Supplemental Material: Feature Attribution Explanation to Detect Harmful Dataset Shift

Anonymous Authors

***Abstract*—Detecting whether a distribution shift has occurred in the data is an important but easily overlooked step when testing, and seemingly small changes in the data distribution may largely affect the performance of the classifier. In this work, we focus on detecting harmful dataset shifts, i.e., shifts that are detrimental to the performance of the classifier. Based on the dataset shift detection framework proposed by Rabanser et al. [1], we use feature attribution explanation (FAE) methods as "dimensionality reduction techniques" to use the gradient information in the model, and a multivariate two-sample detection technique called maximum mean discrepancy (MMD) to detect dataset shifts. The results of experiments using more than twenty shifts on three widely used image datasets show that the feature attribution explanation methods are more effective in identifying harmful data shifts than existing methods. Moreover, experiments on several models of different types/structures show that the ability of our method to identify harmful shifts is virtually indistinguishable across models, i.e., its detection ability is independent of the model used. Code is available at https://anonymous.4open.science/r/FAE-DHDS-861C.**

***Index Terms*—XAI, Dataset Shift, Explainable Artificial Intelligence, Feature Attribution Explanation.**

I. Introduction

This paper presents the supplemental material for the submitted paper "Feature Attribution Explanation to Detect Harmful Dataset Shift".

The paper is organized as follows. Section II shows the trends of p-value with the number of tested samples for each method at each shift individually. Section III shows the trend in p-value as each shift becomes more detrimental to model accuracy across the different tested samples.

II. Detailed Detection Results for Each Harmful Shift

In this section, we show the trend of p-value with the detected sample size for each method at each harmful shift on the MNIST, Fashion-MNIST, and CIFAR10 datasets, one by one.

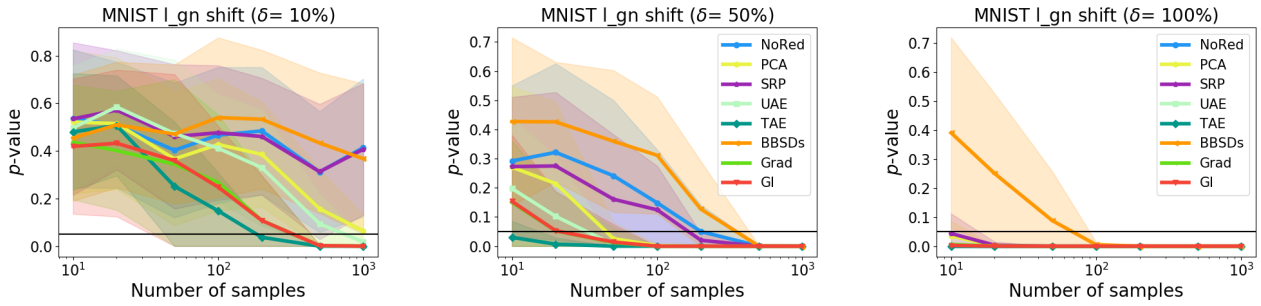

Fig. 1. MNIST p-value trend of harmful l_gn shift.

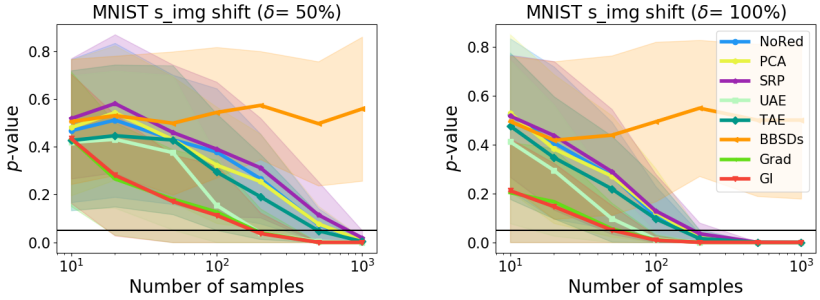

Fig. 2. MNIST p-value trend of harmful s_img shift.

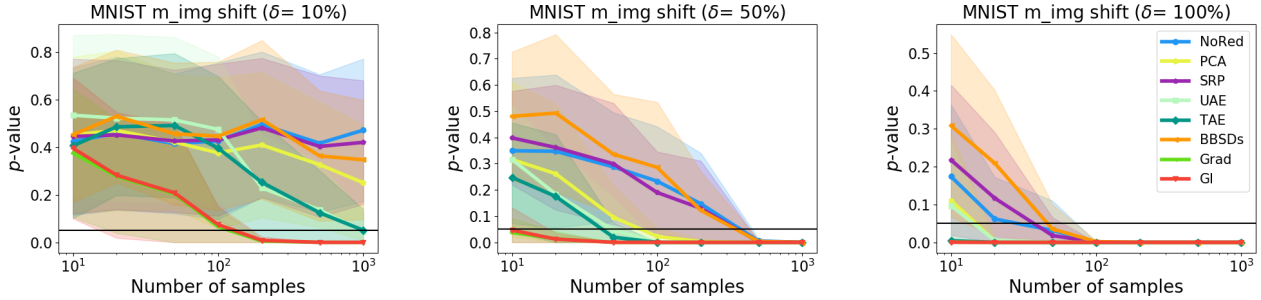

Fig. 3. MNIST p-value trend of harmful m_img shift.

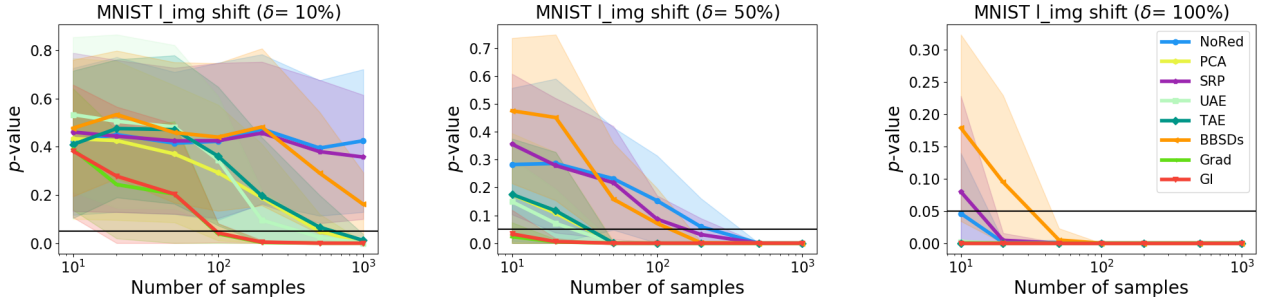

Fig. 4. MNIST p-value trend of harmful l_img shift.

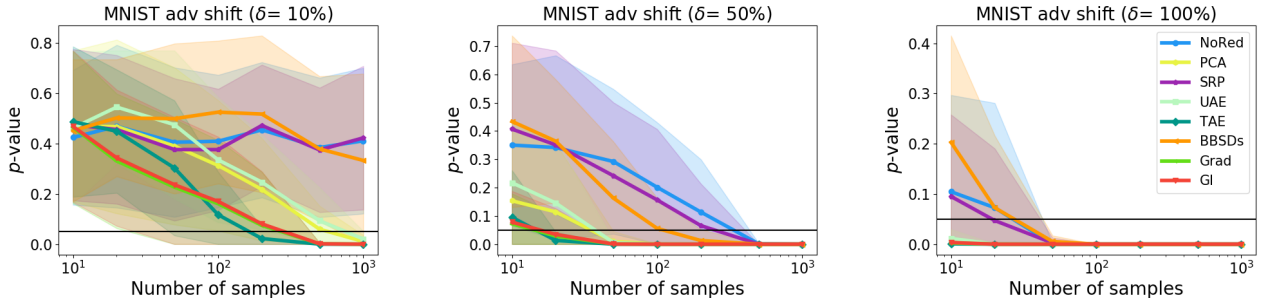

Fig. 5. MNIST p-value trend of harmful adv shift.

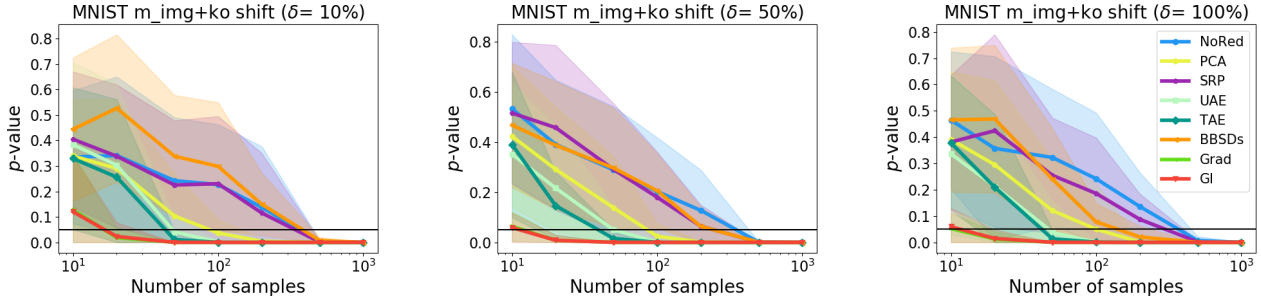

Fig. 6. MNIST p-value trend of harmful $m\_img + ko$ shift.

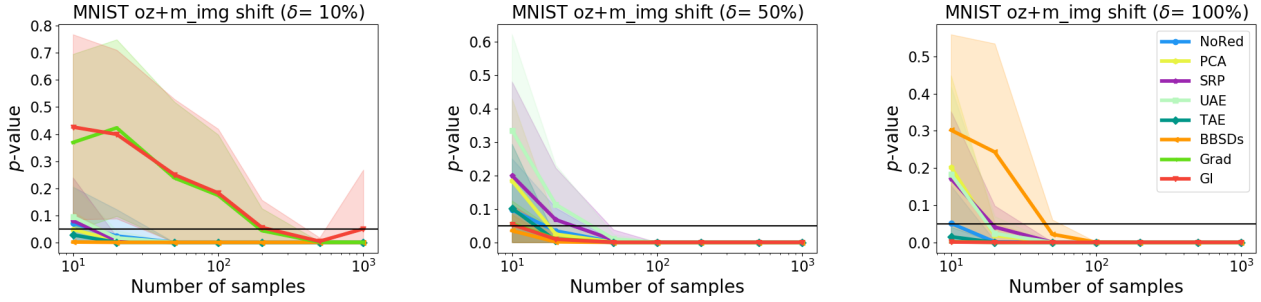

Fig. 7. MNIST p-value trend of harmful $oz + m\_img$ shift.

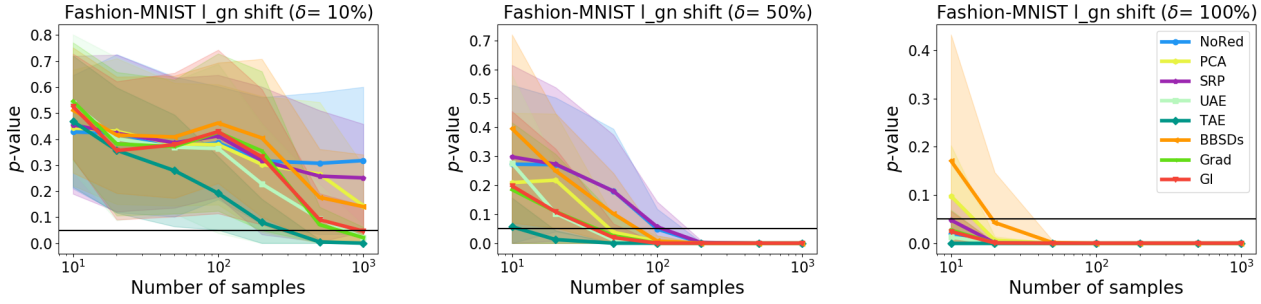

Fig. 8. Fashion-MNIST p-value trend of harmful l_gn shift.

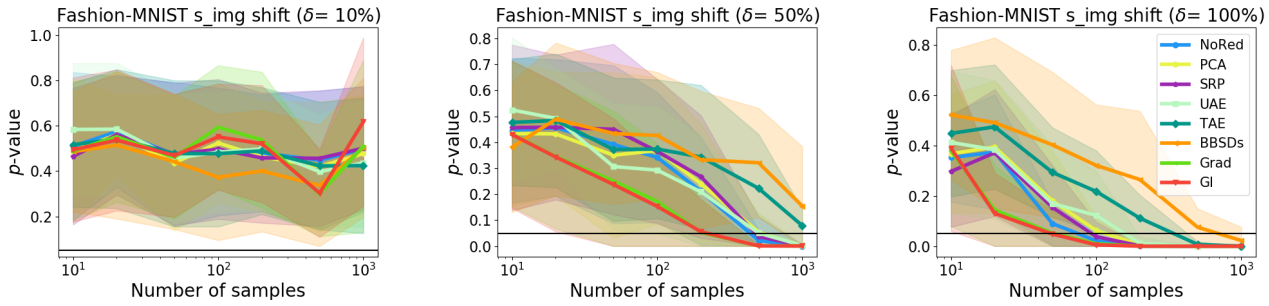

Fig. 9. Fashion-MNIST p-value trend of harmful s_img shift.

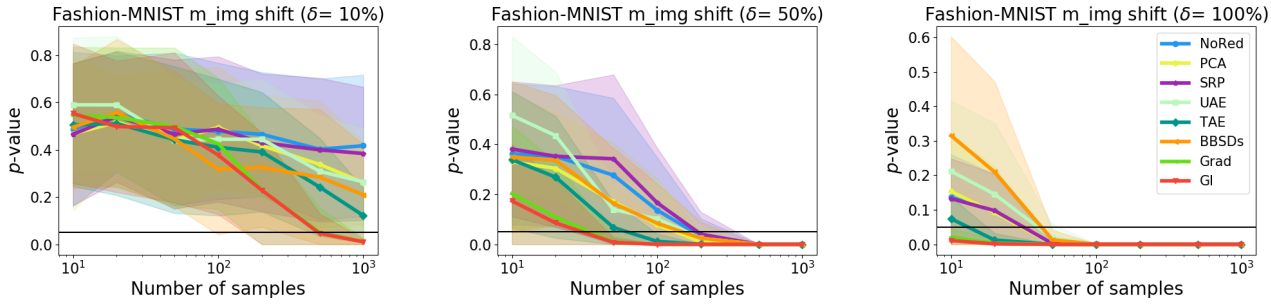

Fig. 10. Fashion-MNIST p-value trend of harmful m_img shift.

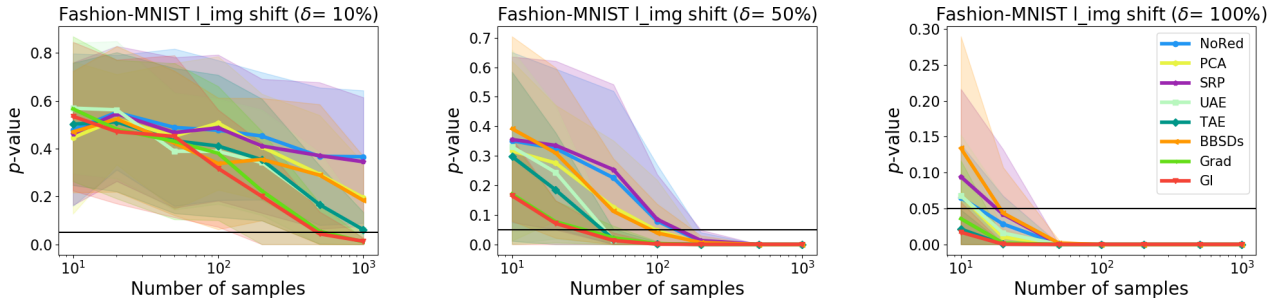

Fig. 11. Fashion-MNIST p-value trend of harmful l_img shift.

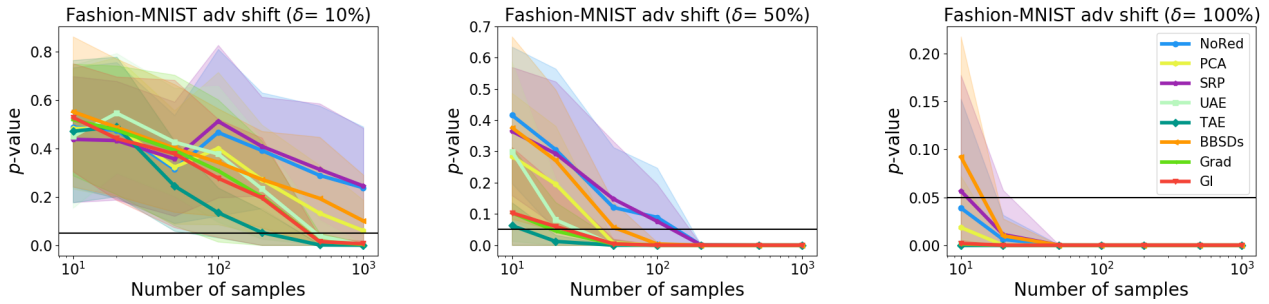

Fig. 12. Fashion-MNIST p-value trend of harmful adv shift.

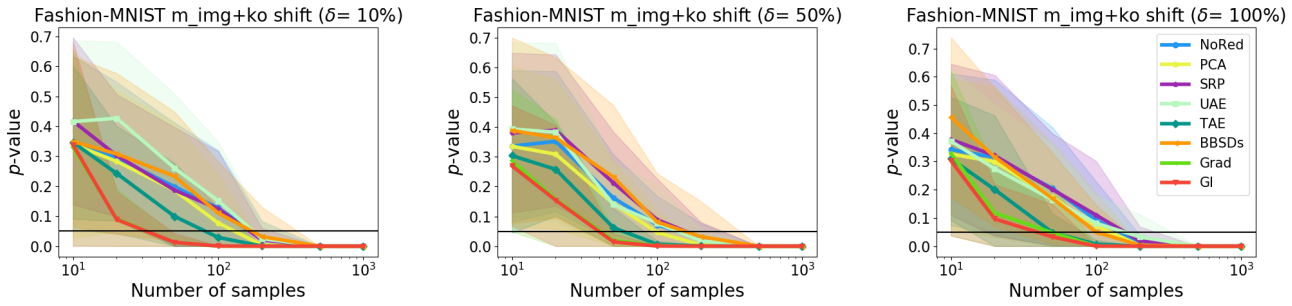

Fig. 13. Fashion-MNIST p-value trend of harmful $m\_img + ko$ shift.

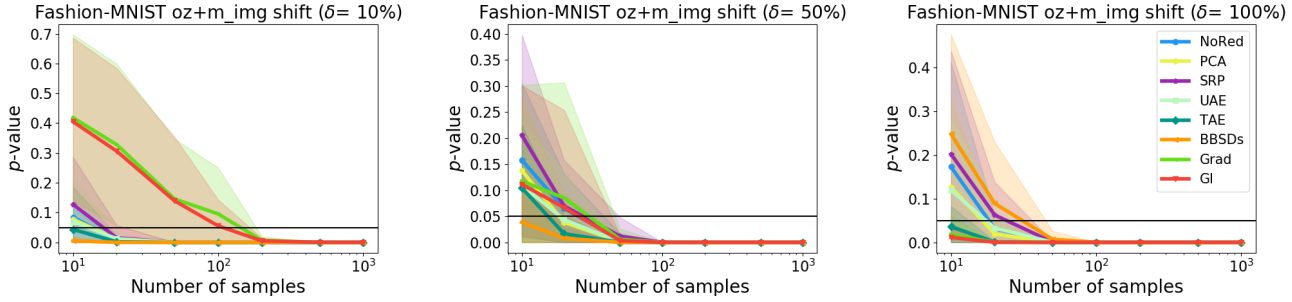

Fig. 14. Fashion-MNIST p-value trend of harmful $oz + m\_img$ shift.

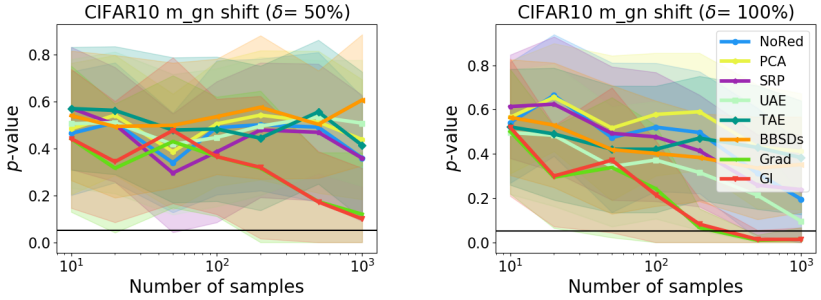

Fig. 15. CIFAR10 p-value trend of harmful m_gn shift.

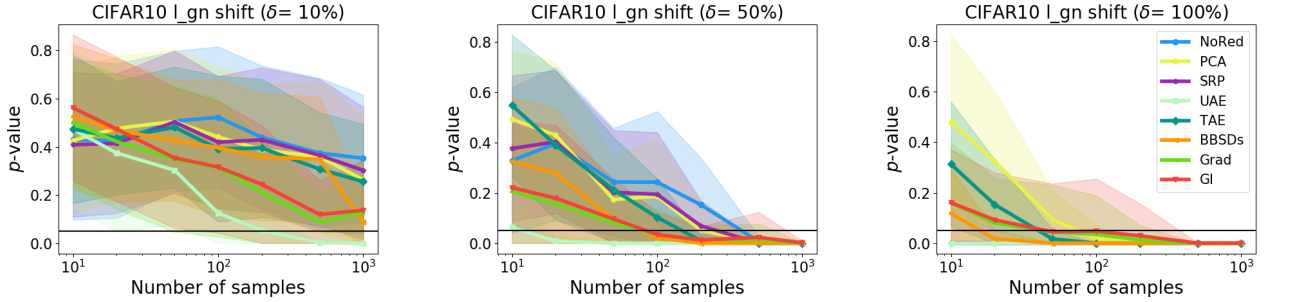

Fig. 16. CIFAR10 p-value trend of harmful l_gn shift.

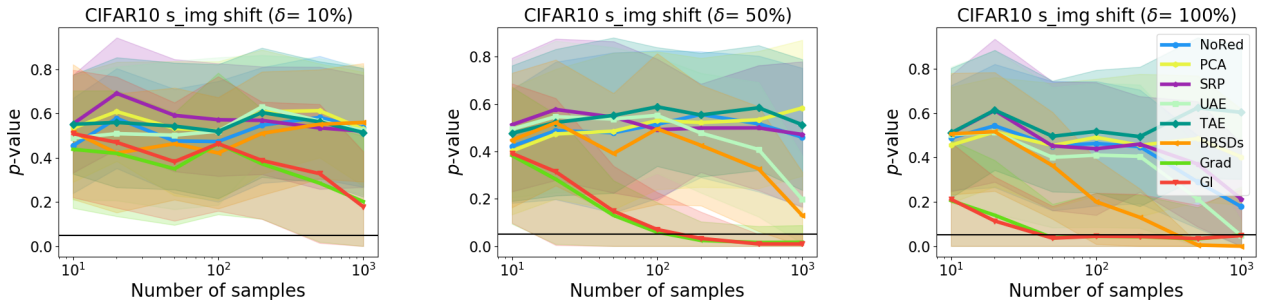

Fig. 17. CIFAR10 p-value trend of harmful s_img shift.

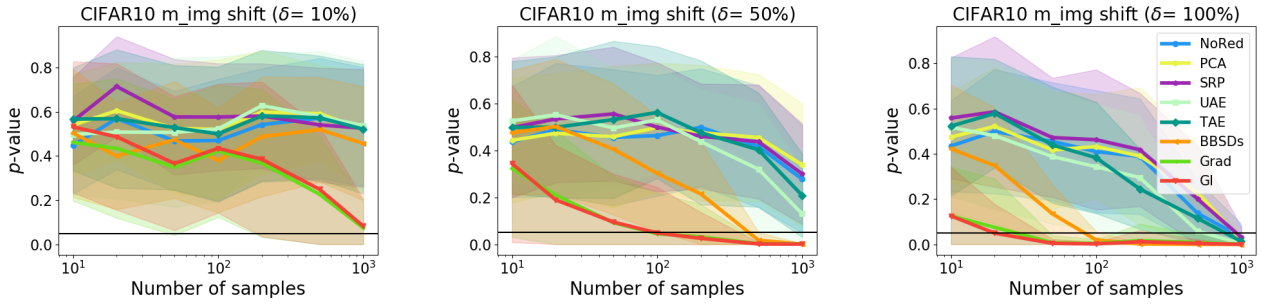

Fig. 18. CIFAR10 p-value trend of harmful m_img shift.

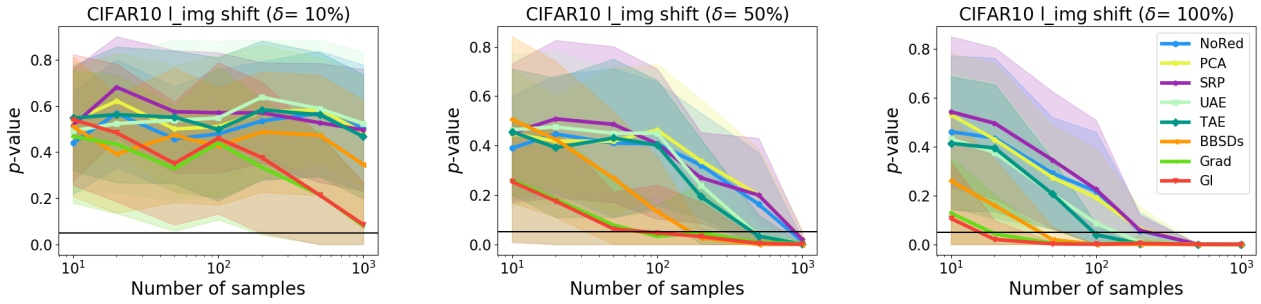

Fig. 19. CIFAR10 p-value trend of harmful l_img shift.

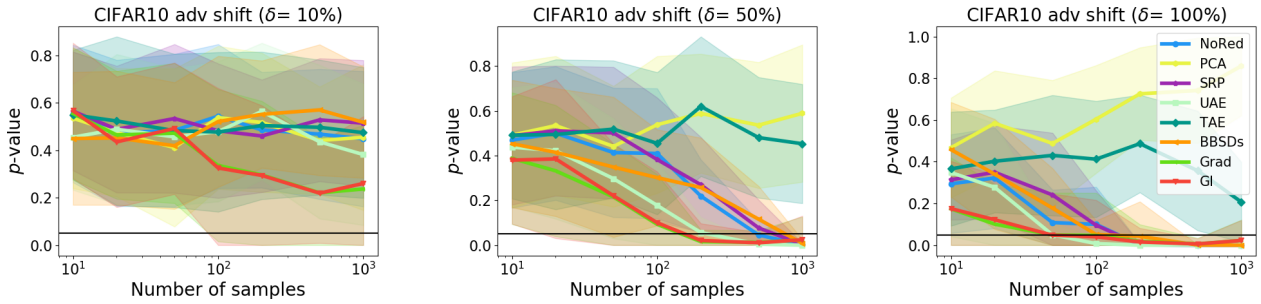

Fig. 20. CIFAR10 p-value trend of harmful adv shift.

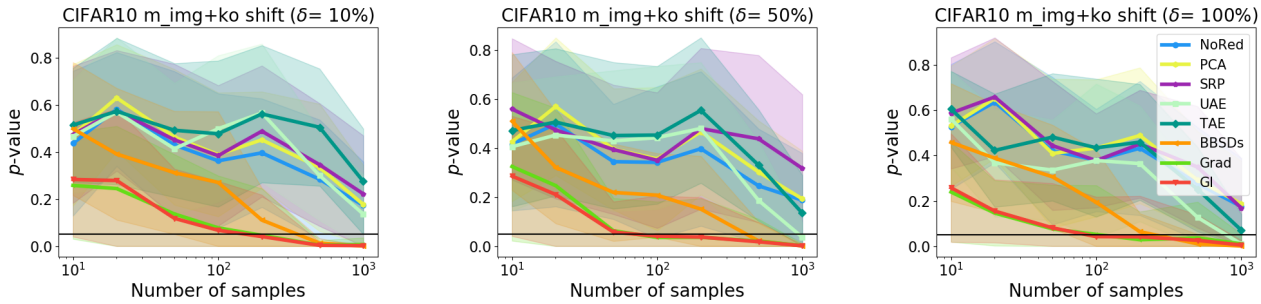

Fig. 21. CIFAR10 p-value trend of harmful $m\_img + ko$ shift.

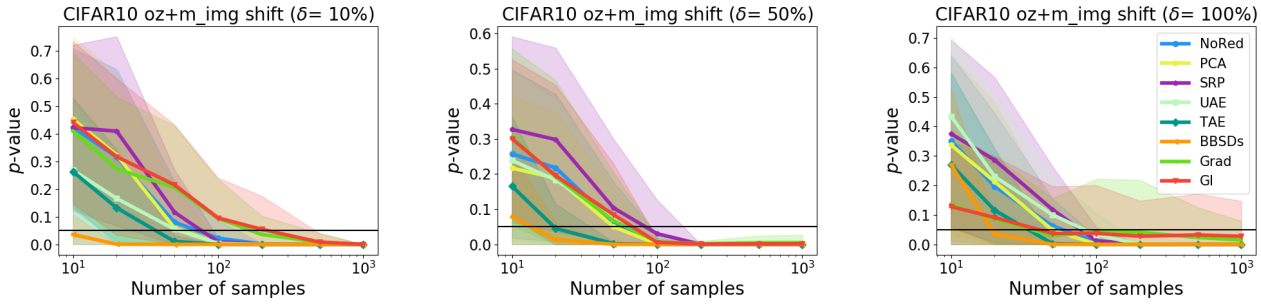

Fig. 22. CIFAR10 p-value trend of harmful $oz + m\_img$ shift.

III. Trend of Each Shift

In this section, we report the trends in p-value as the harmful degree of each shift to model accuracy increases for different detection samples.

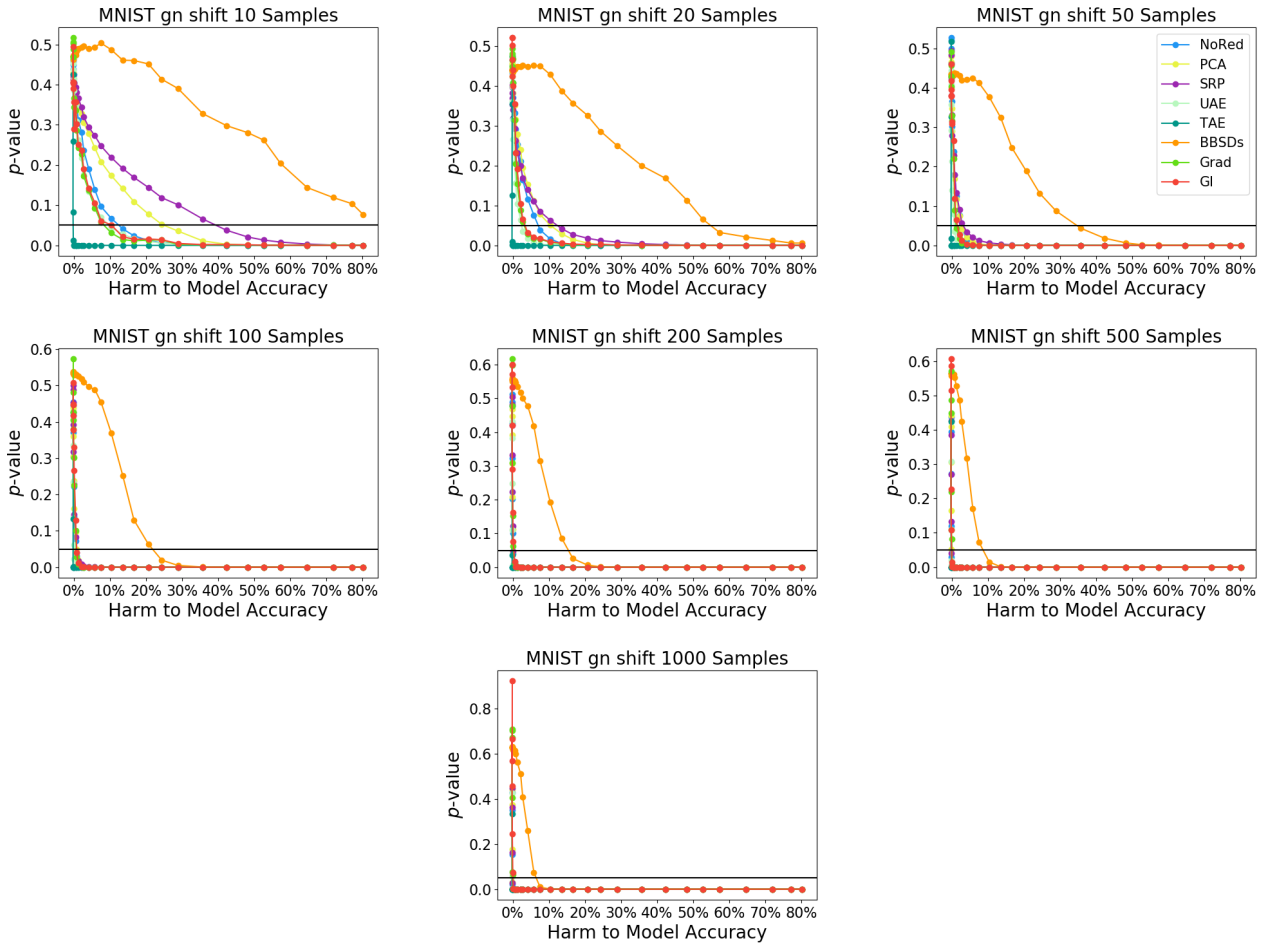

Fig. 23. MNIST p-value trends under different gn shifts

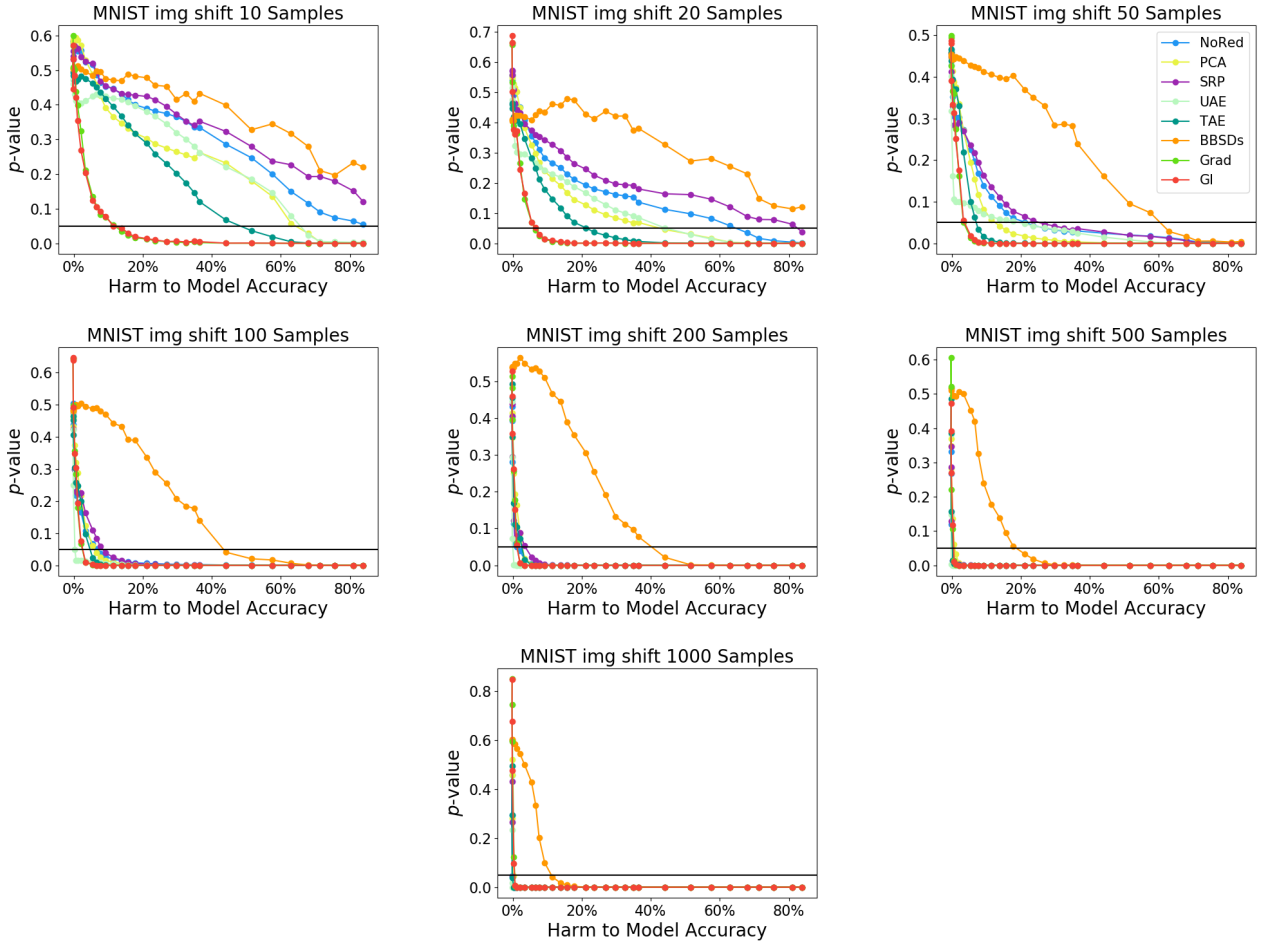

Fig. 24. MNIST p-value trends under different img shifts

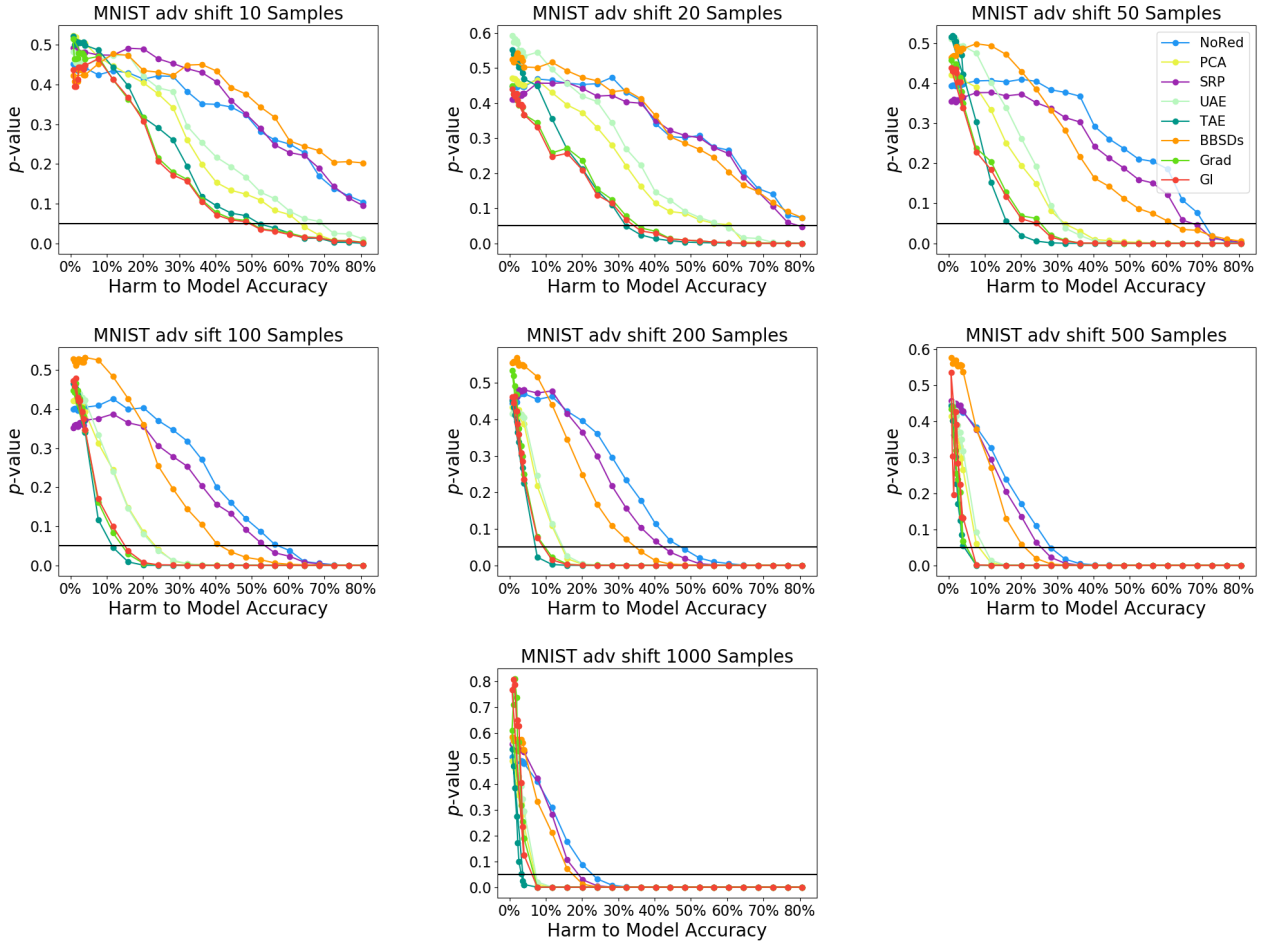

Fig. 25. MNIST p-value trends under different adv shifts

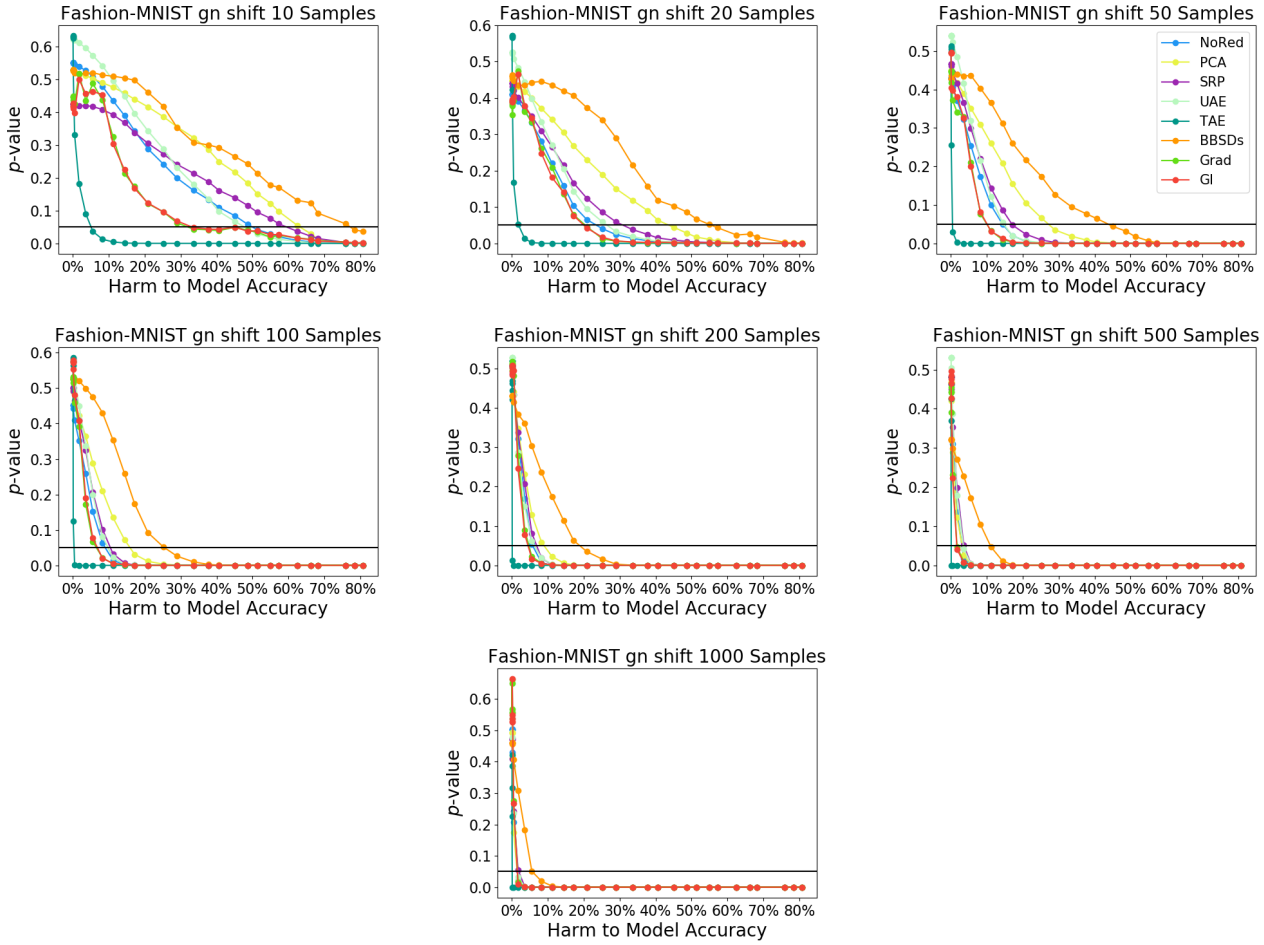

Fig. 26. Fashion-MNIST p-value trends under different gn shifts

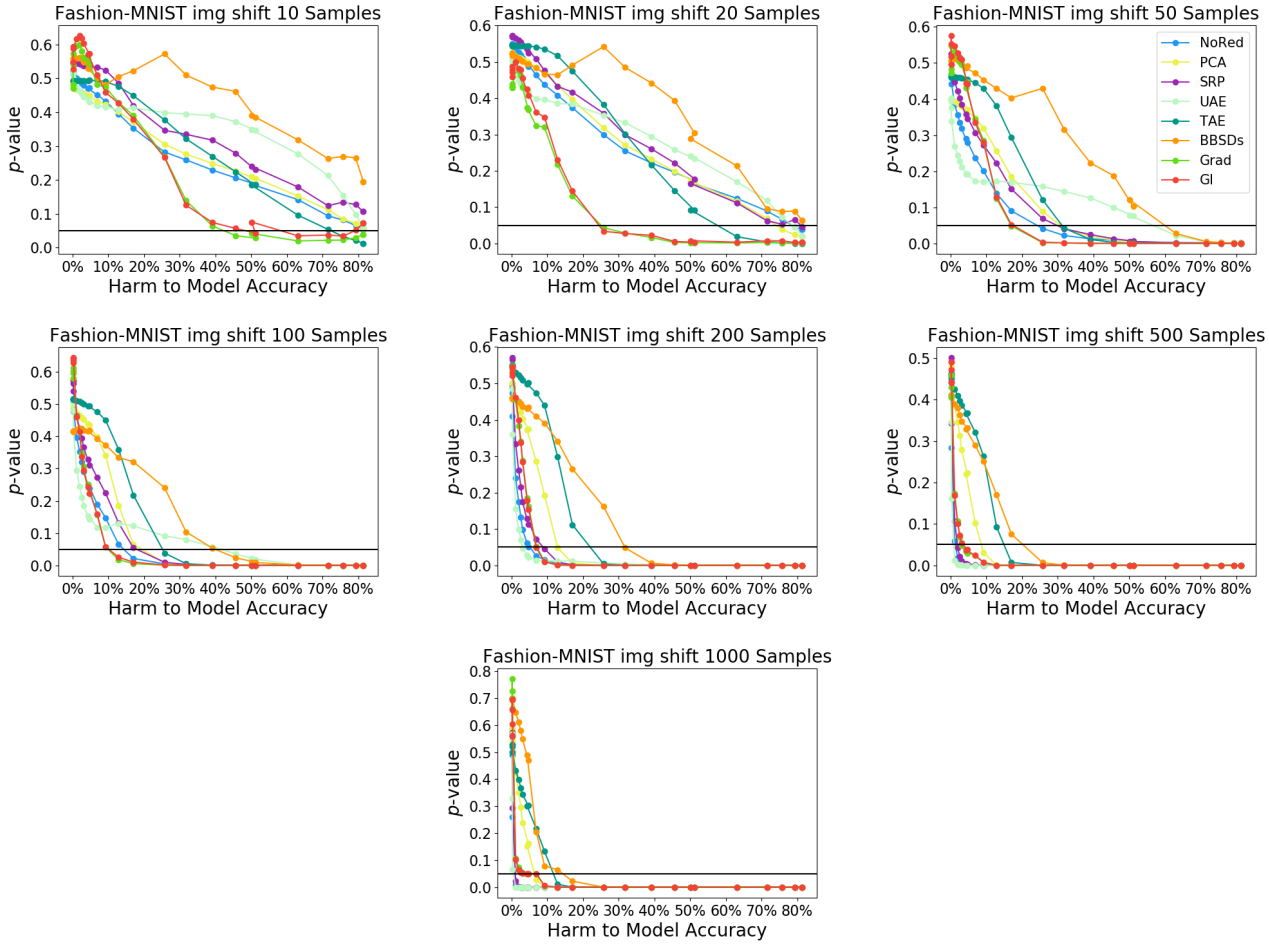

Fig. 27. Fashion-MNIST p-value trends under different img shifts

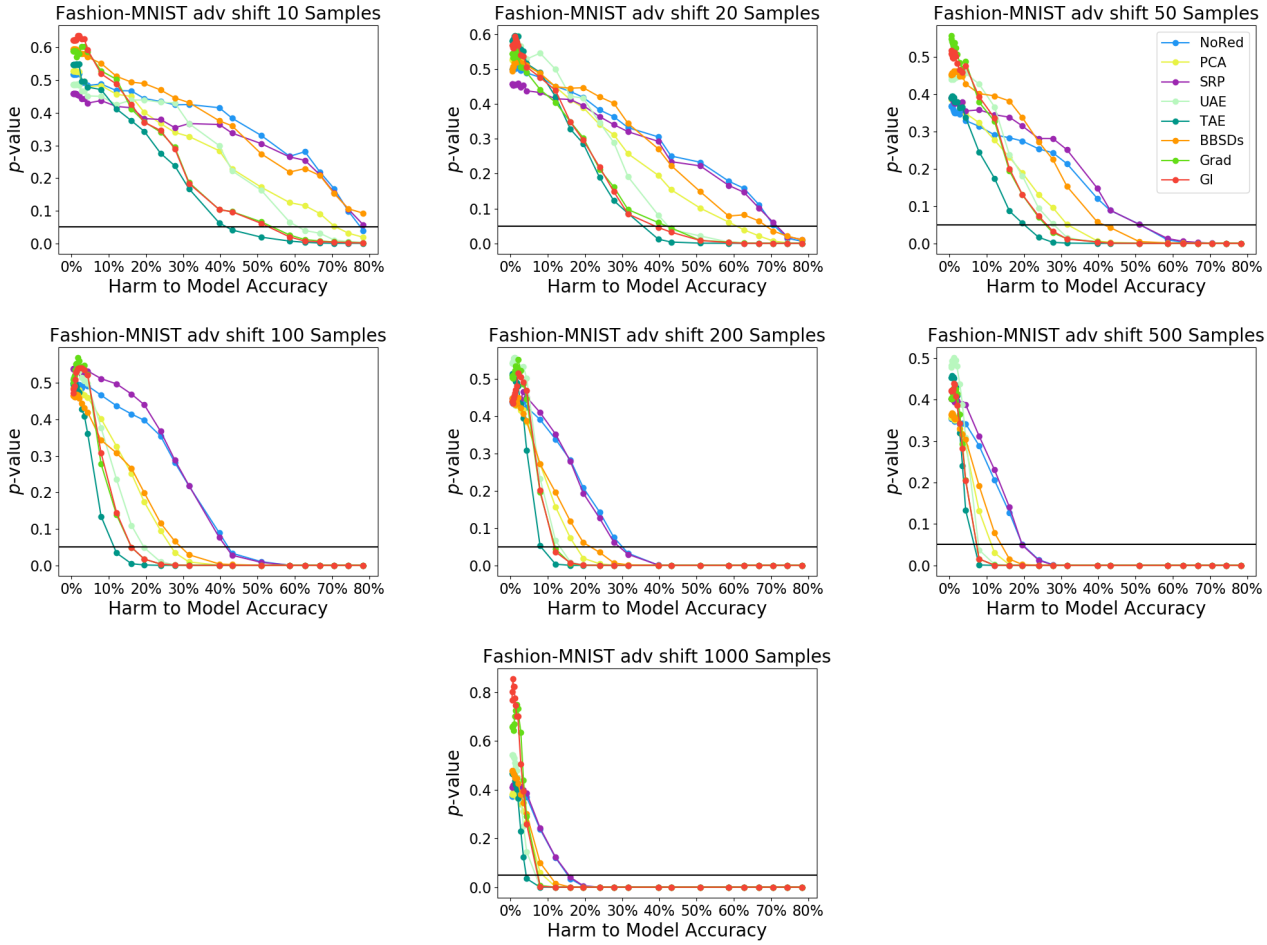

Fig. 28. Fashion-MNIST p-value trends under different adv shifts

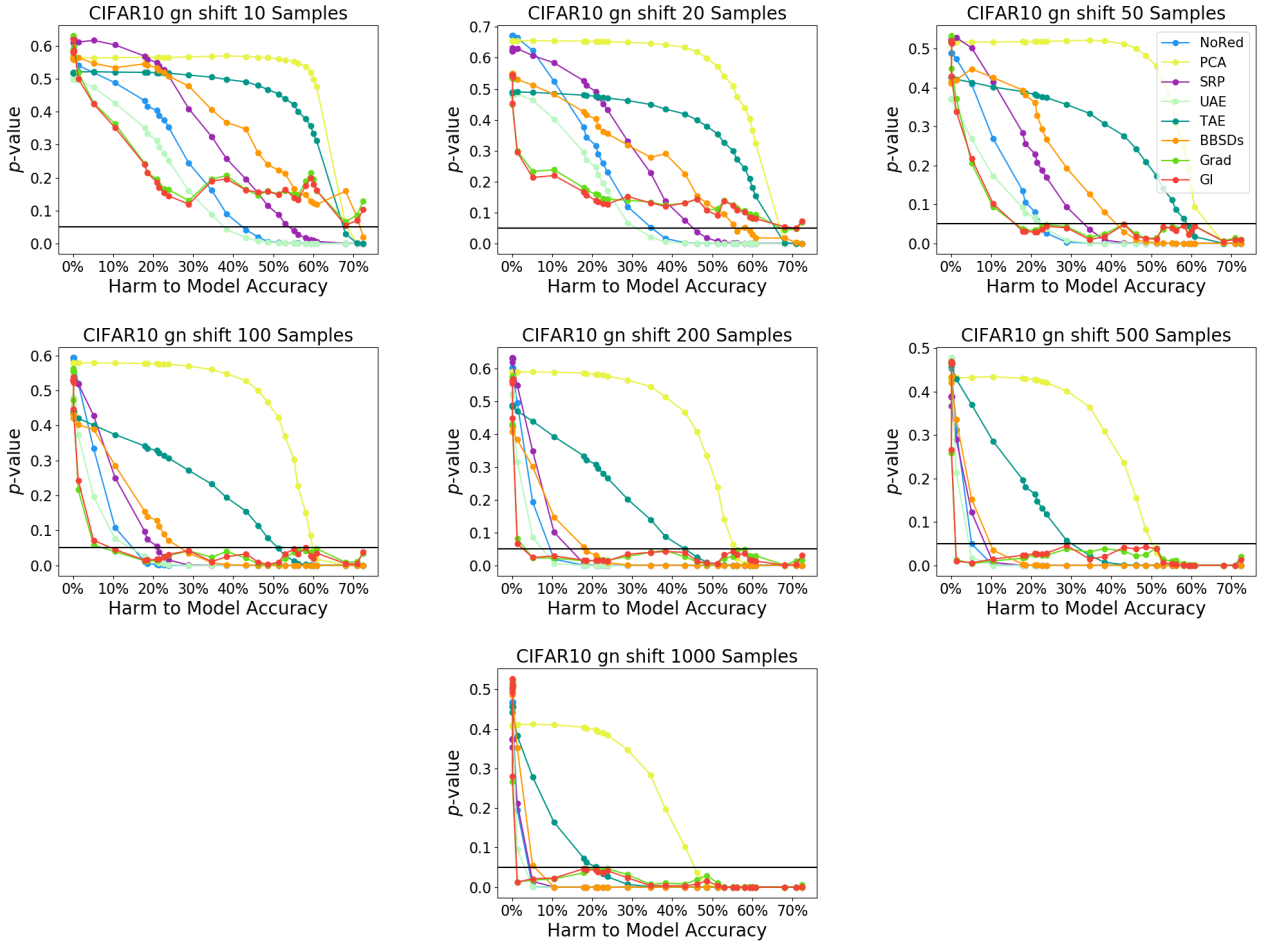

Fig. 29. CIFAR10 p-value trends under different gn shifts

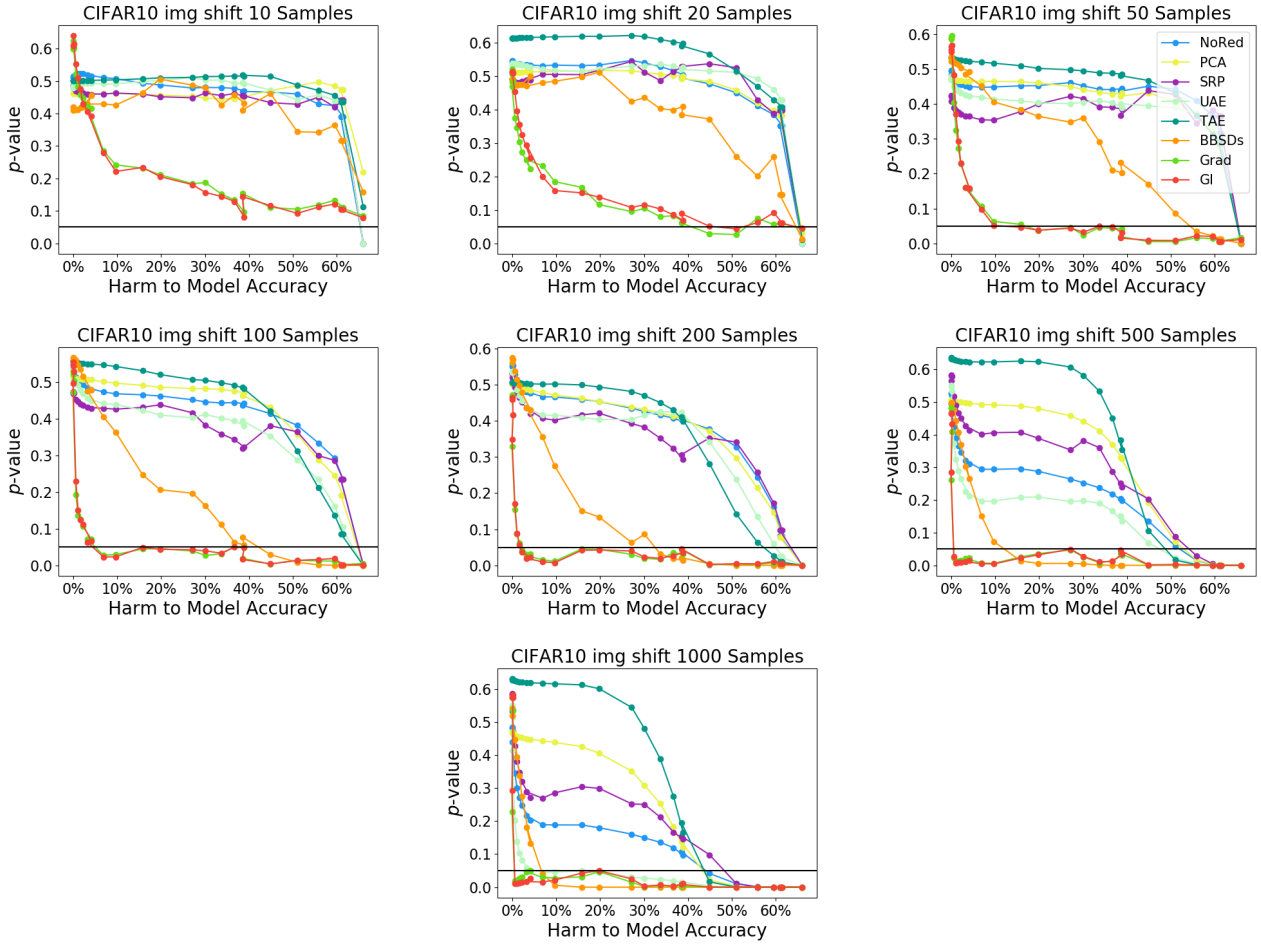

Fig. 30. CIFAR10 p-value trends under different img shifts

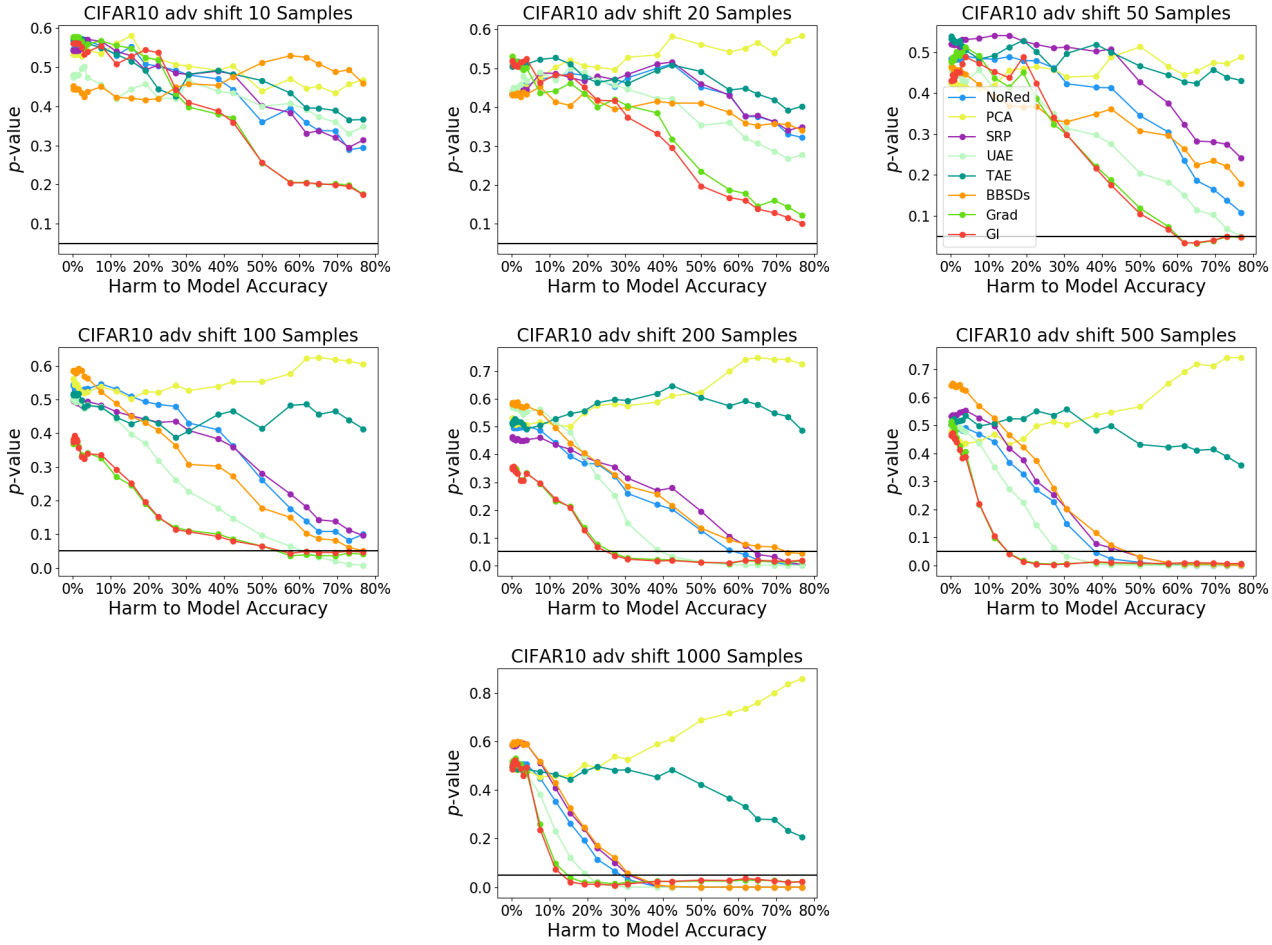

Fig. 31. CIFAR10 p-value trends under different adv shifts

REFERENCES

[1] S. Rabanser, S. Günnemann, and Z. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1396–1408, 2019.