-WP	BORB-LR-WP	BORB-IRF-WP	BORB-IHF-WP
Neutron	0.784 (0.002)	0.232 (0.002)	0.268 (0.000)	0.264 (0.003)	0.242 (0.002)
Fabric8	0.545 (0.003)	0.328 (0.009)	0.275 (0.001)	0.315 (0.005)	0.309 (0.007)
Nova	0.389 (0.006)	0.244 (0.001)	0.266 (0.000)	0.232 (0.001)	0.271 (0.002)
Tomcat	0.882 (0.006)	0.313 (0.006)	0.281 (0.001)	0.336 (0.006)	0.326 (0.005)
Spring-Integration	0.521 (0.010)	0.254 (0.004)	0.273 (0.002)	0.232 (0.008)	0.453 (0.036)
Brackets	0.464 (0.001)	0.256 (0.002)	0.254 (0.001)	0.267 (0.002)	0.376 (0.009)
Camel	0.510 (0.003)	0.295 (0.001)	0.301 (0.000)	0.294 (0.001)	0.230 (0.002)
NPM	0.618 (0.005)	0.407 (0.006)	0.358 (0.002)	0.366 (0.013)	0.474 (0.017)
JGroups	0.543 (0.007)	0.340 (0.004)	0.326 (0.001)	0.261 (0.005)	0.349 (0.007)
Broadleaf	0.580 (0.004)	0.273 (0.002)	0.251 (0.001)	0.263 (0.007)	0.336 (0.005)

Table 11: Average FPR for BORB-CP

Dataset	BORB-NB-CP	BORB-MLP-CP	BORB-LR-CP	BORB-IRF-CP	BORB-IHF-CP
Neutron	0.462 (0.023)	0.240 (0.001)	0.223 (0.000)	0.233 (0.001)	0.264 (0.003)
Fabric8	0.597 (0.018)	0.348 (0.002)	0.309 (0.001)	0.378 (0.001)	0.396 (0.005)
Nova	0.634 (0.012)	0.245 (0.001)	0.243 (0.000)	0.246 (0.000)	0.339 (0.001)
Tomcat	0.512 (0.016)	0.290 (0.002)	0.271 (0.002)	0.291 (0.001)	0.354 (0.005)
Spring-Integration	0.645 (0.026)	0.306 (0.002)	0.334 (0.001)	0.242 (0.001)	0.331 (0.007)
Brackets	0.648 (0.013)	0.258 (0.001)	0.260 (0.001)	0.278 (0.001)	0.336 (0.003)
Camel	0.575 (0.009)	0.360 (0.001)	0.389 (0.000)	0.409 (0.001)	0.428 (0.003)
NPM	0.587 (0.032)	0.348 (0.011)	0.353 (0.001)	0.359 (0.002)	0.354 (0.003)
JGroups	0.666 (0.010)	0.344 (0.004)	0.323 (0.001)	0.261 (0.004)	0.335 (0.007)
Broadleaf	0.593 (0.019)	0.340 (0.002)	0.348 (0.001)	0.271 (0.001)	0.266 (0.003)

Table 12: Average FPR for ORB-WP

Dataset	ORB-OHT-WP	ORB-NB-WP	ORB-MLP-WP	ORB-LR-WP
Neutron	0.179 (0.007)	0.741 (0.060)	0.441 (0.055)	0.491 (0.008)
Fabric8	0.227 (0.009)	0.472 (0.058)	0.408 (0.180)	0.312 (0.007)
Nova	0.235 (0.003)	0.264 (0.024)	0.573 (0.199)	0.421 (0.006)
Tomcat	0.340 (0.004)	0.548 (0.193)	0.522 (0.037)	0.615 (0.009)
Spring-Integration	0.224 (0.014)	0.352 (0.046)	0.335 (0.102)	0.481 (0.034)
Brackets	0.281 (0.007)	0.434 (0.057)	0.393 (0.131)	0.467 (0.011)
Camel	0.109 (0.004)	0.544 (0.064)	0.432 (0.018)	0.485 (0.005)
NPM	0.339 (0.009)	0.573 (0.042)	0.497 (0.040)	0.452 (0.012)
JGroups	0.346 (0.005)	0.423 (0.059)	0.459 (0.031)	0.423 (0.007)
Broadleaf	0.274 (0.010)	0.399 (0.076)	0.452 (0.033)	0.571 (0.010)

Table 13: Average FPR for ORB-CP

Dataset	ORB-OHT-CP	ORB-NB-CP	ORB-MLP-CP	ORB-LR-CP
Neutron	0.163 (0.004)	0.322 (0.045)	0.395 (0.039)	0.440 (0.005)
Fabric8	0.294 (0.004)	0.375 (0.055)	0.418 (0.017)	0.451 (0.008)
Nova	0.225 (0.002)	0.378 (0.040)	0.457 (0.074)	0.508 (0.005)
Tomcat	0.319 (0.002)	0.396 (0.040)	0.472 (0.025)	0.511 (0.010)
Spring-Integration	0.325 (0.004)	0.333 (0.035)	0.425 (0.032)	0.471 (0.013)
Brackets	0.228 (0.005)	0.443 (0.055)	0.474 (0.058)	0.493 (0.007)
Camel	0.369 (0.002)	0.314 (0.034)	0.475 (0.026)	0.562 (0.004)
NPM	0.381 (0.004)	0.453 (0.152)	0.465 (0.026)	0.472 (0.007)
JGroups	0.352 (0.005)	0.416 (0.068)	0.460 (0.030)	0.459 (0.008)
Broadleaf	0.296 (0.004)	0.292 (0.079)	0.401 (0.016)	0.370 (0.006)

Standard deviations are shown in brackets

Table 14: Average FNR for BORB-WP

Dataset	BORB-NB-WP	BORB-MLP-WP	BORB-LR-WP	BORB-IRF-WP	BORB-IHF-WP
Neutron	0.149 (0.001)	0.090 (0.003)	0.075 (0.001)	0.089 (0.010)	0.101 (0.003)
Fabric8	0.164 (0.003)	0.275 (0.009)	0.328 (0.001)	0.243 (0.003)	0.283 (0.009)
Nova	0.070 (0.003)	0.113 (0.001)	0.100 (0.000)	0.140 (0.002)	0.107 (0.002)
Tomcat	0.078 (0.005)	0.282 (0.009)	0.313 (0.001)	0.259 (0.006)	0.328 (0.006)
Spring-Integration	0.331 (0.006)	0.325 (0.005)	0.293 (0.002)	0.347 (0.005)	0.256 (0.011)
Brackets	0.111 (0.001)	0.126 (0.002)	0.146 (0.001)	0.125 (0.002)	0.107 (0.006)
Camel	0.330 (0.002)	0.274 (0.003)	0.254 (0.001)	0.277 (0.003)	0.377 (0.003)
NPM	0.266 (0.004)	0.261 (0.009)	0.293 (0.002)	0.321 (0.022)	0.313 (0.021)
JGroups	0.297 (0.006)	0.301 (0.008)	0.366 (0.002)	0.403 (0.009)	0.357 (0.006)
Broadleaf	0.120 (0.002)	0.244 (0.003)	0.285 (0.002)	0.280 (0.005)	0.266 (0.008)

Table 15: Average FNR for BORB-CP

Dataset	BORB-NB-CP	BORB-MLP-CP	BORB-LR-CP	BORB-IRF-CP	BORB-IHF-CP
Neutron	0.021 (0.003)	0.099 (0.001)	0.086 (0.001)	0.090 (0.001)	0.113 (0.004)
Fabric8	0.303 (0.022)	0.277 (0.003)	0.346 (0.001)	0.236 (0.002)	0.286 (0.004)
Nova	0.061 (0.004)	0.135 (0.001)	0.127 (0.001)	0.139 (0.001)	0.070 (0.001)
Tomcat	0.429 (0.016)	0.334 (0.002)	0.329 (0.002)	0.328 (0.002)	0.337 (0.004)
Spring-Integration	0.307 (0.019)	0.230 (0.004)	0.193 (0.001)	0.296 (0.003)	0.253 (0.005)
Brackets	0.042 (0.005)	0.187 (0.002)	0.171 (0.001)	0.163 (0.002)	0.208 (0.003)
Camel	0.265 (0.004)	0.242 (0.002)	0.209 (0.001)	0.203 (0.001)	0.233 (0.002)
NPM	0.357 (0.037)	0.297 (0.015)	0.289 (0.001)	0.296 (0.003)	0.389 (0.005)
JGroups	0.230 (0.006)	0.300 (0.008)	0.363 (0.001)	0.402 (0.008)	0.366 (0.008)
Broadleaf	0.307 (0.022)	0.232 (0.003)	0.204 (0.001)	0.304 (0.002)	0.308 (0.006)

Table 16: Average FNR for ORB-WP

Dataset	ORB-OHT-WP	ORB-NB-WP	ORB-MLP-WP	ORB-LR-WP
Neutron	0.235 (0.011)	0.149 (0.021)	0.521 (0.157)	0.481 (0.016)
Fabric8	0.392 (0.009)	0.232 (0.063)	0.529 (0.207)	0.641 (0.015)
Nova	0.142 (0.004)	0.137 (0.025)	0.333 (0.267)	0.525 (0.006)
Tomcat	0.347 (0.006)	0.294 (0.166)	0.464 (0.068)	0.385 (0.012)
Spring-Integration	0.634 (0.012)	0.427 (0.032)	0.641 (0.108)	0.442 (0.023)
Brackets	0.166 (0.006)	0.132 (0.049)	0.499 (0.185)	0.440 (0.013)
Camel	0.558 (0.010)	0.307 (0.062)	0.527 (0.060)	0.480 (0.007)
NPM	0.382 (0.025)	0.264 (0.021)	0.469 (0.070)	0.492 (0.022)
JGroups	0.373 (0.009)	0.383 (0.049)	0.541 (0.076)	0.532 (0.011)
Broadleaf	0.349 (0.011)	0.378 (0.109)	0.528 (0.083)	0.389 (0.011)

Table 17: Average FNR for ORB-CP

Dataset	ORB-OHT-CP	ORB-NB-CP	ORB-MLP-CP	ORB-LR-CP
Neutron	0.178 (0.007)	0.049 (0.020)	0.547 (0.102)	0.552 (0.006)
Fabric8	0.328 (0.007)	0.454 (0.075)	0.568 (0.044)	0.527 (0.020)
Nova	0.177 (0.004)	0.069 (0.022)	0.537 (0.134)	0.472 (0.005)
Tomcat	0.351 (0.004)	0.471 (0.043)	0.547 (0.062)	0.486 (0.010)
Spring-Integration	0.264 (0.007)	0.528 (0.048)	0.566 (0.055)	0.515 (0.014)
Brackets	0.250 (0.006)	0.190 (0.045)	0.519 (0.093)	0.498 (0.010)
Camel	0.266 (0.003)	0.466 (0.040)	0.554 (0.081)	0.431 (0.006)
NPM	0.305 (0.010)	0.413 (0.175)	0.555 (0.059)	0.529 (0.026)
JGroups	0.365 (0.010)	0.417 (0.054)	0.537 (0.075)	0.513 (0.011)
Broadleaf	0.272 (0.005)	0.512 (0.113)	0.562 (0.054)	0.605 (0.011)

Standard deviations are shown in brackets

Table 18: Average Precision for BORB-WP

Dataset	BORB-NB-WP	BORB-MLP-WP	BORB-LR-WP	BORB-IRF-WP	BORB-IHF-WP
Neutron	0.253 (0.000)	0.551 (0.002)	0.519 (0.000)	0.519 (0.002)	0.537 (0.002)
Fabric8	0.281 (0.001)	0.360 (0.005)	0.384 (0.001)	0.380 (0.004)	0.372 (0.004)
Nova	0.448 (0.003)	0.553 (0.001)	0.535 (0.000)	0.557 (0.001)	0.528 (0.001)
Tomcat	0.284 (0.001)	0.466 (0.003)	0.482 (0.000)	0.456 (0.004)	0.439 (0.003)
Spring-Integration	0.318 (0.004)	0.491 (0.004)	0.485 (0.002)	0.506 (0.009)	0.375 (0.017)
Brackets	0.371 (0.001)	0.513 (0.002)	0.509 (0.001)	0.503 (0.001)	0.423 (0.005)
Camel	0.255 (0.001)	0.391 (0.001)	0.393 (0.000)	0.391 (0.002)	0.414 (0.002)
NPM	0.204 (0.001)	0.282 (0.003)	0.299 (0.001)	0.286 (0.008)	0.238 (0.005)
JGroups	0.213 (0.002)	0.301 (0.002)	0.289 (0.001)	0.323 (0.005)	0.278 (0.004)
Broadleaf	0.235 (0.001)	0.360 (0.002)	0.366 (0.001)	0.357 (0.006)	0.307 (0.003)

Table 19: Average Precision for BORB-CP

Dataset	BORB-NB-CP	BORB-MLP-CP	BORB-LR-CP	BORB-IRF-CP	BORB-IHF-CP
Neutron	0.399 (0.012)	0.540 (0.001)	0.561 (0.001)	0.549 (0.001)	0.512 (0.002)
Fabric8	0.229 (0.006)	0.346 (0.002)	0.350 (0.001)	0.340 (0.001)	0.315 (0.003)
Nova	0.335 (0.004)	0.546 (0.001)	0.549 (0.000)	0.544 (0.001)	0.483 (0.001)
Tomcat	0.298 (0.003)	0.466 (0.002)	0.485 (0.002)	0.468 (0.001)	0.416 (0.003)
Spring-Integration	0.281 (0.005)	0.478 (0.002)	0.468 (0.000)	0.514 (0.001)	0.451 (0.005)
Brackets	0.313 (0.004)	0.493 (0.001)	0.495 (0.001)	0.481 (0.001)	0.421 (0.002)
Camel	0.250 (0.002)	0.355 (0.001)	0.347 (0.000)	0.337 (0.001)	0.319 (0.002)
NPM	0.191 (0.005)	0.304 (0.003)	0.303 (0.001)	0.297 (0.001)	0.272 (0.002)
JGroups	0.195 (0.002)	0.298 (0.002)	0.292 (0.000)	0.324 (0.004)	0.283 (0.003)
Broadleaf	0.192 (0.004)	0.314 (0.001)	0.317 (0.000)	0.343 (0.001)	0.345 (0.003)

Table 20: Average Precision for ORB-WP

Dataset	ORB-OHT-WP	ORB-NB-WP	ORB-MLP-WP	ORB-LR-WP
Neutron	0.572 (0.007)	0.265 (0.019)	0.254 (0.085)	0.248 (0.005)
Fabric8	0.406 (0.011)	0.294 (0.010)	0.247 (0.094)	0.226 (0.007)
Nova	0.553 (0.003)	0.527 (0.016)	0.277 (0.072)	0.277 (0.003)
Tomcat	0.422 (0.004)	0.338 (0.032)	0.281 (0.037)	0.276 (0.004)
Spring-Integration	0.373 (0.009)	0.374 (0.022)	0.286 (0.065)	0.297 (0.012)
Brackets	0.478 (0.005)	0.383 (0.021)	0.283 (0.067)	0.270 (0.005)
Camel	0.514 (0.006)	0.250 (0.012)	0.222 (0.029)	0.219 (0.002)
NPM	0.283 (0.008)	0.218 (0.009)	0.188 (0.025)	0.195 (0.005)
JGroups	0.275 (0.004)	0.235 (0.013)	0.173 (0.027)	0.188 (0.004)
Broadleaf	0.325 (0.009)	0.241 (0.006)	0.175 (0.032)	0.178 (0.003)

Table 21: Average Precision for ORB-CP

Dataset	ORB-OHT-CP	ORB-NB-CP	ORB-MLP-CP	ORB-LR-CP
Neutron	0.611 (0.004)	0.482 (0.028)	0.264 (0.061)	0.242 (0.003)
Fabric8	0.368 (0.005)	0.271 (0.016)	0.208 (0.020)	0.211 (0.009)
Nova	0.554 (0.002)	0.457 (0.020)	0.260 (0.085)	0.261 (0.002)
Tomcat	0.436 (0.003)	0.337 (0.008)	0.267 (0.031)	0.277 (0.006)
Spring-Integration	0.452 (0.005)	0.341 (0.024)	0.271 (0.037)	0.272 (0.009)
Brackets	0.504 (0.006)	0.362 (0.017)	0.241 (0.055)	0.239 (0.003)
Camel	0.341 (0.002)	0.308 (0.012)	0.197 (0.034)	0.209 (0.002)
NPM	0.283 (0.004)	0.221 (0.016)	0.171 (0.020)	0.177 (0.008)
JGroups	0.274 (0.004)	0.228 (0.015)	0.173 (0.024)	0.182 (0.003)
Broadleaf	0.333 (0.003)	0.255 (0.025)	0.181 (0.023)	0.178 (0.004)

Standard deviations are shown in brackets

Table 22: Average MCC for BORB-WP

Dataset	BORB-NB-WP	BORB-MLP-WP	BORB-LR-WP	BORB-IRF-WP	BORB-IHF-WP
Neutron	0.071 (0.002)	0.591 (0.002)	0.567 (0.001)	0.559 (0.006)	0.572 (0.003)
Fabric8	0.239 (0.002)	0.325 (0.006)	0.334 (0.001)	0.362 (0.005)	0.336 (0.006)
Nova	0.472 (0.004)	0.570 (0.001)	0.558 (0.000)	0.560 (0.002)	0.547 (0.001)
Tomcat	0.057 (0.005)	0.366 (0.004)	0.371 (0.001)	0.364 (0.007)	0.313 (0.005)
Spring-Integration	0.132 (0.009)	0.386 (0.006)	0.394 (0.002)	0.392 (0.011)	0.258 (0.027)
Brackets	0.364 (0.001)	0.535 (0.002)	0.521 (0.001)	0.525 (0.002)	0.439 (0.005)
Camel	0.130 (0.002)	0.358 (0.002)	0.369 (0.001)	0.358 (0.003)	0.344 (0.003)
NPM	0.092 (0.004)	0.255 (0.006)	0.270 (0.002)	0.243 (0.017)	0.163 (0.012)
JGroups	0.123 (0.005)	0.277 (0.005)	0.240 (0.001)	0.272 (0.008)	0.227 (0.006)
Broadleaf	0.233 (0.002)	0.379 (0.003)	0.370 (0.002)	0.361 (0.007)	0.304 (0.005)

Table 23: Average MCC for BORB-CP

Dataset	BORB-NB-CP	BORB-MLP-CP	BORB-LR-CP	BORB-IRF-CP	BORB-IHF-CP
Neutron	0.447 (0.016)	0.576 (0.001)	0.604 (0.001)	0.590 (0.001)	0.539 (0.003)
Fabric8	0.083 (0.019)	0.306 (0.003)	0.286 (0.001)	0.312 (0.002)	0.257 (0.005)
Nova	0.293 (0.009)	0.550 (0.001)	0.559 (0.001)	0.546 (0.001)	0.515 (0.001)
Tomcat	0.053 (0.007)	0.344 (0.003)	0.368 (0.003)	0.348 (0.002)	0.278 (0.004)
Spring-Integration	0.044 (0.014)	0.415 (0.003)	0.420 (0.001)	0.424 (0.002)	0.371 (0.007)
Brackets	0.294 (0.010)	0.484 (0.002)	0.493 (0.001)	0.482 (0.002)	0.390 (0.003)
Camel	0.133 (0.006)	0.325 (0.002)	0.327 (0.001)	0.315 (0.001)	0.275 (0.003)
NPM	0.044 (0.015)	0.275 (0.006)	0.277 (0.001)	0.267 (0.002)	0.201 (0.004)
JGroups	0.085 (0.007)	0.274 (0.004)	0.245 (0.001)	0.274 (0.007)	0.232 (0.005)
Broadleaf	0.077 (0.014)	0.325 (0.002)	0.339 (0.001)	0.335 (0.002)	0.337 (0.004)

Table 24: Average MCC for ORB-WP

Dataset	ORB-OHT-WP	ORB-NB-WP	ORB-MLP-WP	ORB-LR-WP
Neutron	0.535 (0.007)	0.109 (0.057)	0.033 (0.179)	0.023 (0.011)
Fabric8	0.334 (0.014)	0.240 (0.015)	0.060 (0.074)	0.040 (0.013)
Nova	0.554 (0.003)	0.529 (0.009)	0.094 (0.148)	0.047 (0.005)
Tomcat	0.282 (0.007)	0.152 (0.030)	0.012 (0.087)	0.000 (0.012)
Spring-Integration	0.143 (0.009)	0.200 (0.030)	0.025 (0.099)	0.069 (0.026)
Brackets	0.476 (0.006)	0.371 (0.013)	0.095 (0.132)	0.079 (0.010)
Camel	0.353 (0.005)	0.123 (0.025)	0.033 (0.063)	0.028 (0.005)
NPM	0.219 (0.018)	0.128 (0.022)	0.026 (0.066)	0.043 (0.012)
JGroups	0.218 (0.008)	0.148 (0.013)	-0.001 (0.067)	0.034 (0.009)
Broadleaf	0.298 (0.013)	0.170 (0.024)	0.015 (0.079)	0.031 (0.009)

Table 25: Average MCC for ORB-CP

Dataset	ORB-OHT-CP	ORB-NB-CP	ORB-MLP-CP	ORB-LR-CP
Neutron	0.601 (0.004)	0.538 (0.034)	0.050 (0.120)	0.007 (0.007)
Fabric8	0.314 (0.008)	0.140 (0.032)	0.012 (0.043)	0.017 (0.020)
Nova	0.537 (0.004)	0.483 (0.018)	0.006 (0.173)	0.017 (0.004)
Tomcat	0.300 (0.004)	0.120 (0.011)	-0.017 (0.064)	0.003 (0.013)
Spring-Integration	0.367 (0.009)	0.128 (0.040)	0.008 (0.072)	0.012 (0.019)
Brackets	0.464 (0.007)	0.313 (0.016)	0.007 (0.118)	0.008 (0.007)
Camel	0.297 (0.004)	0.186 (0.017)	-0.023 (0.078)	0.006 (0.005)
NPM	0.242 (0.009)	0.107 (0.043)	-0.015 (0.049)	-0.001 (0.019)
JGroups	0.218 (0.008)	0.128 (0.019)	0.002 (0.060)	0.022 (0.008)
Broadleaf	0.335 (0.005)	0.158 (0.050)	0.028 (0.051)	0.019 (0.009)

Standard deviations are shown in brackets

References

1. Cabral, G.G., Minku, L.L., Shihab, E., Mujahid, S.: Class imbalance evolution and verification latency in just-in-time software defect prediction. In: 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), pp. 666–676. IEEE (2019)
2. Zhu, Q.: On the performance of matthews correlation coefficient (MCC) for imbalanced dataset. Pattern Recognition Letters **136**, 71–80 (2020)