Feature Attribution Explanation to Detect Harmful Dataset Shift

Ziming Wang[1,2], Changwu Huang[2,1], and Xin Yao[2,1]
[1]Research Institute of Trustworthy Autonomous Systems (RITAS),
Southern University of Science and Technology, Shenzhen, China
[2]Guangdong Key Laboratory of Brain-inspired Intelligent Computation,
Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China
wangzm2021@mail.sustech.edu.cn, huangwc3@sustech.edu.cn, xiny@sustech.edu.cn

2023 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-8867-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCNN54540.2023.10191221

***Abstract*—Detecting whether a distribution shift has occurred in the dataset is a critical aspect when implementing machine learning models, as even a small shift in the data distribution may largely affect the performance of a machine learning model and thus cause the deployed model to fail. In this work, we focus on detecting harmful dataset shifts, i.e., shifts that are detrimental to the performance of the machine learning model. The existing methods usually detect whether there is a shift between two datasets according to the following framework: first carrying out dimensionality reduction on the datasets, then determining whether dataset shift exists according to the two-sample statistical test(s) on the reduced datasets. The knowledge contained in the model trained on the dataset is not utilized in the above described dataset shift detection framework. To address this, this paper proposes to take advantage of explainable artificial intelligence (XAI) techniques to exploit the knowledge in trained models when detecting harmful dataset shifts. Specifically, we employ the feature attribution explanation (FAE) method to capture the knowledge in the model and combine it with a widely-used two-sample test method, i.e., maximum mean difference (MMD), to detect harmful dataset shifts. The experimental results on more than twenty different shifts in three widely used image datasets demonstrate that the proposed method is more effective in identifying harmful dataset shifts than existing methods. Moreover, experiments on several different models show that the method is robust and effective over different models, i.e., its detection performance is not sensitive to the model used.**

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

I. Introduction

Machine learning (ML) models often assume that the data encountered after the model is deployed and the training data are independently and identically distributed (i.i.d.). However, in real-world environments, the data distribution often changes after deployment due to various factors, resulting in significant differences from the initial training data distribution and ultimately leading to serious degradation of model performance [1]. The difference between training and test data (or the application data) is known as dataset shift [2]. Therefore, we need methods to detect dataset shifts to ensure that the model works as expected and that its predictions are reliable.

In the classification problem, the dataset shift is defined as $P_{trn}(y,\boldsymbol{x}) \neq P_{tst}(y,\boldsymbol{x})$, where $\boldsymbol{x}$ and y represent the input features and target variable (the class variable), respectively, and $P(y,\boldsymbol{x})$ denotes the joint distribution of $\boldsymbol{x}$ and y [2]. Further, the dataset shift can be classified into three categories, which are described below and also listed in Table I [2]:

- **Covariate Shift**: Changes and differences in the distribution of input variables between training and test data.
- **Label Shift (Prior Probability Shift)**: Changes and differences in the distribution of target variable (i.e., class variable) between training and test data.
- **Concept Shift**: Changes in the relationship between input variables and class variables.

TABLE I
THREE TYPES OF DATASET SHIFTS.

Classification	Definition
Covariate Shift	$P_{trn}(\boldsymbol{x}) \neq P_{tst}(\boldsymbol{x})$ but $P_{trn}(y\|\boldsymbol{x}) = P_{tst}(y\|\boldsymbol{x})$
Label Shift	$P_{trn}(y) \neq P_{tst}(y)$ but $P_{trn}(\boldsymbol{x}\|y) = P_{tst}(\boldsymbol{x}\|y)$
Concept Shift	$P_{trn}(y\|\boldsymbol{x}) \neq P_{tst}(y\|\boldsymbol{x})$ but $P_{trn}(\boldsymbol{x}) = P_{tst}(\boldsymbol{x})$

This study focuses on detecting harmful dataset shifts, especially harmful covariate and label shifts. Harmful shifts refer to shifts that are detrimental to the predictive performance of the model [3], [4]. In other words, the aim is to accurately and efficiently detect dataset shifts that have a detrimental impact on the predictive performance of machine learning models, regardless of their ability to detect shifts that do not negatively affect the model performance [5].

The statistical test compares the distributions of the training and test sets to detect dataset shifts, but usually ignores the information in the model [6]. In contrast, the black box shift detection (BBSD) method [7] which uses the output of the model to detect dataset shifts. However, it only utilizes the output information of the model, but not the knowledge inside the model well. In recent years, explainable artificial intelligence (XAI) has gained much attention for providing insight into the internal decision logic of artificial intelligence (AI) systems [8]–[10], and the feature attribution explanation (FAE)

This work was supported by the National Natural Science Foundation of China (Grant No. 62250710682), the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No.2017ZT07X386), the Shenzhen Science and Technology Program (Grant No.KQTD2016112514355531), and the Research Institute of Trustworthy Autonomous Systems.
Corresponding author: Xin Yao (xiny@sustech.edu.cn)

method has been touted as the most dynamic, widely used, and well-researched XAI method to date [8], [11]. Among them, gradient-based local FAE methods use the gradient information in the model to describe how much each input feature contributes to the output of the model for a given data point [12]. Therefore, in this paper, we use the gradient-based FAE method to exploit knowledge inside the model and combine it with a multivariate two-sample test technique called maximum mean discrepancy (MMD) for detecting harmful dataset shifts. A large number of experiments on MNIST, Fashion-MNIST, and CIFAR-10 datasets show that our method can be used more effectively to detect harmful dataset shifts than other methods, and its detection results are not sensitive to the model used.

The main contributions of this paper are as follows:

- We have studied two gradient-based FAE methods with MMD for detecting harmful dataset shifts. The results of experiments conducted on more than twenty different dataset shifts in three widely used image datasets demonstrate that our method outperforms state-of-the-art methods in identifying harmful dataset shifts.
- Furthermore, using five different types or structures of models, we have found that although our method is model-based, different models do not affect the effectiveness of our method in detecting harmful shifts, i.e., our method is robust and insensitive to the model used.

The remainder of this paper is organized as follows. Section II describes the existing techniques for detecting dataset shifts. Section III presents our proposed method for detecting harmful shifts using the FAE method. The related experimental setup and results are given in Section IV, and conclusions are given in Section V.

II. Related Work

Two-sample hypothesis tests are used to test whether two groups of samples are significantly different or whether they come from the same overall distribution, such as the traditional t-test and the Kolmogorov-Smirnov test. However, such tests require strong parametric assumptions, and they are difficult to analyze with high-dimensional data [13]. Therefore, in recent years, kernel-based non-parametric two-sample tests have become increasingly popular, especially MMD (for more details on MMD, see Section 3.2 of [3]). Further, Liu *et al.* [13] proposed deep kernel MMD, which uses deep kernels to address the limitation of MMD by simple kernels (e.g., Gaussian kernels) in the face of complex distributions. However, the statistical power of kernel-based two-sample test methods decays severely in high-dimensional environments [3], [14]. Therefore, Rabanser *et al.* [3] proposed a pipeline for detecting dataset shifts by combining dimensionality reduction techniques and two-sample test techniques, as shown in Fig. 1 below.

The specific detection process proposed by Rabanser *et al.* [3] for detecting dataset shifts is as follows: (1) the two datasets $\boldsymbol{X}$ and $\boldsymbol{X}^{'}$ are dimensionalized by the dimensionality reduction technique; (2) the distance between the distributions represented by the two datasets after dimensionality reduction is calculated by the two-sample test; (3) the probability that the samples in the two datasets come from the same distribution is judged based on the distance. In addition, the BBSD method proposed by Lipton *et al.* [7] uses the output of the classifier for detecting label shifts. And the experiments in the literature [3] demonstrate that BBSD can also be used to detect covariate shifts.

Ultimately, the following eight dimensionality reduction methods were considered in the literature [3]: No Reduction (NoRed), Principal Components Analysis (PCA), Sparse Random Projection (SRP), Trained Autoencoder (TAE), Untrained Autoencoder (UAE), BBSD using the softmax output (BBSDs), BBSD using the hard-threshold predictions (BBSDh), and Domain Classifier. Then, the appropriate statistical hypothesis test is selected based on the representation produced by each dimensionality reduction technique (i.e., uni- or multi-dimensional, continuous or discrete), which includes the following four: MMD, Kolmogorov-Smirnov Test+Bonferroni Correction, Chi-Squared Test, and Binomial Test. And according to the corresponding statistical hypothesis test results to determine whether there is a difference between the two datasets.

However, few existing methods make use of the information in the deployed model f. A notable exception is that BBSD [7] makes use of its output information. However, the output information contains a limited amount of information, so this paper uses the FAE method to obtain the gradient information inside the model and uses it in combination with the MMD detection technique for detecting the harmful dataset shifts.

Although a preliminary exploration of using a combination of the Grad*Input (GI) method (an FAE method) and a two-sample test for detecting covariate shifts was proposed [6], only two covariate shifts were considered and no study of different models was considered, it is difficult to draw a definite and solid conclusion from such a study. Therefore, we build on this work by conducting comprehensive experiments combining two FAE methods with the two-sample test on more than twenty different shifts (not only covariate shifts but also label shifts and combinations of both). We find that they do not excel on all types of dataset shifts but only outperform existing methods in detecting shifts that are detrimental to model performance (i.e., the harmful dataset shifts). Furthermore, by comparing two different FAE methods, we further confirm that the effectiveness of the method is indeed due to the use of information in the model rather than in the input data itself. In addition, since the FAE method is model-based, we further explore the impact of different models on our methods, i.e., the robustness of the method.

III. FAE to Detect Harmful Dataset Shifts

In this section, we first clearly define the problem we consider, then describe the concept and definition of the FAE method, and finally present the process used in this paper to detect harmful dataset shifts by combining the FAE method with the two-sample test.

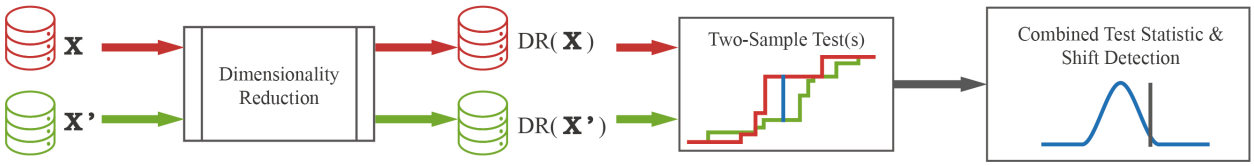

Fig. 1. The dataset shift detection pipeline proposed by Rabanser *et al.* [3]. And the figure is drawn based on [3].

A. Problem Statement and Formulation

Given labeled dataset $D = (\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ with $\boldsymbol{x} \sim p(\boldsymbol{x})$, a neural network classification model f trained on dataset D (i.e., $\hat{y} = f(\boldsymbol{x})$), and unlabeled dataset $\boldsymbol{X}^{'} = \{\boldsymbol{x}_1^{'}, \ldots, \boldsymbol{x}_m^{'}\}$ with $\boldsymbol{x}^{'} \sim q(\boldsymbol{x}^{'})$. Our task is to determine whether there is difference between $p(\boldsymbol{x})$ and $q(\boldsymbol{x}^{'})$. Formally, $H_0 : p(\boldsymbol{x}) = q(\boldsymbol{x}^{'})$ and $H_A : p(\boldsymbol{x}) \neq q(\boldsymbol{x}^{'})$.

In particular, we focus on detecting harmful dataset shifts. That is, we want to detect when $q(\boldsymbol{x}^{'})$ is shifted compared to $p(\boldsymbol{x})$ and the shift may cause performance deterioration of model f.

B. Feature Attribution Explanation (FAE) Method

The FAE methods indicate how much each input feature contributes to the model's output for a given data point [12]. Consider that f is a model that maps an input $\boldsymbol{x} = [x_1, \cdots, x_d]^T \in \mathbb{R}^d$ to an output $f(\boldsymbol{x})$. In the classification problem, $f(\boldsymbol{x})$ is the probability that $\boldsymbol{x}$ belongs to the class that needs to be explained. In general, FAE methods provide a vector $\boldsymbol{\omega} = [\omega_1, \cdots, \omega_d]^T \in \mathbb{R}^d$, where each value ω_i represents the importance of the corresponding feature x_i to the model's prediction $f(\boldsymbol{x})$. Further, FAE methods can be classified as perturbation-based and gradient-based according to how the explanation is generated [15]. In this paper, we use two gradient-based FAE methods, namely Gradient Saliency (Grad) [16] and Grad*Input (GI) [17], as described below.

Gradient Saliency (Grad) [16] calculates the gradient of $f(\boldsymbol{x})$ over $\boldsymbol{x}$ to attribute the importance of each input feature to the model output, where $\boldsymbol{x}$ and f denote the input features and the model, respectively. The explanation of the Grad method is defined as

$$\boldsymbol{\omega}^{Grad}(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}, \tag{1}$$

where $\partial f(\boldsymbol{x})$ represents the derivative of $f(\boldsymbol{x})$.

Grad*Input (GI) [17] was originally proposed as a technique to improve the clarity of the attribution graph. Its explanation result is obtained by multiplying the input features on top of the Grad. The explanation of the GI method is defined as

$$\boldsymbol{\omega}^{GI}(\boldsymbol{x}) = \boldsymbol{x} \cdot \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}. \tag{2}$$

In terms of the focus of this paper, the only difference between the two methods is that the Grad method contains only the gradient information in the model, while the GI method additionally contains the information in the input data.

C. Harmful Dataset Shift Detection based on Feature Attribution Explanation (FAE)

Given a data point, gradient-based FAE uses the gradient information in the model to attribute how each input feature of the data point affects the model decision [12]. In this paper, we combine the FAE method with a two-sample test technique for detecting harmful dataset shifts. Specifically, our method consists of the following three steps:

(1) For the two datasets $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ and $\boldsymbol{X}^{'} = \{\boldsymbol{x}_1^{'}, \ldots, \boldsymbol{x}_k^{'}\}$ used to detect the presence of discrepancies/shifts, their explanation results $\boldsymbol{W}(\boldsymbol{X}) = \{\boldsymbol{\omega}(\boldsymbol{x}_1), \ldots, \boldsymbol{\omega}(\boldsymbol{x}_k)\}$ and $\boldsymbol{W}(\boldsymbol{X}^{'}) = \{\boldsymbol{\omega}(\boldsymbol{x}_1^{'}), \ldots, \boldsymbol{\omega}(\boldsymbol{x}_k^{'})\}$ are calculated by the FAE method, where $\boldsymbol{\omega}(\boldsymbol{x})$ is the explanation result returned by a gradient-based FAE method (such as the Grad and GI methods described in Section III-B) for the data point $\boldsymbol{x}$.

(2) The difference between $\boldsymbol{W}(\boldsymbol{X})$ and $\boldsymbol{W}(\boldsymbol{X}^{'})$ was evaluated by a two-sample test method (e.g., the MMD used in this paper).

(3) The evaluation results are used to determine whether there is a difference between the two datasets $\boldsymbol{X}$ and $\boldsymbol{X}^{'}$, i.e., whether dataset $\boldsymbol{X}^{'}$ has been shifted relative to $\boldsymbol{X}$.

The FAE-based pipeline for harmful dataset shifts detection is shown in Fig. 2 below.

IV. Experimental Studies

In this section, we first present the relevant settings for our experiments, and then comprehensively compare the effectiveness of our method with other methods for detecting harmful shifts by more than twenty different shifts on three datasets, and discuss whether the detection effectiveness of our method is influenced by the model used. The codes used in the experimental studies are available at https://github.com/oddwang/FAE-DHDS.

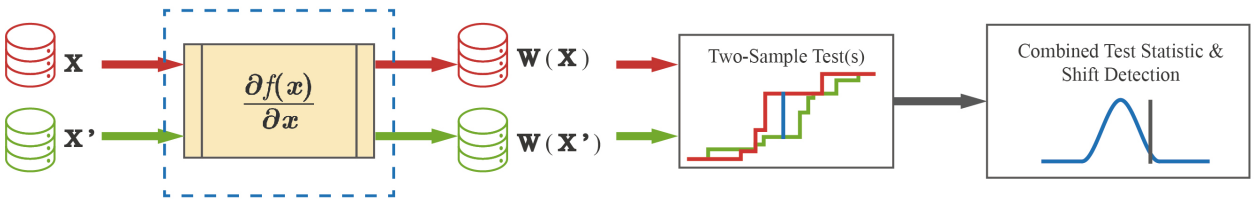

Fig. 2. FAE-based pipeline for harmful dataset shift detection. The Grad method is used as an example of the FAE method.

A. Experimental Setup

Our experiments were conducted on three datasets, MNIST [18], Fashion-MNIST [19], and CIFAR-10 [20]. And the chosen classification model f is mainly ResNet-50 [21] using stochastic gradient descent (SGD) with momentum over 200 epochs with early stopping. In addition, related experiments were conducted on ResNet-18, ResNet-34 [21], DesNet-121 [22], and InceptionV3 [23] using the same parameter settings to explore whether the effectiveness of our method in identifying harmful dataset shifts differs from one model to another.

The methods we compare are the six dimensionality reduction methods used in the literature [3] that are able to use MMD as a statistical hypothesis test: NoRed, PCA, SRP, TAE, UAE, BBSDs [7], which is also consistent with the literature [6]. And this paper uses two gradient-based FAE methods, Grad [16] and GI [17], to combine with the two-sample test technique. Referring to the setup in the literature [3], a convolutional architecture with three convolutional layers and one fully connected layer is used for UAE and TAE (the parameter settings remain the same as before). In addition, for PCA, SRP, UAE, and TAE, we reduce the input to 32 dimensions, which is also consistent with the literature [3].

To evaluate the ability of each method to detect harmful dataset shifts, we randomly divide the dataset into training, validation, and test sets. The training data are only used to construct the model, while various simulated covariate shifts, label shifts, and combinations of both were applied to the test set to construct the shifted test set and perform shift detection with the validation set. Referring to the literature [3], three percentages affecting the test set $\delta \in \{10\%, 50\%, 100\%\}$ are considered for each shift type. In addition, a further subdivision is performed for Gaussian noise and image shifts, considering three different intensities. Specifically, the types of shifts considered are as follows [3]:

- **Gaussian noise (*gn*)**: Gaussian noise with standard deviation $\sigma \in \{1, 10, 100\}$ (denoted as s_gn, m_gn, l_gn) was applied to a certain percentage δ of the test set.
- **Image (*img*)**: A combination of random rotations $\{10°, 40°, 90°\}$, (x, y) axis translation percentages $\{5\%, 20\%, 40\%\}$, and image zooms $\{10\%, 20\%, 40\%\}$ (denoted as s_img, m_img, l_img) was applied to a certain percentage δ of the test set.
- **Adversarial (*adv*)**: The adversarial samples generated by the fast gradient signed method (FGSM) [24] replace a certain percentage δ of the test set.
- **Knock-out (*ko*)**: Remove a certain percentage δ of class 0 samples from the test set [7].
- **Image + knock-out (*m_img+ko*)**: Using a combination of m_img shift with fixed $\delta_1 = 0.5$ and ko shift with δ.
- **Only-zero + image (*oz+m_img*)**: Using a combination that includes only images from class 0 of the test set (i.e., oz shift) and m_img shift with δ.

In summary, there are ten different types of shifts, each applied to three different percentages of the samples in the test set, making a total of 30 different shifts. However, it is worth noting that since this paper focuses on identifying harmful dataset shifts, not all 30 different shifts are considered, but only the harmful ones. Nowadays, various methods have been proposed to identify harmful shifts without the help of the real label of the test set [3], [4]. However, since the key of this paper is to verify the ability of each method to identify harmful shifts, therefore, in this paper, we draw on the labels of the test set and use a threshold of 0.8% to distinguish between harmful and harmless shifts, i.e., if the impairment of the model accuracy after applying a shift to the test set exceeds 0.8% (relative to the accuracy of the test set without the shift), we consider it as a harmful shift. We show the impact of the ResNet-50 model on accuracy after adding various shifts to each dataset, as shown in the table II below, and all the harmless shifts are underlined.

From Table II, we can see that the impact of different shifts on the model varies widely, but the same shift has a similar impact on the model on different datasets. And most of the shifts are detrimental to the performance of the model. In general, 20, 23 and 21 shifts of our ResNet-50 model on the MNIST, Fashion-MNIST and CIFAR-10 datasets, respectively, are considered as harmful shifts.

We use MMD with squared exponential kernel function [25] as a two-sample test method to calculate the distance from the validation set and the test set after applying the shift for different sample sizes $k \in \{10, 20, 50, 100, 200, 500, 1000\}$, and obtain the p-value of H_0 by performing a permutation test on the generated kernel matrix. In this paper, the significance level $\alpha = 0.05$, which means that H_0 can be rejected when

TABLE II

THE EFFECT OF ADDING VARIOUS SHIFTS TO THE RESNET-50 MODEL ON THE ACCURACY OF THE MODEL ON THE MNIST, FASHION-MNIST, AND CIFAR-10 DATASETS. THE VALUES IN THE TABLE INDICATE THE DEGREE OF ACCURACY DEGRADATION AFTER ADDING THE CORRESPONDING SHIFTS. ACCURACY DECREASES OF LESS THAN 0.8% ARE HIGHLIGHTED WITH AN UNDERLINE.

Shift Type		MNIST	Fashion-MNIST	CIFAR-10
s_gn	10%	0.02%	0.44%	-0.02%
	50%	0.10%	0.44%	-0.02%
	100%	0.03%	0.50%	0.09%
m_gn	10%	0.01%	0.48%	0.05%
	50%	0.20%	0.64%	0.85%
	100%	0.05%	0.77%	1.03%
l_gn	10%	3.00%	6.23%	6.19%
	50%	14.9%	29.1%	30.6%
	100%	29.5%	57.4%	60.5%
s_img	10%	0.30%	2.13%	1.97%
	50%	2.10%	8.63%	9.95%
	100%	3.80%	16.6%	19.5%
m_img	10%	7.00%	7.25%	4.74%
	50%	34.8%	34.2%	23.0%
	100%	69.3%	69.0%	46.6%
l_img	10%	8.70%	8.48%	6.20%
	50%	43.3%	42.0%	31.3%
	100%	86.4%	84.3%	63.9%
adv	10%	7.90%	8.26%	7.80%
	50%	39.8%	39.4%	38.5%
	100%	80.7%	78.4%	76.9%
ko	10%	0.06%	0.42%	0.05%
	50%	0.04%	0.33%	0.08%
	100%	0.04%	0.18%	0.27%
$m\_img + ko$	10%	35.1%	34.3%	23.3%
	50%	35.4%	34.5%	23.2%
	100%	34.7%	33.9%	23.9%
$oz + m\_img$	10%	7.10%	9.27%	1.39%
	50%	33.8%	37.0%	16.1%
	100%	70.0%	67.7%	37.4%

p-value ≤ 0.05. In addition, to ensure the reliability of the results, we performed 20 independent runs for each shift.

B. Experimental Result Analysis and Discussions

In this section, we first analyze and discuss the performance of each method to identify harmful shifts under different datasets and shifts. Then, we further discuss whether the performance of our methods is stable when different models are used (i.e., whether the detection effect is sensitive to the model).

1) Performance Comparison between Our Approach and the Existing Methods: We first evaluated the average success rate (p-value ≤ 0.05 is considered successful) and p-value of each method for all harmful shifts identified on each of the three datasets independently and present the average results on the three datasets, as shown in Tables III and IV below. And the trend plots of the average p-value of each method detecting all harmful shifts on each dataset with the change of detection samples are shown separately in Fig. 3 below.

Further, for each dataset, we show the trends of p-value with the number of detected samples for various dimensionality reduction methods at $\delta = 50\%$ under l_gn, l_img and $m\_img + ko$ shifts, as shown in Fig. 4 below. Due to the space limitation, we cannot show the p-value change trends for all the harmful shifts here.

TABLE III

THE AVERAGE DETECTION SUCCESS RATE OF DIFFERENT METHODS IN ALL HARMFUL SHIFTS ON THE THREE DATASETS. FOR EACH GIVEN SAMPLE SIZE, THE METHODS WITH THE HIGHEST AND SECOND HIGHEST SUCCESS RATES ARE HIGHLIGHTED IN BOLD AND UNDERLINED, RESPECTIVELY.

Method	Number of samples from validation and test sets						
	10	20	50	100	200	500	1000
NoRed	23%	28%	40%	45%	54%	67%	72%
PCA	23%	31%	47%	55%	60%	67%	73%
SRP	20%	27%	38%	47%	54%	66%	71%
UAE	26%	36%	51%	60%	67%	77%	85%
TAE	31%	40%	53%	60%	64%	71%	76%
BBSDs	19%	27%	41%	52%	60%	71%	76%
Grad	**41**%	54%	**68**%	74%	83%	**91**%	94%
GI	39%	**54**%	67%	**75**%	**83**%	91%	**94**%

TABLE IV

THE AVERAGE p-VALUE OF DIFFERENT METHODS IN ALL HARMFUL SHIFTS ON THE THREE DATASETS. FOR EACH GIVEN SAMPLE SIZE, THE METHODS WITH THE SMALLEST AND SECOND SMALLEST p-VALUE ARE HIGHLIGHTED IN BOLD AND UNDERLINED, RESPECTIVELY.

Method	Number of samples from validation and test sets						
	10	20	50	100	200	500	1000
NoRed	0.336	0.333	0.249	0.226	0.195	0.139	0.122
PCA	0.342	0.323	0.223	0.209	0.187	0.146	0.118
SRP	0.369	0.354	0.268	0.226	0.192	0.145	0.123
UAE	0.340	0.293	0.202	0.175	0.145	0.089	0.059
TAE	0.315	0.277	0.209	0.177	0.160	0.125	0.092
BBSDs	0.374	0.326	0.236	0.195	0.169	0.123	0.103
Grad	0.257	0.185	0.128	**0.096**	0.066	0.030	0.026
GI	**0.252**	**0.183**	**0.125**	0.100	**0.065**	**0.028**	**0.023**

From a comprehensive analysis of Tables III and IV and Fig. 3, it can be observed that, in general, the two linear dimensionality reduction methods, PCA and SRP, do not improve their ability to identify harmful shifts compared to the no-reduction (NoRed) method. The two nonlinear dimensionality reduction methods, TAE and UAE, can identify the harmful shifts better than the previous three methods. The BBSDs method, which utilizes the output information from the model, exhibits an inconsistent performance with better results on the CIFAR-10 dataset and poorer results on the MNIST and Fashion-MNIST datasets. And the Grad and GI methods are more accurate and effective than all the other methods in identifying the harmful shifts on each dataset under various sample sizes.

As we have performed more than twenty different shifts on each dataset, only a portion of the experimental results on different shifts are shown in Fig. 4 due to the page limit. Specifically on each shift, on the MNIST dataset, Grad and GI methods performed less well than TAE on l_gn shift; on the Fashion-MNIST dataset, Grad and GI methods were inferior to TAE on adv and l_gn shifts; on the CIFRA-10 dataset, Grad and GI methods did not perform as well as UAE on l_gn shift; in addition, Grad and GI methods have average performance

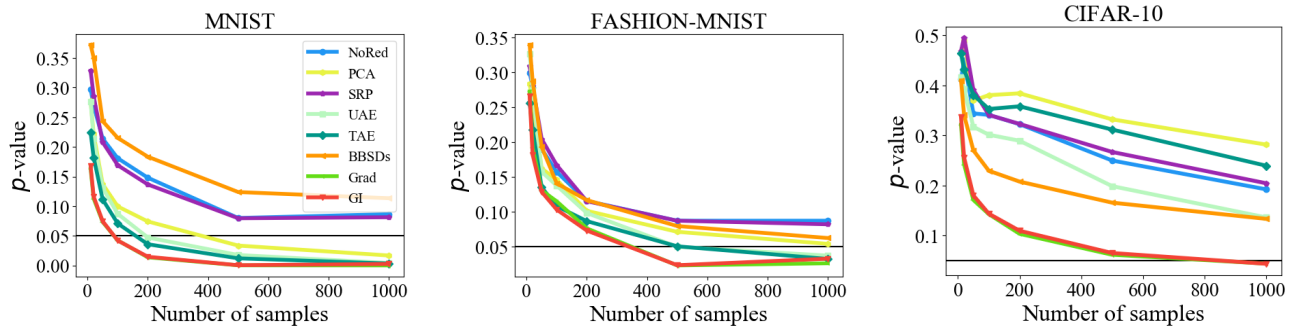

Fig. 3. Trend plots of the average p-values of all harmful shifts over the three datasets separately for different dimensionality reduction techniques with changes in the detection samples.

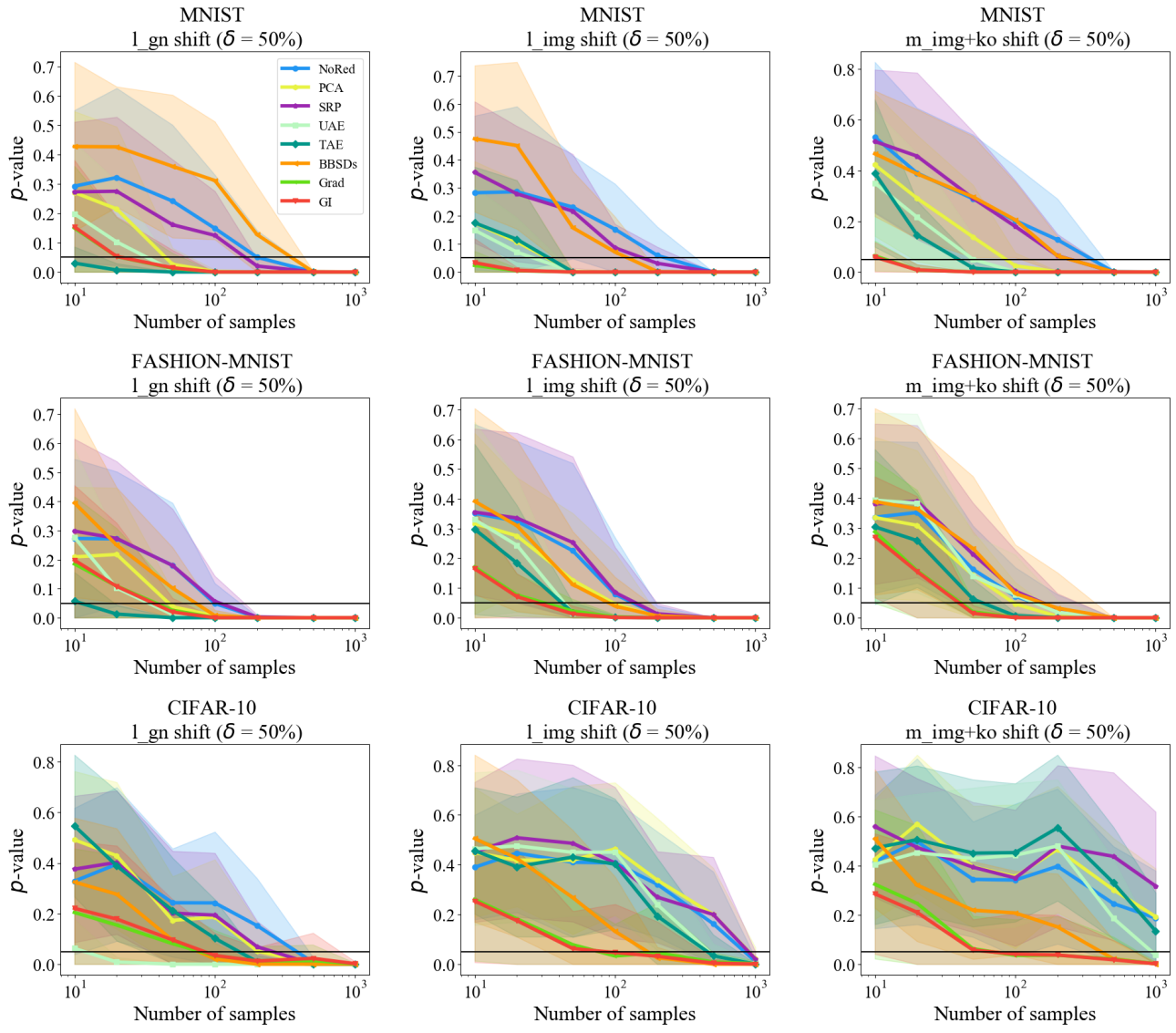

Fig. 4. p-value trends of various dimensionality reduction methods at $\delta = 50\%$ for l_gn shift, l_img shift and $m\_img + ko$ shift.

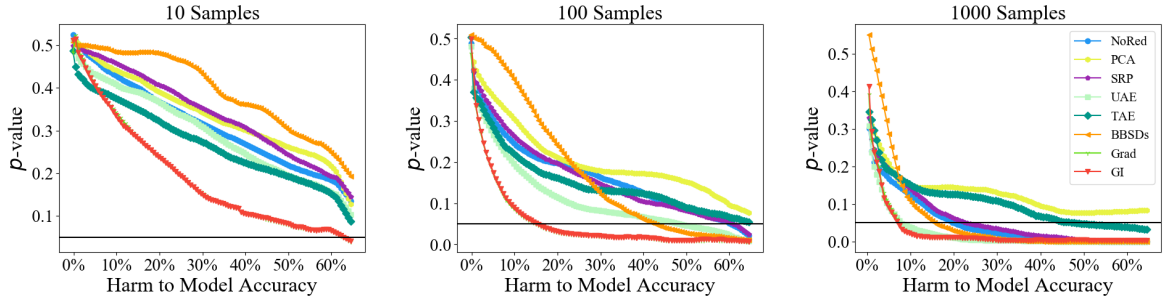

Fig. 5. Average results of the trend of p-value with increasing harmfulness of the nine shifts for different numbers of tested samples.

on $oz + m\_img$ shift. Otherwise, in all other cases, the GI and Grad methods outperform all other methods on all other shifts, i.e., in general, they can identify shifts faster and more accurately than other methods on the vast majority of harmful shifts.

In addition, our results reveal that Grad and GI have almost the consistent detection success rate and p-values. This consistency is not limited to the overall average results but is reflected in the results obtained on each individual shift. And the only difference between the two methods is that the Grad method utilizes only the gradient information from the model, while the GI method also includes the original input information. From this, we can also conclude that it is the information in the model that can help to identify the harmful shifts more effectively. And adding the original input information to the information in the model utilized does not improve the performance of detecting harmful shifts.

Furthermore, although for experimental convenience, in the above we use 0.8% to distinguish between harmful and harmless shifts. However, in fact, a blanket division of harmful and harmless shifts using 0.8% as the threshold is not reasonable enough. This is because the definitions and thresholds of harmful and harmless shifts will be different when the requirements of the users, the models and datasets used, and the shifts suffered are different. Therefore, to ensure the accuracy of the conclusions we obtained, we selected several shifts for further subdivision into about 30 different degrees/proportions of shifts so as to observe the trend of the p-value with the change in the degree of harm to the model accuracy from the shifts. Specifically, we constructed various degrees of gn and img shifts on $\delta = 100\%$ (i.e., different standard deviations σ for gn shifts and different degrees of image rotation, axis translation, and zoom for img shifts) and replaced samples in the test set with various proportions of adversarial samples (i.e., adv shifts with different values of δ) on each of the three datasets. Their impact on model accuracy ranges from 0% to between 66% and 84%, depending on the specific dataset and shift type (e.g., 30 different types of gn shift after subdivision on MNIST dataset has an impact on model accuracy from 0% to 80.3%), to observe the trend of the p-value of each method for detecting harmful shifts when the shift becomes more detrimental to the model accuracy.

Finally, we report the average results of the trend of the change in p-value with the increase of the harmful degree of the nine shifts (three shifts on each of the three datasets) on the model accuracy at 10, 100 and 1000 test samples, as shown in Fig. 5.

From Fig. 5 we can see that when the shifts have no effect on the model accuracy, the p-values of the methods are very similar, and both have difficulty identifying the occurrence of the shifts. However, as the harm of the shifts on the accuracy of the model increases, the p-value values of the Grad and GI methods decrease rapidly compared to the other methods. This indicates, on the one hand, that the conclusion in the literature [6] is not accurate enough: i.e., that the FAE method combined with the statistical test is not good at identifying all covariate shifts, but only on those dataset shifts that are detrimental to model performance (not only covariate shifts), and on the other hand further validates our previous conclusions that the method is more effective in identifying harmful shifts, which are not affected by the chosen threshold of 0.8%.

2) Sensitivity and Robustness Analysis of Our Approach against Different Model: Since FAE methods like Grad and GI need to use information from the models to assist in detecting harmful shifts, we wanted to further observe whether different models would affect the detection ability of our method. Therefore, we used Grad and GI methods on five different models, ResNet-18, ResNet-34, ResNet-50, DesNet-121, and InceptionV3 to detect harmful shifts and compared the average results of detection success rate and p-value on different models as shown in Fig. 6 below. It is worth mentioning that using 0.8% as the threshold to distinguish between harmful and harmless shifts has subtle differences in which shifts are harmful to the five models used, so to be convenient for comparison, we uniformly use the harmful shifts on ResNet-50 to present the detection results for each model. However, since the differences are very small, they do not affect the final results and conclusions.

From Fig. 6, we can see that the performance of the GI and Grad methods in identifying harmful shifts on different models is stable and is not affected by the difference of the used model. Therefore, we believe that although our detection method is model-based, the ability of our method to detect harmful shifts is not sensitive to the model used. Hence, our approach could be applied to harmful dataset shift detection for different ML models.

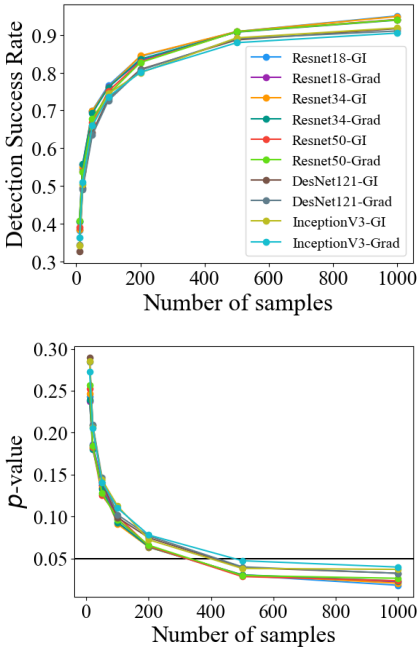

Fig. 6. Success rate and p-value of detecting harmful shifts under different models of Grad and GI methods.

V. Conclusion

In this paper, we use two FAE methods to extract the information involved in the trained model and combine it with MMD to detect harmful shifts. Experiments on more than twenty different covariate shifts and label shifts across three datasets demonstrate that our method can detect harmful shifts more effectively than the other state-of-the-art methods. And by comparing the results of the two different FAE methods, we further confirm that the effectiveness of the method is due to the utilization of the knowledge inside the model. In addition, although our method is model-based, they perform very stably and work well under different models, i.e., their ability to detect harmful shifts is not sensitive to different models.

In the future, we plan to (1) compare our method with other correlation methods [26] and (2) expose shifts have occurred in the data from an explanatory perspective.

References

[1] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," *Advances in Neural Information Processing Systems*, vol. 32, pp. 14 003–14 014, 2019.

[2] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[3] 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.

[4] T. Ginsberg, Z. Liang, and R. G. Krishnan, "A learning based hypothesis test for harmful covariate shift," *arXiv:2212.02742*, 2022.

[5] A. Komolafe, "Retraining model during deployment: Continuous training and continuous testing," 2022. [Online]. Available: https://neptune.ai/blog/retraining-model-during-deployment-continuous-training-continuous-testing

[6] S. Castle, R. Schwarzenberg, and M. Pourvali, "Detecting covariate drift with explanations," in *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 2021, pp. 317–322.

[7] Z. Lipton, Y.-X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3122–3130.

[8] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[9] C. Huang, Z. Zhang, B. Mao, and X. Yao, "An overview of artificial intelligence ethics," *IEEE Transactions on Artificial Intelligence*, pp. 1–21, 2022, doi: 10.1109/TAI.2022.3194503.

[10] Q. Zhang, J. Liu, Z. Zhang, J. Wen, B. Mao, and X. Yao, "Mitigating unfairness via evolutionary multi-objective ensemble learning," *IEEE Transactions on Evolutionary Computation*, 2022, doi: 10.1109/TEVC.2022.3209544.

[11] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 648–657.

[12] U. Bhatt, A. Weller, and J. M. Moura, "Evaluating and aggregating feature-based model explanations," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2021, pp. 3016–3022.

[13] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland, "Learning deep kernels for non-parametric two-sample tests," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6316–6326.

[14] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman, "On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[15] C. Molnar, *Interpretable machine learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[16] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv:1312.6034*, 2013.

[17] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv:1605.01713*, 2016.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," *arXiv:1708.07747*, 2017.

[20] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Technical Report*, 2009. [Online]. Available: https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[25] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[26] J. Lust and A. P. Condurache, "A survey on assessing the generalization envelope of deep neural networks: predictive uncertainty, out-of-distribution and adversarial samples," *arXiv preprint arXiv:2008.09381*, 2020.