Study on Exploring Disagreements in XAI Techniques for Software Defect Prediction: An In-Depth Evaluation

Dear Survey Participants,

We are excited to inform you about a user study on eXplainable Artificial Intelligence (XAI) techniques and their application in software defect prediction. As part of this study, we will explore the impact of feature selection on reducing disagreement among XAI techniques.

In recent years, XAI has emerged as a critical area of research to enhance the transparency and interpretability of machine learning models. However, one challenge often arises in applying XAI techniques is the potential for disagreement among different methods' explanations of model predictions. This disagreement can occur due to various factors, including the choice of features used by each technique to generate explanations.

Feature selection plays a crucial role in determining which aspects of the input data are considered by XAI techniques when generating explanations. By carefully selecting features, researchers and practitioners can mitigate disagreement among different XAI techniques and improve the consistency and reliability of their explanations.

In this user study, we aim to investigate the effect of feature selection on reducing disagreement among XAI techniques. We will provide you (participants) with a set of scenarios and ask (them) to interpret model predictions using different XAI techniques with various feature selection techniques. Through this study, we hope to gain valuable insights into the role of feature selection in improving the consistency and trustworthiness of XAI explanations.

Your participation in this study is invaluable, and your feedback will contribute to advancing our understanding of XAI techniques and their practical application. Thank you for your participation, and we look forward to your valuable insights.

The survey will take about 10–15 minutes for you to complete. All the responses will be anonymized, and we will not share your personal information with anyone else.

Thank you very much for your participation!

Sincerely,
Saumendu Roy
Doctoral Researcher, Department of Computer Science
University of Saskatchewan, Canada

Contact Email: saumendu.roy@usask.ca, saumocse3j6@gmail.com

* Indicates required question

1. Company/University *

2. Country *

3. How many years are you working on Software Development/ Machine Learning/ Explainability? *

Mark only one oval.

- ◯ Less than 1 year
- ◯ 1 year
- ◯ 2 years
- ◯ 3 years
- ◯ More than 3 years
- ◯ Other: _____

4. Please select your role (Options are below) *

Mark only one oval.

- ◯ Postdoctoral Researcher
- ◯ PhD Researcher
- ◯ Researcher (Master's) / Master's Student
- ◯ Software Developer
- ◯ End-user
- ◯ Other

5. Have you used explainability methods in your work before? *

Mark only one oval.

- ◯ Yes
- ◯ No

6. Which explainability methods do you use in your day-to-day workflow? *

Check all that apply.

- ☐ LIME
- ☐ SHAP
- ☐ BreakDown
- ☐ PyExplainer
- ☐ Other
- ☐ Other: _____

7. Do you encounter disagreement in the explanations generated by state-of-the-art methods in your daily workflow?

Mark only one oval.

- ◯ Yes
- ◯ No

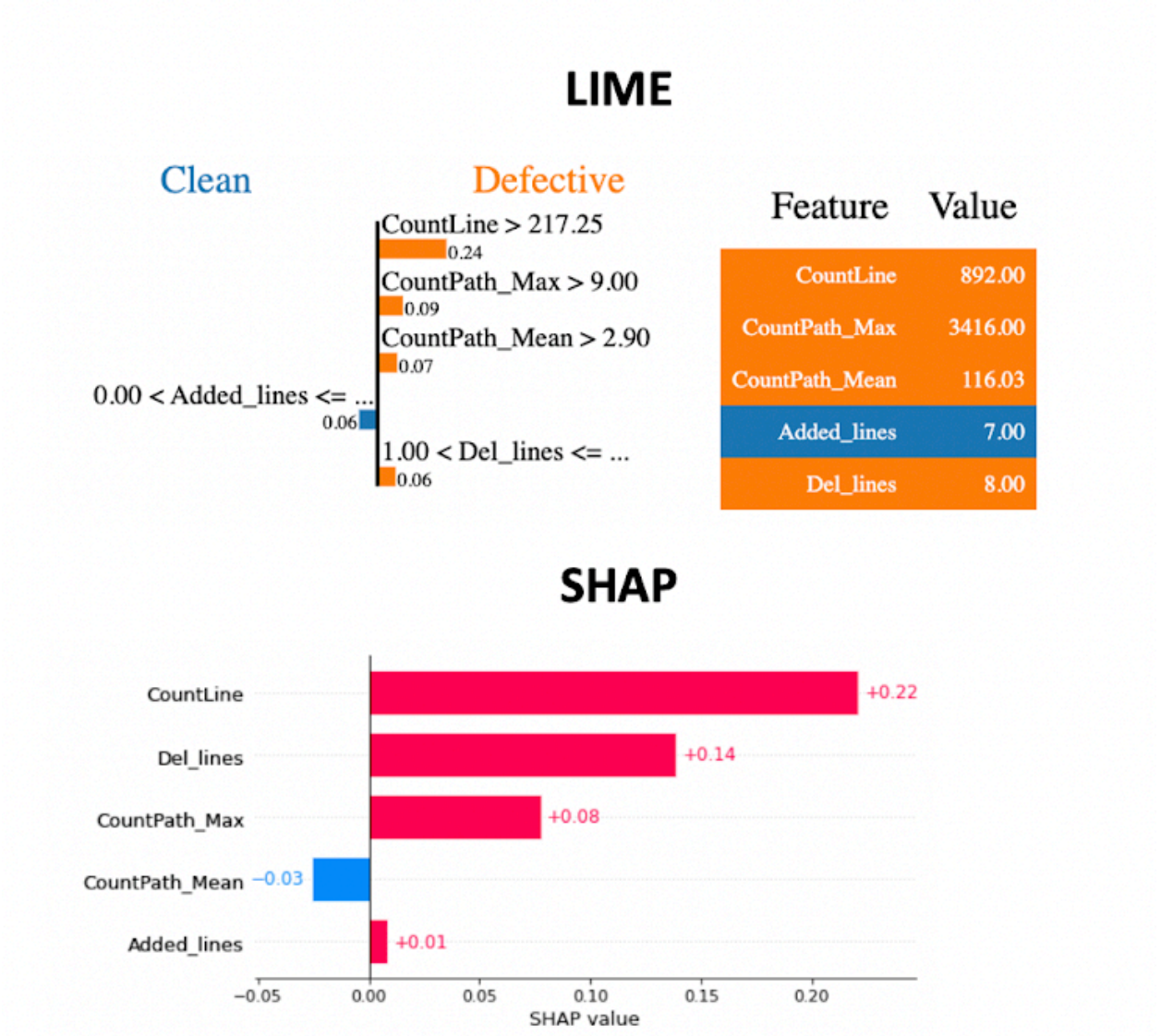
Disagreements between LIME and SHAP

Disagreement: If the features of two explanations are not agreed with the other, is called disagreement. The below example is a real example of disagreement. We have considered three metrics to compute the disagreements. They are Feature Agreement (FA), Sign Agreement (SA), and Rank Agreement (RA).

Feature Agreement (FA):

If we consider Feature Agreement (FA) for LIME and SHAP in the below figure, we can notice that the same five features are present for both of the techniques (LIME and SHAP). That means, all the five features are available with both of the techniques.

So, we can say FA=1.



8. Is it understandable of Feature Agreement (FA)? *

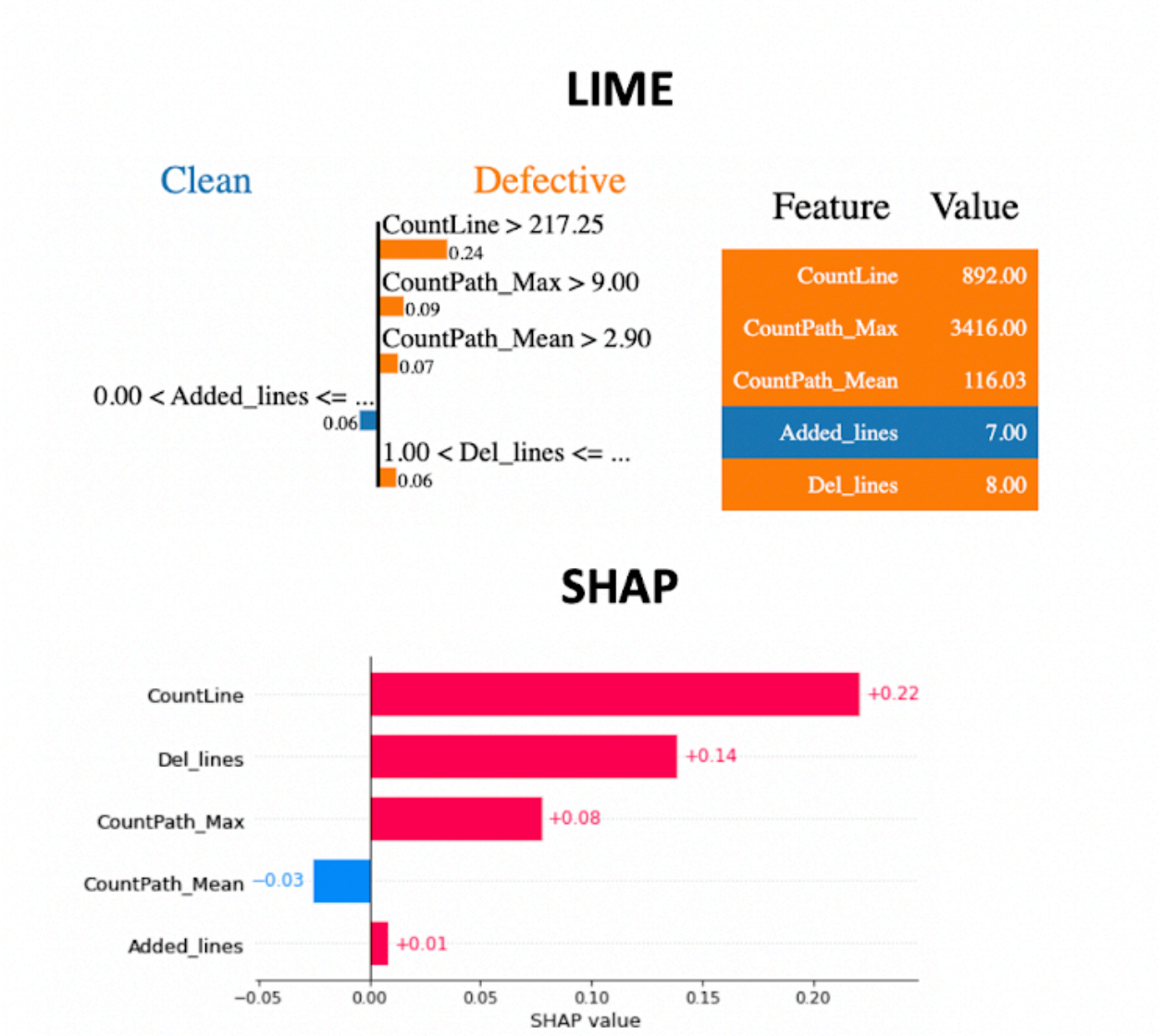
Mark only one oval.

- Yes
- No
- Maybe

Sign Agreement (SA):

If we consider Sign Agreement (SA) for LIME and SHAP in the below figure, we can notice that three features have same sign direction for the both cases (LIME and SHAP). More specifically; "CountLine", "Del_lines", "CountPath_Max" has positive attribution to be defective. But, "CountPath_Mean" and "Added_lines" have opposite sign direction for LIME and SHAP. So, we can say that among five features three features agree with each and other. That's why SA will be (3/5).

So, SA = 3/5



9. Is it understandable of Sign Agreement (SA)? *

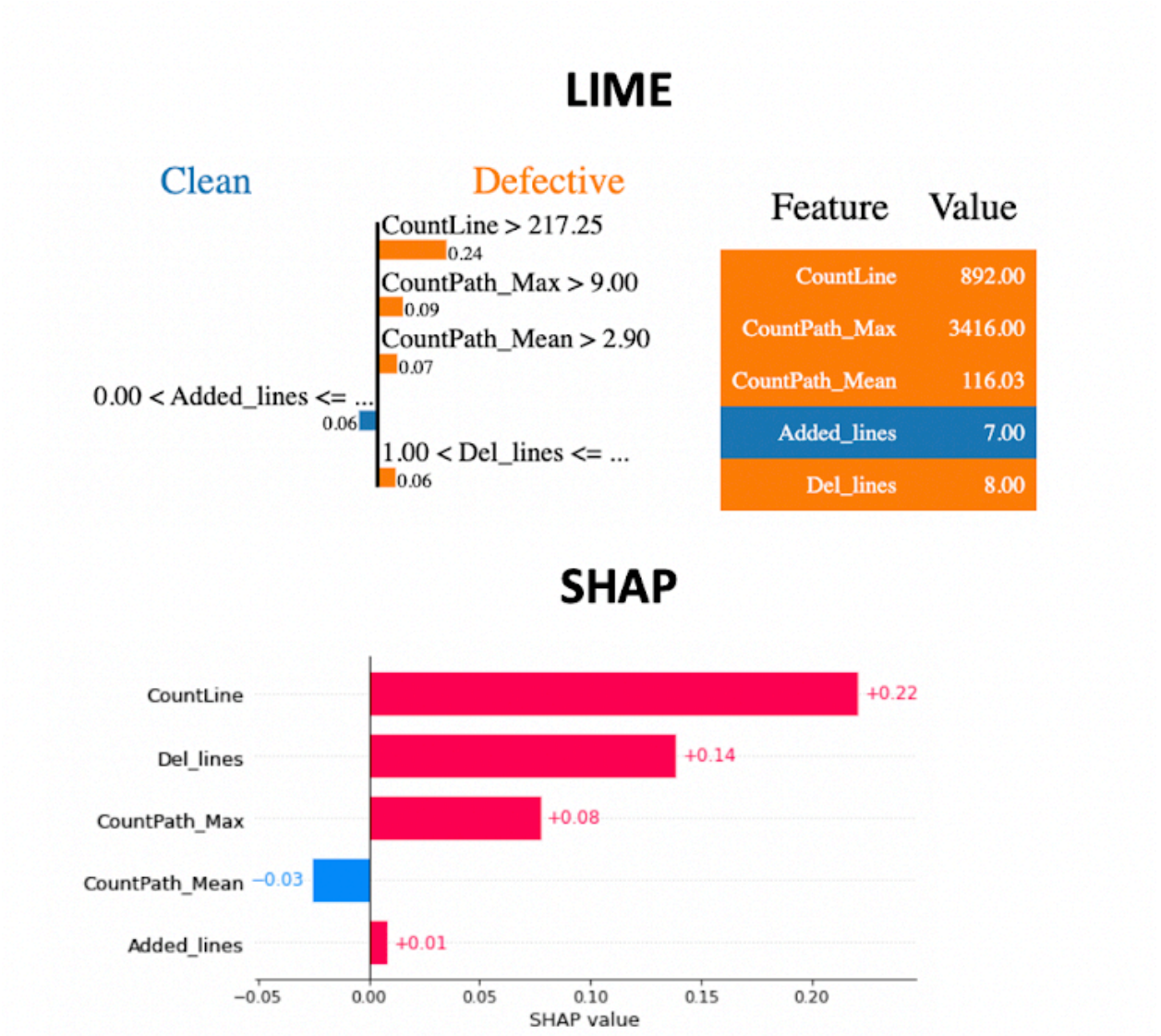
Mark only one oval.

- Yes
- No
- Maybe

Rank Agreement (RA):

Rank agreement means order of ranking. If we consider Rank Agreement (RA) for LIME and SHAP in the below figure, we can notice that only the feature "CountLine" sharing the same ranking for both of the cases (LIME and SHAP). The feature "CountLine" sharing the first rank for both explainers. But, the rest of the features does not have same ranking for both cases. So, we can say that among five features, only one features agree with each and other. That's why RA will be (1/5).

So, RA = 1/5



10. Is it understandable of Rank Agreement (RA)? *

Mark only one oval.

- Yes
- No
- Maybe

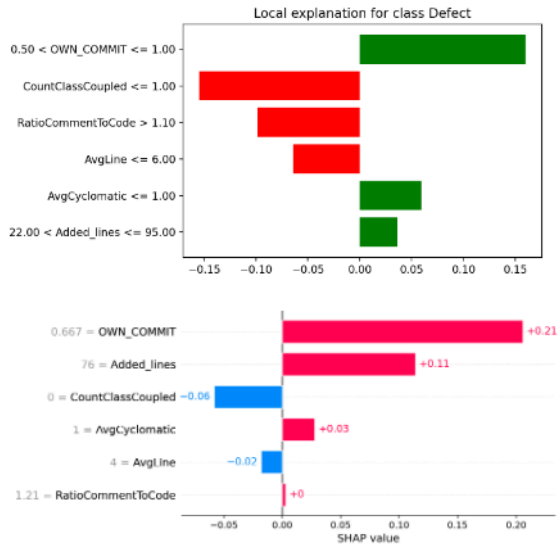
Scenario: (Complete example for three metrics of disagreements)

Examples of disagreements between LIME and SHAP explainers when attempting to understand why two files are predicted as defect.

(Left figure), we observe the top-3 most important features are not the same between the both methods.

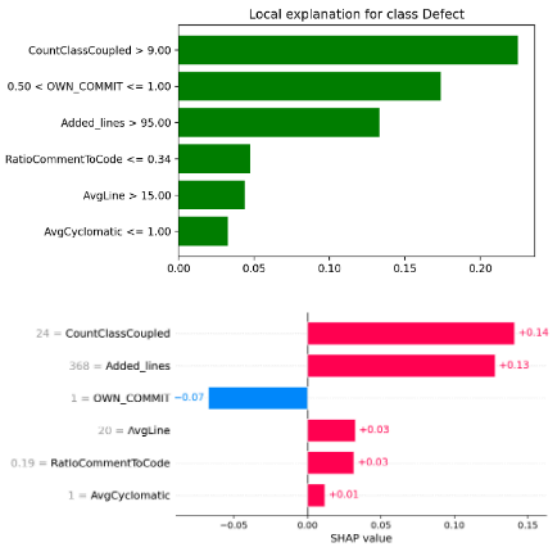
(Right figure) the top-3 features of LIME and SHAP contain the same features, however two of these features do not have the same rank, and the sign. Where, the feature OWN_COMMIT is positive in LIME and negative in SHAP.

Example 1
The Top-3 features are not the same



Metrics : FA=2/3, SA=2/3, RA=1/3

Example 2
The Top-3 features are the same but their rank and sign disagree



Metrics : FA=1,SA=2/3, RA=1/3

11. Does the Random Forest (RF) method consistently beat other machine learning models before
and after data balancing, justifying its preference for use in feature selection strategies to choose the best
approach?

Table 3
Before Data Balancing

Dataset	#Files	Model	Accuracy (Before CV)	Precision	Recall	F1 Score	AUC	Accuracy (After CV)	Selected Model
jruby-1.1	511	XGB	0.95	0.39	0.41	0.40	0.77	0.95	XGB
jruby-1.4.0	684	RF	0.91	0.19	0.21	0.19	0.84	0.94	XGB
jruby-1.5.0	791	DT	0.94	0.15	0.16	0.15	0.86	0.96	XGB
jruby-1.7.0.preview1	**1129**	**RF**	**0.96**	**0.52**	**0.39**	**0.44**	**0.72**	**0.96**	**RF**
hive-0.9.0	991	RF	0.96	0.34	0.29	0.30	0.80	0.95	RF
hive-0.10.0	1092	DT	0.95	0.48	0.54	0.50	0.85	0.94	RF
hive-0.12.0	**1863**	**RF**	**0.97**	**0.40**	**0.43**	**0.42**	**0.78**	**0.97**	**RF**
hbase-0.94.0	741	RF	0.83	0.12	0.13	0.12	0.82	0.85	RF
hbase-0.95.0	**1168**	**RF**	**0.94**	**0.16**	**0.15**	**0.14**	**0.84**	**0.92**	**RF**
hbase-0.95.2	1283	XGB	0.84	0.21	0.16	0.16	0.82	0.85	XGB
groovy-1_5_7	**529**	**SVM**	**0.96**	**0.96**	**0.96**	**0.96**	**0.50**	**0.97**	**SVM, RF, XGB**
groovy-1_6_BETA_1	574	RF	0.98	0.62	0.70	0.64	0.73	0.96	DT
groovy-1_6_BETA_2	618	RF	0.96	0.29	0.27	0.27	0.82	0.96	XGB
derby-10.2.1.6	1374	DT	0.75	0.12	0.13	0.12	0.85	0.79	XGB
derby-10.3.1.4	1544	RF	0.79	0.13	0.11	0.11	0.87	0.80	RF
derby-10.5.1.1	**1893**	**RF**	**0.89**	**0.16**	**0.15**	**0.14**	**0.86**	**0.89**	**RF**
camel-1.4.0	1060	RF	0.94	0.28	0.30	0.29	0.78	0.94	RF
camel-2.9.0	4984	XGB	0.98	0.28	0.29	0.28	0.84	0.98	XGB
camel-2.10.0	5539	XGB	0.98	0.37	0.35	0.34	0.86	0.98	RF, XGB
camel-2.11.0	**6192**	**DT**	**0.99**	**0.44**	**0.41**	**0.42**	**0.85**	**0.99**	**RF, DT**
activemq-5.0.0	1318	RF	0.94	0.27	0.24	0.25	0.83	0.92	RF
activemq-5.1.0	1379	XGB	0.97	0.34	0.29	0.29	0.82	0.97	XGB
activemq-5.2.0	1428	RF	0.95	0.19	0.18	0.17	0.87	0.95	RF
activemq-5.3.0	1656	RF	0.93	0.24	0.25	0.24	0.79	0.92	RF
activemq-5.8.0	**2394**	**RF**	**0.97**	**0.28**	**0.28**	**0.27**	**0.84**	**0.97**	**RF**
wicket-1.3.0-beta2	1234	RF	0.97	0.34	0.36	0.35	0.77	0.97	RF, XGB
wicket-1.3.0-incubating-beta-1	1170	XGB	0.97	0.28	0.31	0.29	0.81	0.97	XGB
wicket-1.5.3	**1804**	**DT**	**0.98**	**0.67**	**0.48**	**0.49**	**0.68**	**0.98**	**RF**
lucene-2.3.0	563	XGB	0.92	0.36	0.34	0.34	0.79	0.90	XGB
lucene-2.9.0	957	RF	0.90	0.35	0.26	0.26	0.82	0.91	RF
lucene-3.0.0	935	RF	0.95	0.48	0.42	0.43	0.78	0.95	RF, XGB
lucene-3.1	**1964**	**XGB**	**0.98**	**0.45**	**0.48**	**0.46**	**0.78**	**0.98**	**XGB**

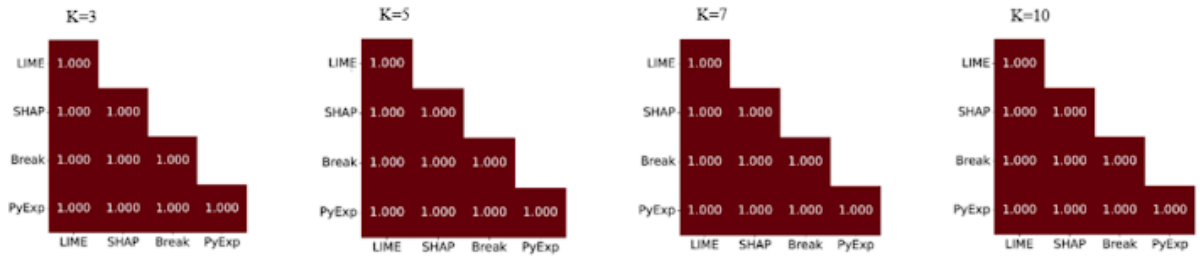
Table 5
After Data Balancing

Dataset	#Files	Model	Accuracy (Before CV)	Precision	Recall	F1 Score	AUC	Accuracy (After CV)	Selected Model
jruby-1.1	4005	RF	0.95	0.43	0.31	0.33	0.77	0.99	RF
jruby-1.4.0	5040	RF	0.93	0.18	0.19	0.18	0.83	1.0	RF, DT, XGB
jruby-1.5.0	5912	XGB	0.93	0.15	0.16	0.15	0.87	1.0	XGB
jruby-1.7.0.preview1	**9603**	**RF**	**0.96**	**0.52**	**0.41**	**0.45**	**0.73**	**1.0**	**RF**
hive-0.9.0	7020	RF	0.96	0.34	0.29	0.30	0.80	1.0	RF
hive-0.10.0	9640	RF	0.95	0.24	0.26	0.25	0.75	0.99	RF
hive-0.12.0	**8575**	**XGB**	**0.97**	**0.38**	**0.41**	**0.40**	**0.80**	**1.0**	**RF, XGB**
hbase-0.94.0	10030	XGB	0.85	0.13	0.18	0.14	0.86	1.0	RF, XGB, DT
hbase-0.95.0	**10800**	**RF**	**0.94**	**0.15**	**0.14**	**0.13**	**0.83**	**0.99**	**RF**
hbase-0.95.2	16439	XGB	0.84	0.20	0.24	0.21	0.84	0.99	XGB
groovy-1_5_7	**4608**	**DT**	**0.98**	**0.15**	**0.12**	**0.13**	**0.82**	**0.98**	**DT, MLP**
groovy-1_6_BETA_1	5190	XGB	0.97	0.59	0.67	0.59	0.73	1.0	XGB
groovy-1_6_BETA_2	5085	RF	0.96	0.31	0.30	0.31	0.79	1.0	RF
derby-10.2.1.6	19173	RF	0.77	0.10	0.11	0.11	0.85	1.0	RF
derby-10.3.1.4	22512	RF	0.78	0.009	0.10	0.009	0.86	1.0	RF
derby-10.5.1.1	**27693**	**RF**	**0.89**	**0.14**	**0.13**	**0.13**	**0.87**	**0.99**	**RF**
camel-1.4.0	6888	RF	0.93	0.34	0.31	0.32	0.77	0.99	RF
camel-2.9.0	48510	XGB	0.98	0.26	0.27	0.26	0.85	1.0	RF
camel-2.10.0	43024	RF	0.98	0.48	0.29	0.34	0.67	0.99	RF
camel-2.11.0	**48488**	**RF**	**0.99**	**0.44**	**0.43**	**0.43**	**0.80**	**1.0**	**RF**
activemq-5.0.0	14495	RF	0.93	0.17	0.18	0.17	0.83	0.99	RF
activemq-5.1.0	12700	RF	0.97	0.41	0.43	0.32	0.82	1.0	XGB
activemq-5.2.0	15204	XGB	0.96	0.26	0.24	0.23	0.88	1.0	XGB
activemq-5.3.0	25024	RF	0.93	0.23	0.24	0.22	0.83	0.98	RF
activemq-5.8.0	**24750**	**RF**	**0.98**	**0.31**	**0.29**	**0.28**	**0.84**	**1.0**	**RF**
wicket-1.3.0-beta2	8022	RF	0.96	0.34	0.34	0.34	0.76	1.0	RF
wicket-1.3.0-incubating-beta-1	8800	XGB	0.97	0.27	0.29	0.28	0.80	1.0	XGB
wicket-1.5.3	**15615**	**RF**	**0.98**	**0.67**	**0.50**	**0.49**	**0.67**	**0.98**	**RF**
lucene-2.3.0	4620	RF	0.92	0.45	0.37	0.38	0.79	1.0	RF
lucene-2.9.0	7670	RF	0.90	0.37	0.31	0.32	0.82	0.99	RF
lucene-3.0.0	7443	RF	0.95	0.48	0.43	0.44	0.76	1.0	RF
lucene-3.1	**15144**	**XGB**	**0.97**	**0.35**	**0.30**	**0.32**	**0.66**	**1.0**	**RF**

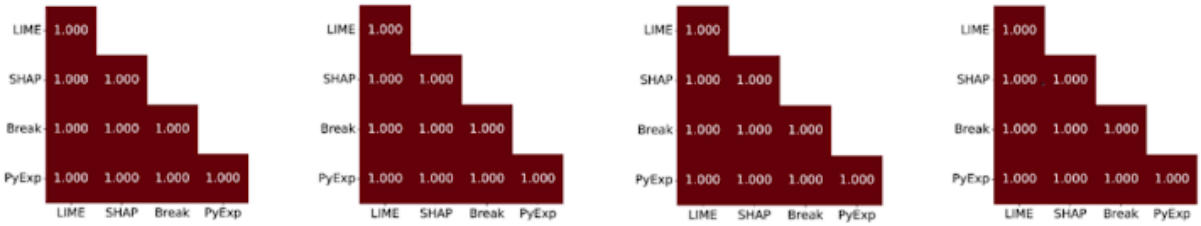
Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Other: _____

12. Feature Agreement: Feature Agreement among four explainers on Chi-square Test and Correlation Matrix with Heatmap for a buggy file. What are you observing - Is the Feature Agreement (FA) same for Chi-square Test and Correlation Matrix with Heatmap? *



Chi-square Test

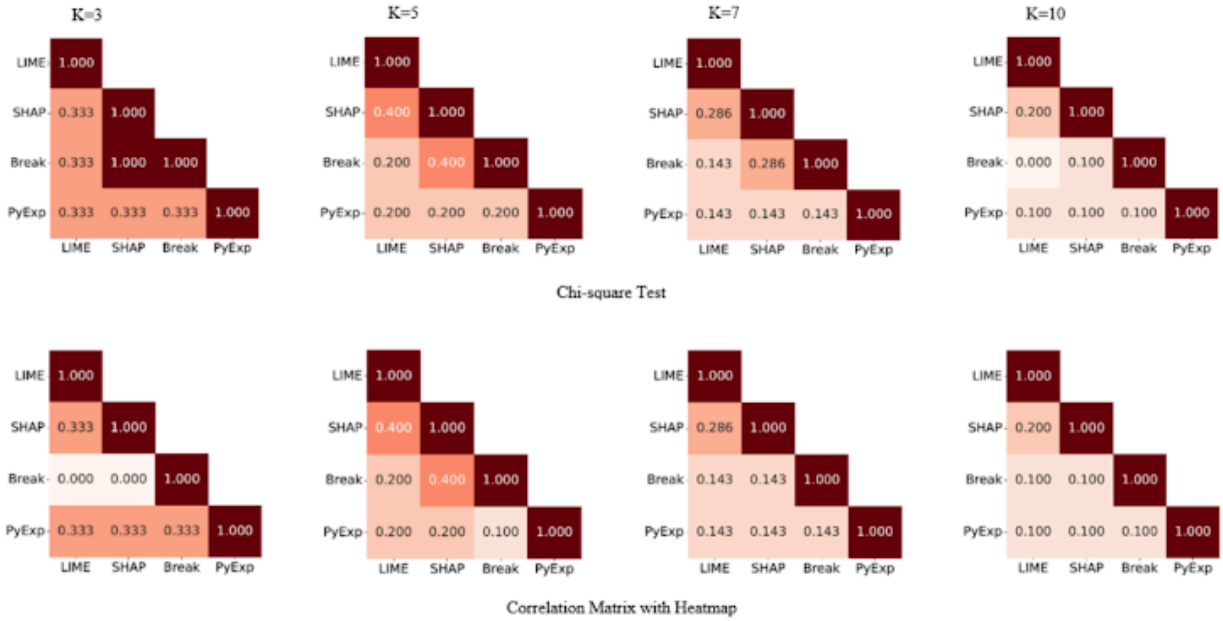


Correlation Matrix with Heatmap

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Other: _____

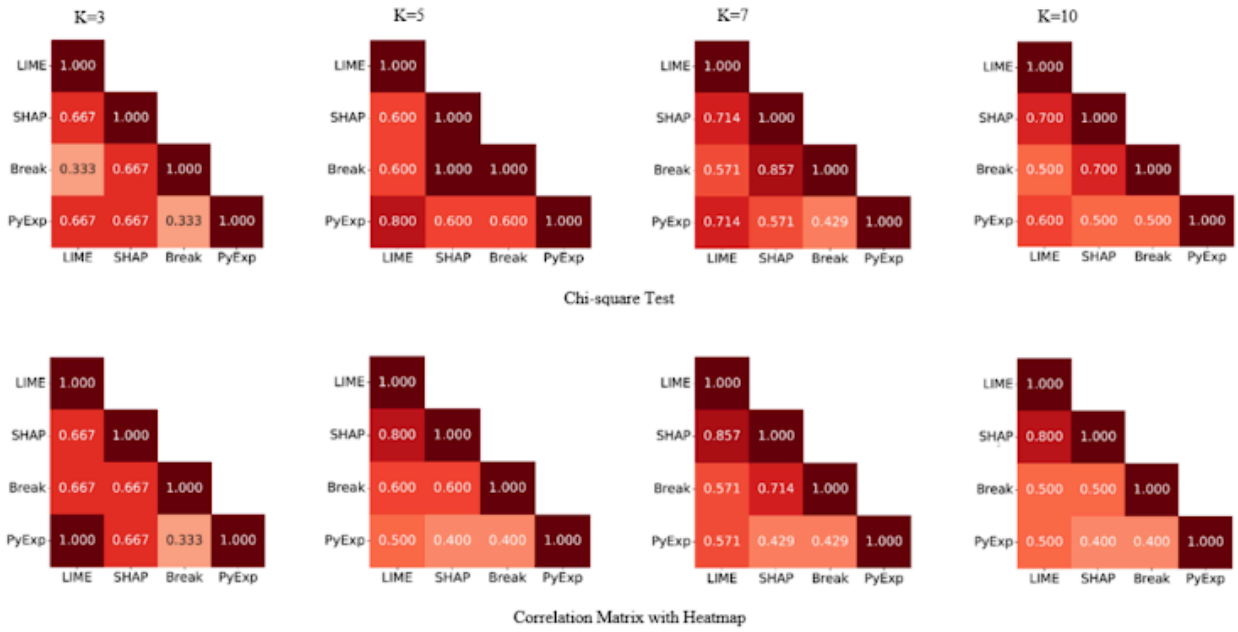
13. Rank Agreement : Rank Agreement among four explainers on Chi-square Test and Correlation Matrix with Heatmap for buggy file. Which technique has comparatively low disagreement?



Mark only one oval.

- ◯ Chi-square Test
- ◯ Correlation Matrix with Heatmap
- ◯ Other: _____

14. Sign Agreement: Sign Agreement among four explainers on Chi-square Test and Correlation Matrix with Heatmap for buggy file. Which technique has comparatively low disagreement?



Mark only one oval.

- ◯ Chi-square Test
- ◯ Correlation Matrix with Heatmap
- ◯ Other: _____

15. How does using feature selection methods like the Chi-square Test and the Correlation Matrix with Heatmap affect the outcomes of disagreements in XAI techniques for predicting software bugs? The Chi-square Test and Correlation Matrix with Heatmap show less disagreement when top-k features are 5 and 3, respectively. To what extent do you think the argument shown above agree or disagree with each other? *

Mark only one oval.

	1	2	3	4	
Con	◯	◯	◯	◯	Completely disagree

16. Which feature selection approach is better at minimizing disagreements between the Chi-square Test and Correlation Matrix with Heatmap when utilizing XAI approaches for software defect prediction? *

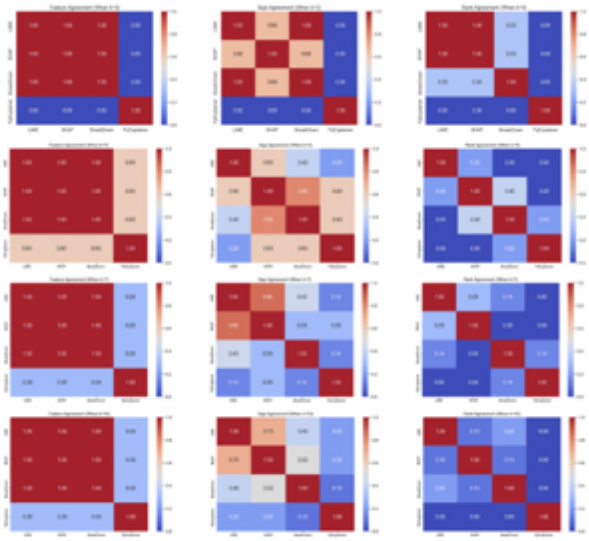


Fig. 11. Disagreements among four explainers on Chi-square Test for buggy file

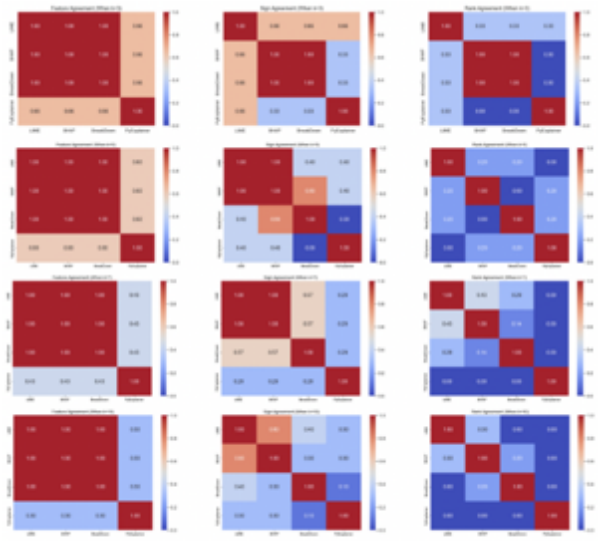


Fig. 12. Disagreements among four explainers on Correlation Matrix with Heatmap for buggy file

Mark only one oval.

- ◯ Chi-square Test
- ◯ Correlation Matrix with Heatmap
- ◯ Other: _____

17. Is the feature engineering, especially in feature selection, truly reduces the disagreement between explanations? *

Mark only one oval.

	1	2	3	4	
Con	◯	◯	◯	◯	Completely disagree

18. **Do you have any suggestions that can help to reduce the disagreement between explainers?** *

Thank you for your time and valuable feedback. Wish you good luck.