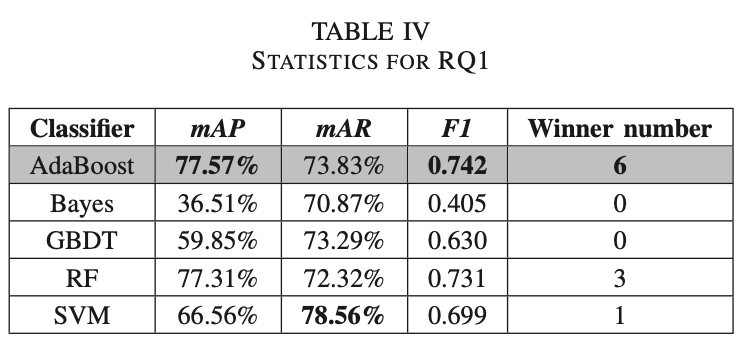
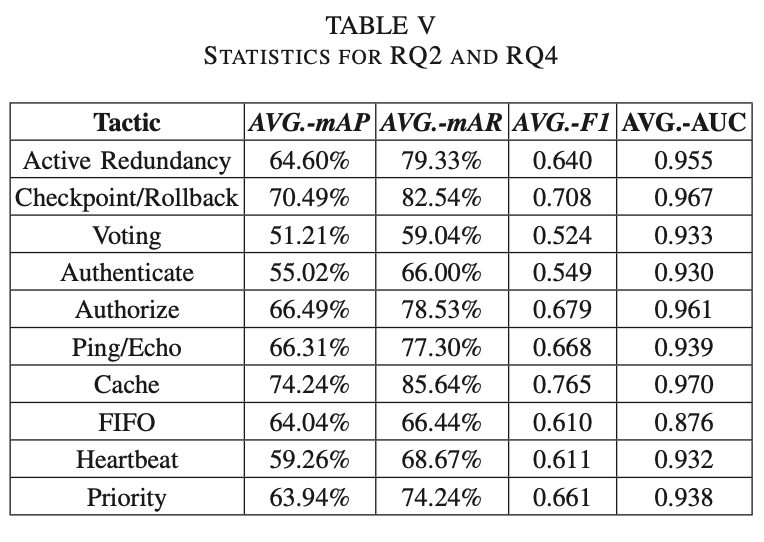
**RQ1 Which classifier shows better performance than others?**



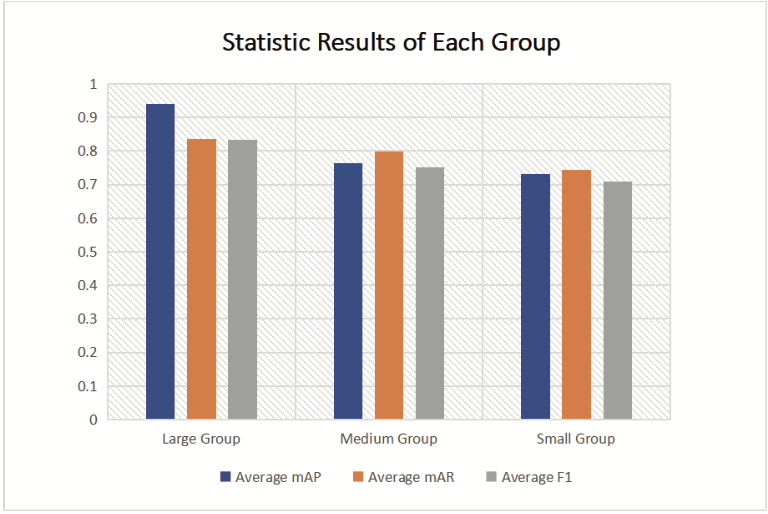
The performance of different classifiers varies on the same tactic dataset. In terms of the average score, the AdaBoost classifier outperforms the other four classifiers. Among the five classifiers, in terms of F1 score, the AdaBoost classifier outperforms on six datasets, the RF classifier out- performs on three datasets, the SVM classifier only wins on the Authenticate dataset. The Bayes classifier and the GBDT classifier do not win on any tactic dataset.

**RQ2: What is the performance of the classifiers on different tactics?**

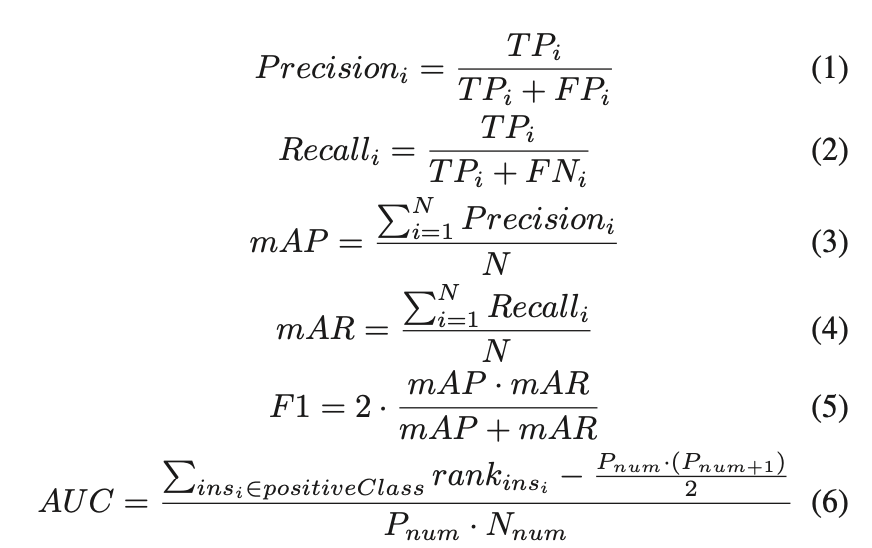


The performance of our tactic method detectors varies with tactics. The detector is more effective on the projects with Cache and Checkpoint/Rollback tactics, of which the average F1 scores are over 0.7. The average F1 scores of the other eight tactics are over 0.61. The detector is less effective on the projects with Voting and Authenticate tactics, of which the average F1 scores are under 0.55.

**RQ3: What is the impact of the dataset size on the performance of the classifiers?**



We calculate the average mAP, mAR, and F1 values of each group and show the statistics. Note that we take the best classifier per tactic in the statistics. It shows that the average mAP, mAR and F1 values increase as the size of the dataset increases. The large group is much better than the other groups. It is not difficult to see that the size of the dataset has a certain positive influence on the performance of our tactic detector. The detection performance improves as the size of the dataset increases. Therefore, if we continue to enlarge the dataset, the detection performance will be improved.



We use the average accuracy rate (mAP) calculated using Precisioni of each class, the average recall rate (mAR) calculated using Recalli, F1 value and the AUC as the evaluation metrics. The calculation formulae of the metrics are provided hereafter, where TPi is the number of true positive data, FPi is the number of false-positive data, FNi is the number of false-negative data, N is the total types of tactic behavior methods, rankinsi is the number of ith sample(Ordered by probability score from large to small), Pnum is the number of positive data and Nnum is the number of negative data.

**RQ4: What is the performance of our approach on class imbalance test data?**

AUC can reflect the ability of classifiers to predict the positive samples and negative samples considering the impact of class imbalance. It shows that the average AUC scores of all tactics except FIFO are over 0.9. It can prove that our approach is effective on the class imbalance data.