

## A APPENDIX

### A.1 ADDITIONAL INFORMATION ON THE EXPERIMENTAL PROCEDURES.

**Datasets.** The number of samples in the clean sets (*i.e.*, the test sets) of the datasets we investigated are as follows:

- MNIST, FMNIST and CIFAR-10: 10'000,
- SVHN: 26'032.
- Imagenette: 3'925.

A sample from each dataset can be seen in Figure 4.

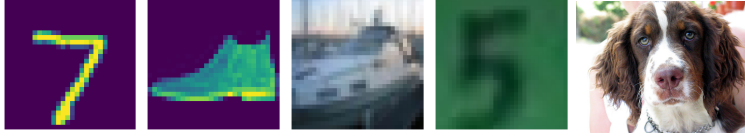


Figure 4: Sample images from all datasets used in the paper. From left to right: MNIST, FashionMNIST, CIFAR-10, SVHN and Imagenette.

**Shifts.** In order to illustrate the effect of the shift types described in Section 5.1 of the main article, we show the effects of the shifts and their intensities on the MNIST dataset in Figure 5. For the detailed parameters of each shift intensity (per dataset) we refer to the associated code.

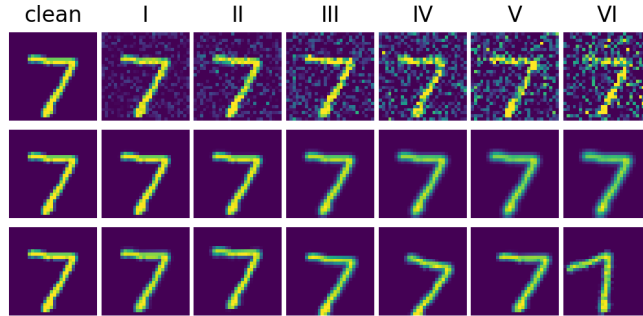


Figure 5: Illustration of intensities of the shift types — Gaussian noise (top row), Gaussian blur (middle row) and Image shift (bottom row) — on a sample from the MNIST dataset.

Shift	Int.	Feat.	MNIST						
			Sample size						
			10	20	50	100	200	500	1000
GN	II	MD	2.7 ± 0.8	7.7 ± 1.3	11.2 ± 1.6	29.3 ± 2.3	55.9 ± 2.5	94.7 ± 1.1	99.9 ± 0.1
		CV	0.0 ± 0.2	0.1 ± 0.1	0.0 ± 0.2	0.1 ± 0.1	0.3 ± 0.3	1.5 ± 0.6	6.1 ± 1.2
	IV	MD	8.6 ± 1.4	26.1 ± 2.2	60.9 ± 2.5	93.3 ± 1.3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
		CV	0.1 ± 0.2	1.0 ± 0.5	3.8 ± 1.0	17.6 ± 1.9	58.4 ± 2.5	99.3 ± 0.4	100.0 ± 0.0
	VI	MD	16.5 ± 1.9	54.2 ± 2.5	93.5 ± 1.3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
		CV	3.7 ± 1.0	20.4 ± 2.0	65.6 ± 2.4	98.9 ± 0.5	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
GB	II	MD	1.9 ± 0.7	2.7 ± 0.8	3.3 ± 0.9	4.3 ± 1.0	9.9 ± 1.5	22.2 ± 2.1	40.7 ± 2.5
		CV	0.0 ± 0.2	0.1 ± 0.1	0.0 ± 0.2	0.0 ± 0.2	0.0 ± 0.2	0.1 ± 0.1	0.1 ± 0.1
	IV	MD	4.8 ± 1.1	13.1 ± 1.7	30.3 ± 2.3	63.1 ± 2.4	93.9 ± 1.2	100.0 ± 0.0	100.0 ± 0.0
		CV	0.0 ± 0.2	0.0 ± 0.2	0.1 ± 0.2	0.3 ± 0.3	1.5 ± 0.6	11.3 ± 1.6	44.9 ± 2.5
	VI	MD	9.2 ± 1.5	25.1 ± 2.2	57.5 ± 2.5	92.4 ± 1.3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
		CV	0.1 ± 0.1	0.5 ± 0.4	1.4 ± 0.6	5.1 ± 1.1	22.1 ± 2.1	88.5 ± 1.6	100.0 ± 0.0
IS	II	MD	3.5 ± 0.9	8.6 ± 1.4	15.1 ± 1.8	32.7 ± 2.4	66.3 ± 2.4	98.0 ± 0.7	100.0 ± 0.0
		CV	0.0 ± 0.2	0.0 ± 0.2	0.0 ± 0.2	0.1 ± 0.2	0.7 ± 0.4	6.9 ± 1.3	28.4 ± 2.3
	IV	MD	5.6 ± 1.2	18.5 ± 2.0	42.1 ± 2.5	78.5 ± 2.1	98.0 ± 0.7	100.0 ± 0.0	100.0 ± 0.0
		CV	0.1 ± 0.2	0.9 ± 0.5	2.4 ± 0.8	15.5 ± 1.8	50.0 ± 2.5	99.5 ± 0.3	100.0 ± 0.0
	VI	MD	10.4 ± 1.5	34.0 ± 2.4	72.6 ± 2.3	98.9 ± 0.5	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
		CV	1.3 ± 0.6	7.3 ± 1.3	31.7 ± 2.4	83.3 ± 1.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0

Table 4: Power of the statistical test with MAGDiff (abbreviated as MD) and CV representations for the shift types Gaussian noise (GN), Gaussian blur (GB) and Image shift (IS), three different shift intensities (II, IV, VI) and fixed  $\delta = 0.5$  for the MNIST dataset. The estimated 95%-confidence intervals are indicated.

## A.2 ADDITIONAL EXPERIMENTAL RESULTS

**Sample size.** To further support our claims, we include comprehensive results of the power with respect to the sample size for the MNIST, Imagenette and CIFAR-10 datasets in Tables 4, 5 and 6. We provide all results for the shift intensities II, IV and VI, for all shift types, and fixed  $\delta = 0.5$  for MNIST, CIFAR-10, respectively  $\delta = 1.0$  for Imagenette (the  $\delta$  were chosen so that the task is comparatively easy at high shift intensity and hard at low shift intensity for both methods).

**Shift intensity.** In Figures 6, 7 and 8, we collect the plots of the estimated powers of the test for multiple cases, in addition to the one presented in the main article. Note that the only situation in which MAGDiff is very slightly outperformed by the baseline CV, is the case of FMNIST, when we consider MAGDiff representations of layer  $l_{-1}$ . In all other cases, shift detection using MAGDiff representations clearly outperforms the baseline of CV by a large margin.

**Model accuracy.** In Figure 9 we show the impact of the shift type and intensity on the model accuracy. It is interesting to note that, even in cases where the model accuracy is only minimally impacted (*e.g.*, for Gaussian blur on the MNIST and FMNIST datasets), our method can still reliably detect the presence of the shift.

**Norm variations.** As mentioned in the main paper, many variations of MAGDiff are conceivable. Here, we present some experimental results for variations on the type of norm that is used to construct the MAGDiff representations. In Figure 10 we show the results where, instead of the Frobenius-norm, we consider the *spectral norm* as well as the operator norm  $\|\cdot\|_\infty$  induced by the sup-norm on vectors. The spectral norm is equal to the largest singular value and  $\|\cdot\|_\infty$  is defined by:

$$\|M\|_\infty := \sup_{x \neq 0} \frac{\|Mx\|_\infty}{\|x\|_\infty} = \max_{1 \leq i \leq m} \sum_{j=1}^n |m_{ij}|$$

for  $M \in \mathbb{R}^{m \times n}$ . Comparing to Figure 6, we observe that the results for the Frobenius-norm and the spectral norm are almost identical. However, while the results for the  $\|\cdot\|_\infty$  are still better (in almost all cases) than those of the baseline CV, they are less powerful than those of the Frobenius norm.

Shift	Int.	Feat.	Imagenette						
			Sample size						
			10	20	50	100	200	500	1000
GN	II	MD	0.7 $\pm$ 0.4	2.2 $\pm$ 0.7	6.3 $\pm$ 1.2	16.7 $\pm$ 1.9	46.2 $\pm$ 2.5	93.3 $\pm$ 1.3	99.9 $\pm$ 0.1
		CV	2.4 $\pm$ 0.8	4.5 $\pm$ 1.1	2.7 $\pm$ 0.8	4.1 $\pm$ 1.0	4.2 $\pm$ 1.0	7.3 $\pm$ 1.3	10.3 $\pm$ 1.5
	IV	MD	0.9 $\pm$ 0.5	3.6 $\pm$ 0.9	7.3 $\pm$ 1.3	22.1 $\pm$ 2.1	60.7 $\pm$ 2.5	98.5 $\pm$ 0.6	100.0 $-$ 0.0
		CV	1.6 $\pm$ 0.6	4.1 $\pm$ 1.0	3.2 $\pm$ 0.9	3.9 $\pm$ 1.0	4.7 $\pm$ 1.1	7.5 $\pm$ 1.3	10.1 $\pm$ 1.5
	VI	MD	0.9 $\pm$ 0.5	3.6 $\pm$ 0.9	7.3 $\pm$ 1.3	22.1 $\pm$ 2.1	60.7 $\pm$ 2.5	98.5 $\pm$ 0.6	100.0 $-$ 0.0
		CV	2.2 $\pm$ 0.7	4.9 $\pm$ 1.1	3.5 $\pm$ 0.9	5.3 $\pm$ 1.1	6.8 $\pm$ 1.3	15.5 $\pm$ 1.8	33.5 $\pm$ 2.4
GB	II	MD	0.5 $\pm$ 0.3	2.5 $\pm$ 0.8	4.1 $\pm$ 1.0	15.3 $\pm$ 1.8	40.6 $\pm$ 2.5	91.7 $\pm$ 1.4	99.9 $\pm$ 0.2
		CV	2.1 $\pm$ 0.7	3.7 $\pm$ 1.0	3.2 $\pm$ 0.9	3.7 $\pm$ 1.0	7.0 $\pm$ 1.3	11.3 $\pm$ 1.6	18.2 $\pm$ 2.0
	IV	MD	1.0 $\pm$ 0.5	3.5 $\pm$ 0.9	9.1 $\pm$ 1.5	29.3 $\pm$ 2.3	67.9 $\pm$ 2.4	99.1 $\pm$ 0.5	100.0 $-$ 0.0
		CV	2.3 $\pm$ 0.8	4.5 $\pm$ 1.0	3.7 $\pm$ 1.0	4.5 $\pm$ 1.1	6.3 $\pm$ 1.2	13.1 $\pm$ 1.7	26.9 $\pm$ 2.2
	VI	MD	1.3 $\pm$ 0.6	5.0 $\pm$ 1.1	17.2 $\pm$ 1.9	50.5 $\pm$ 2.5	89.9 $\pm$ 1.5	100.0 $-$ 0.0	100.0 $-$ 0.0
		CV	2.1 $\pm$ 0.7	3.5 $\pm$ 0.9	3.1 $\pm$ 0.9	5.0 $\pm$ 1.1	6.9 $\pm$ 1.3	18.5 $\pm$ 2.0	46.8 $\pm$ 2.5
IS	II	MD	0.5 $\pm$ 0.4	1.1 $\pm$ 0.5	1.9 $\pm$ 0.7	4.7 $\pm$ 1.1	13.6 $\pm$ 1.7	44.5 $\pm$ 2.5	83.1 $\pm$ 1.9
		CV	2.5 $\pm$ 0.8	3.7 $\pm$ 1.0	3.9 $\pm$ 1.0	3.1 $\pm$ 0.9	4.3 $\pm$ 1.0	6.3 $\pm$ 1.2	9.9 $\pm$ 1.5
	IV	MD	0.3 $\pm$ 0.3	1.9 $\pm$ 0.7	2.9 $\pm$ 0.9	9.1 $\pm$ 1.5	28.1 $\pm$ 2.3	75.1 $\pm$ 2.2	98.3 $\pm$ 0.7
		CV	1.7 $\pm$ 0.7	3.0 $\pm$ 0.9	3.6 $\pm$ 0.9	3.9 $\pm$ 1.0	4.8 $\pm$ 1.1	8.3 $\pm$ 1.4	12.8 $\pm$ 1.7
	VI	MD	0.6 $\pm$ 0.4	2.5 $\pm$ 0.8	5.0 $\pm$ 1.1	14.9 $\pm$ 1.8	44.3 $\pm$ 2.5	93.5 $\pm$ 1.2	99.9 $\pm$ 0.1
		CV	2.0 $\pm$ 0.7	3.9 $\pm$ 1.0	1.9 $\pm$ 0.7	4.6 $\pm$ 1.1	6.1 $\pm$ 1.2	8.3 $\pm$ 1.4	15.5 $\pm$ 1.8

Table 5: Power of the statistical test with MAGDiff (abbreviated as MD) and CV representations for the shift types Gaussian noise (GN), Gaussian blur (GB) and Image shift (IS), three different shift intensities (II, IV, VI) and fixed  $\delta = 1$  for the Imagenette dataset. The estimated 95%-confidence intervals are indicated.

Shift	Int.	Feat.	CIFAR-10						
			Sample size						
			10	20	50	100	200	500	1000
GN	II	MD	1.8 $\pm$ 0.7	5.3 $\pm$ 1.1	16.7 $\pm$ 1.9	47.0 $\pm$ 2.5	86.9 $\pm$ 1.7	100.0 $-$ 0.0	100.0 $-$ 0.0
		CV	2.3 $\pm$ 0.8	4.5 $\pm$ 1.1	6.9 $\pm$ 1.3	19.1 $\pm$ 2.0	38.3 $\pm$ 2.5	88.3 $\pm$ 1.6	99.9 $\pm$ 0.1
	IV	MD	2.5 $\pm$ 0.8	11.1 $\pm$ 1.6	36.7 $\pm$ 2.4	81.1 $\pm$ 2.0	99.3 $\pm$ 0.4	100.0 $-$ 0.0	100.0 $-$ 0.0
		CV	2.5 $\pm$ 0.8	6.7 $\pm$ 1.3	11.7 $\pm$ 1.6	29.9 $\pm$ 2.3	63.4 $\pm$ 2.4	99.3 $\pm$ 0.4	100.0 $-$ 0.0
	VI	MD	2.7 $\pm$ 0.8	14.7 $\pm$ 1.8	49.2 $\pm$ 2.5	91.2 $\pm$ 1.4	99.9 $\pm$ 0.1	100.0 $-$ 0.0	100.0 $-$ 0.0
		CV	2.7 $\pm$ 0.8	7.3 $\pm$ 1.3	14.2 $\pm$ 1.8	37.5 $\pm$ 2.5	77.3 $\pm$ 2.1	99.9 $\pm$ 0.2	100.0 $-$ 0.0
GB	II	MD	0.8 $\pm$ 0.5	2.8 $\pm$ 0.8	6.6 $\pm$ 1.3	18.9 $\pm$ 2.0	49.5 $\pm$ 2.5	93.2 $\pm$ 1.3	100.0 $-$ 0.0
		CV	2.6 $\pm$ 0.8	4.0 $\pm$ 1.0	3.4 $\pm$ 0.9	6.8 $\pm$ 1.3	11.1 $\pm$ 1.6	30.4 $\pm$ 2.3	58.9 $\pm$ 2.5
	IV	MD	1.8 $\pm$ 0.7	5.7 $\pm$ 1.2	19.5 $\pm$ 2.0	49.7 $\pm$ 2.5	89.6 $\pm$ 1.5	99.9 $\pm$ 0.1	100.0 $-$ 0.0
		CV	2.5 $\pm$ 0.8	6.4 $\pm$ 1.2	7.3 $\pm$ 1.3	13.9 $\pm$ 1.7	35.9 $\pm$ 2.4	84.0 $\pm$ 1.9	99.8 $\pm$ 0.2
	VI	MD	2.1 $\pm$ 0.7	6.3 $\pm$ 1.2	23.4 $\pm$ 2.1	62.5 $\pm$ 2.4	96.1 $\pm$ 1.0	100.0 $-$ 0.0	100.0 $-$ 0.0
		CV	3.0 $\pm$ 0.9	8.9 $\pm$ 1.4	14.7 $\pm$ 1.8	44.2 $\pm$ 2.5	85.5 $\pm$ 1.8	100.0 $-$ 0.0	100.0 $-$ 0.0
IS	II	MD	0.3 $\pm$ 0.3	1.3 $\pm$ 0.6	3.6 $\pm$ 0.9	6.7 $\pm$ 1.3	19.6 $\pm$ 2.0	60.1 $\pm$ 2.5	92.6 $\pm$ 1.3
		CV	2.5 $\pm$ 0.8	3.3 $\pm$ 0.9	2.5 $\pm$ 0.8	4.4 $\pm$ 1.0	7.9 $\pm$ 1.4	16.5 $\pm$ 1.9	31.5 $\pm$ 2.4
	IV	MD	0.6 $\pm$ 0.4	2.4 $\pm$ 0.8	3.9 $\pm$ 1.0	16.1 $\pm$ 1.9	39.9 $\pm$ 2.5	88.2 $\pm$ 1.6	99.9 $\pm$ 0.1
		CV	2.0 $\pm$ 0.7	3.9 $\pm$ 1.0	2.9 $\pm$ 0.9	6.7 $\pm$ 1.3	10.8 $\pm$ 1.6	25.4 $\pm$ 2.2	53.1 $\pm$ 2.5
	VI	MD	1.3 $\pm$ 0.6	3.9 $\pm$ 1.0	8.9 $\pm$ 1.4	22.8 $\pm$ 2.1	57.4 $\pm$ 2.5	97.7 $\pm$ 0.8	100.0 $-$ 0.0
		CV	2.0 $\pm$ 0.7	4.6 $\pm$ 1.1	4.4 $\pm$ 1.0	9.5 $\pm$ 1.5	15.5 $\pm$ 1.8	44.5 $\pm$ 2.5	83.2 $\pm$ 1.9

Table 6: Power of the statistical test with MAGDiff (abbreviated as MD) and CV representations for the shift types Gaussian noise (GN), Gaussian blur (GB) and Image shift (IS), three different shift intensities (II, IV, VI) and fixed  $\delta = 0.5$  for the CIFAR-10 dataset. The estimated 95%-confidence intervals are indicated.

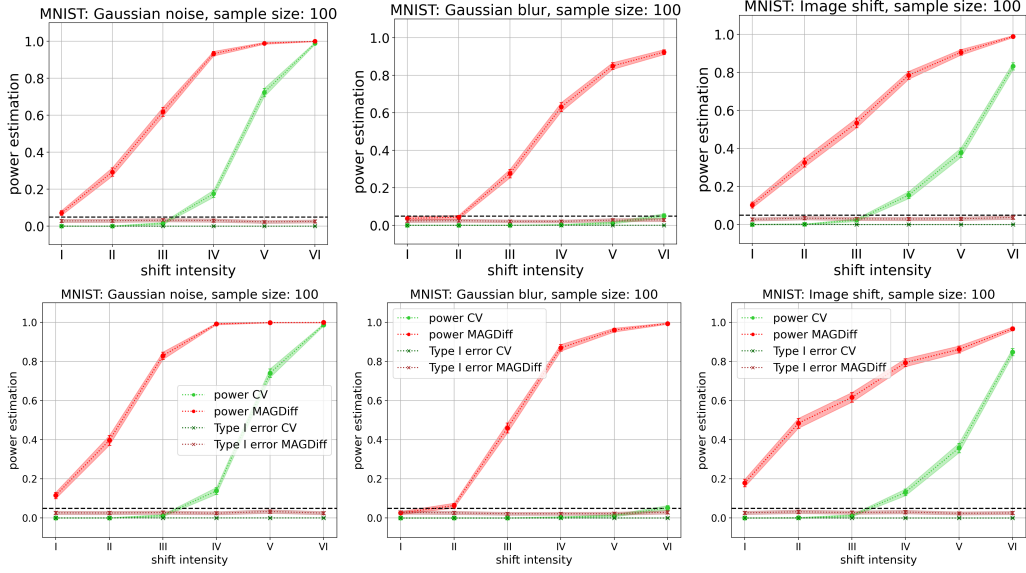


Figure 6: Power and type I error of the test with `MAGDiff` (red) and `CV` (green) representations w.r.t. the shift intensity for various shift types on the MNIST dataset with  $\delta = 0.5$ , sample size 100, for layers  $\ell_{-1}$  (top row) and  $\ell_{-3}$  (bottom row).

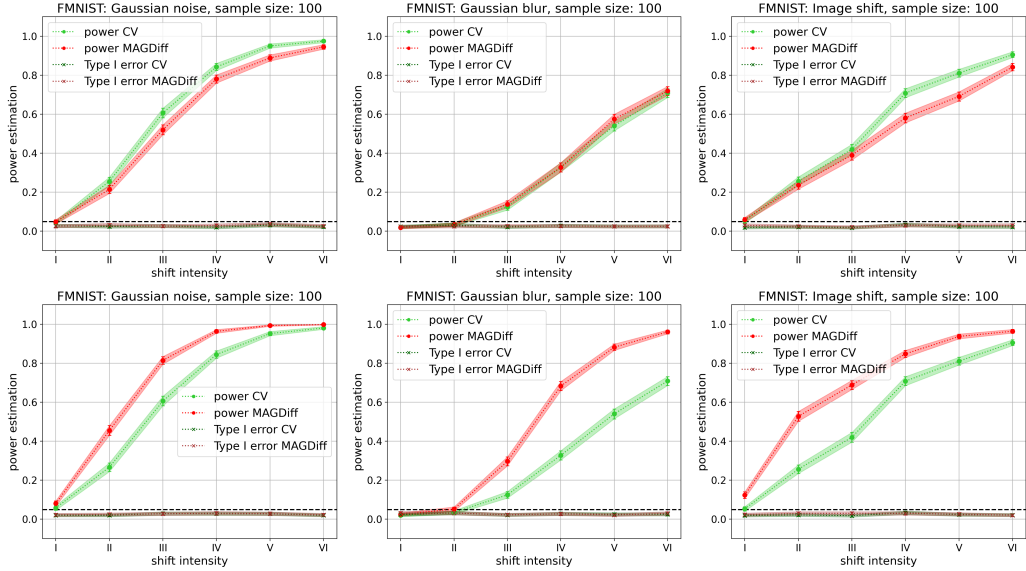


Figure 7: Power and type I error of the test with `MAGDiff` (red) and `CV` (green) representations w.r.t. the shift intensity for various shift types on the FMNIST dataset with  $\delta = 0.5$ , sample size 100, for layers  $\ell_{-1}$  (top row) and  $\ell_{-3}$  (bottom row).

### A.3 THEORETICAL OBSERVATIONS REGARDING THE PRESERVATION OF SHIFT DISTRIBUTIONS BY CONTINUOUS FUNCTIONS

In the main article, we mentioned the fact that under generic conditions, two distinct distributions remain distinct under the application of a non-constant continuous function (though this does not necessarily translate to good quantitative guarantees). In this section, we make this assertion more formal and provide an elementary proof.

Let  $X$  be a separable metric space, and denote by  $\mathcal{P}(X)$  the set of probability measures on  $X$  equipped with its Borel  $\sigma$ -algebra. Let  $C_b(X)$  be the real bounded continuous functions on  $X$ . We

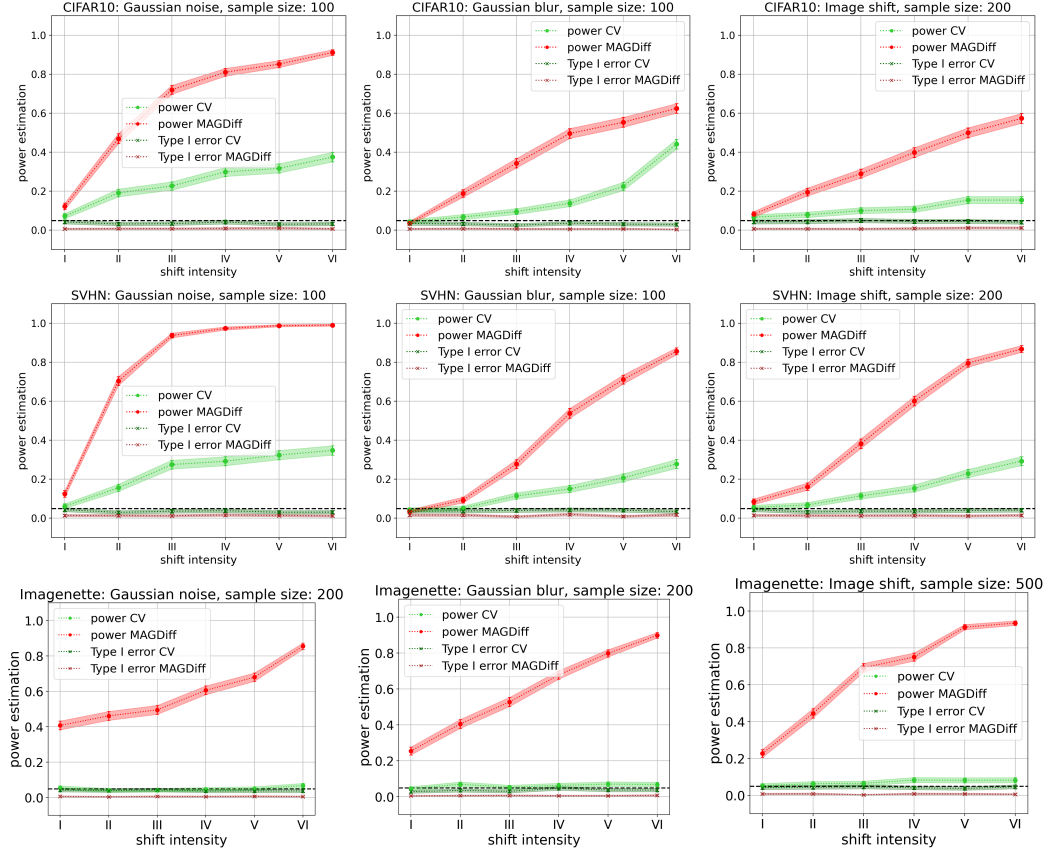


Figure 8: Power and type I error of the test with `MAGDiff` (red) and `CV` (green) representations w.r.t. the shift intensity for various shift types on the CIFAR-10, SVHN (with  $\delta = 0.5$ ) and Imagenette (with  $\delta = 1$ ) datasets. Sample sizes and values of  $\delta$  were chosen to make the plots as expressive as possible (low power for low shift intensity, high power for high shift intensity), as the difficulty of the task varies depending on the shift type and dataset.

consider the weak convergence topology on  $\mathcal{P}(X)$ ; remember that a subbase for this topology is given by the sets

$$U_{f,a,b} := \left\{ \mu \in \mathcal{P}(X) \mid \int_X f d\mu \in ]a, b[ \right\},$$

for  $f \in C_b(X)$  and  $a < b \in \mathbb{R}$  (see for example Kallianpur (1961)).

Now let  $X, Y$  be two such separable metric spaces with their Borel  $\sigma$ -algebra. Any measurable map  $F : X \rightarrow Y$  induces a map

$$\begin{aligned} F_* : \mathcal{P}(X) &\rightarrow \mathcal{P}(Y) \\ \mu &\mapsto F_*(\mu), \end{aligned}$$

where  $F_*(\mu)$  is the pushforward of  $\mu$  by  $F$ , that is the measure on  $\mathcal{P}(Y)$  characterized by  $F_*(\mu)(A) = \mu(F^{-1}(A))$  for any Borel set  $A \subset Y$ .

**Fact 1.** *If  $F : X \rightarrow Y$  is continuous, then  $F_* : \mathcal{P}(X) \rightarrow \mathcal{P}(Y)$  is continuous for the weak convergence topology.*

*Proof.* Given  $f \in C_b(X)$  and  $a < b \in \mathbb{R}$ , we see that  $F_*^{-1}(U_{f,a,b}) = U_{f \circ F, a, b}$ , which is enough to conclude by the definition of subbases.  $\square$

The following result follows from standard arguments; we give an elementary proof for the convenience of the reader.

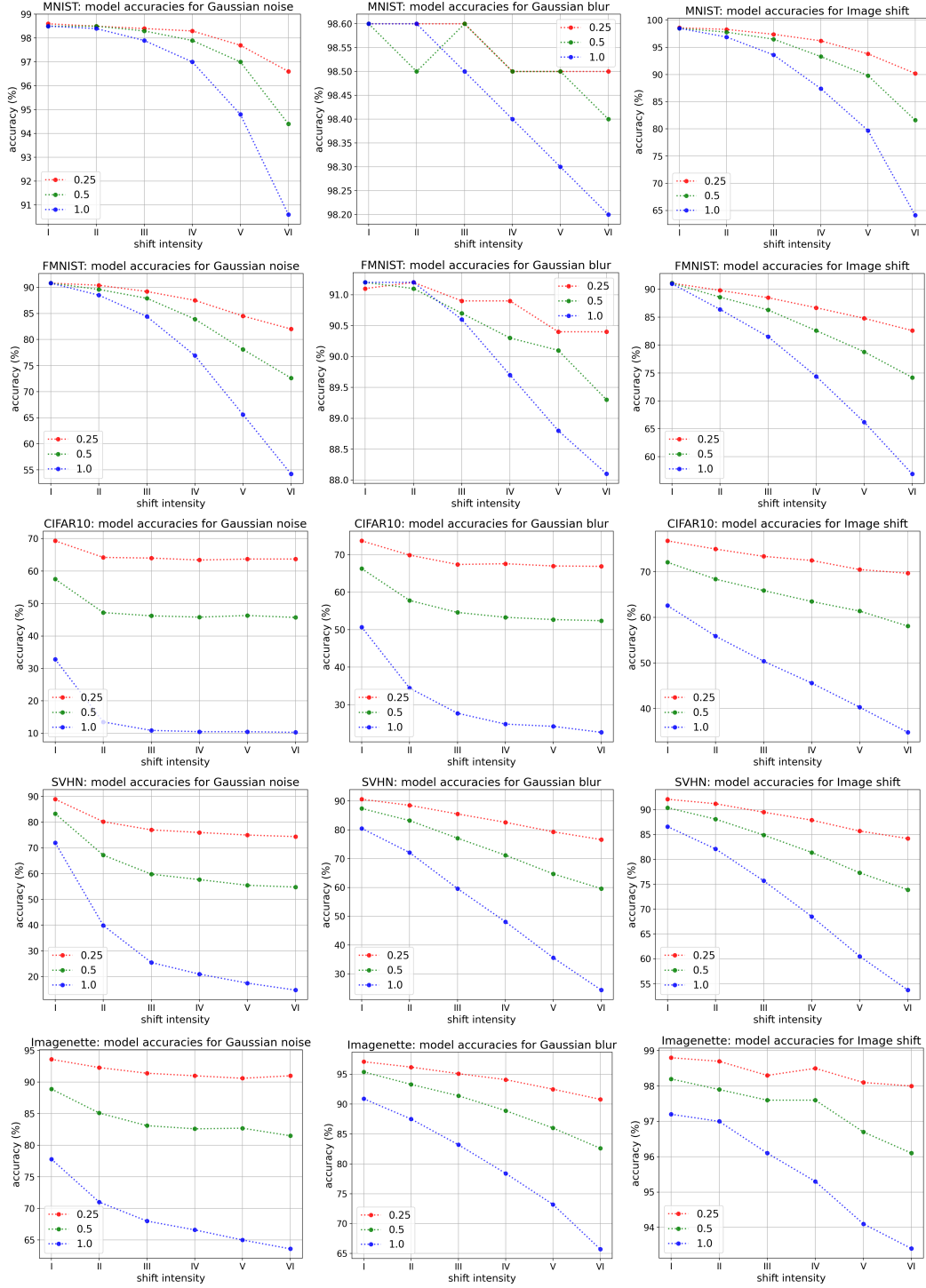


Figure 9: The impact of the shift type and intensity on the model accuracy for  $\delta = 1.0$  (blue),  $\delta = 0.5$  (green) and  $\delta = 0.25$  (red).

**Proposition 1.** Let  $F : X \rightarrow \mathbb{R}$  be continuous and non-constant for  $X$  a separable metric space, and let  $\nu \in F_*(\mathcal{P}(X)) \subset \mathcal{P}(\mathbb{R})$ . Then the complement  $F_*^{-1}(\{\nu\})^c = \mathcal{P}(X) \setminus F_*^{-1}(\{\nu\})$  of the set  $F_*^{-1}(\{\nu\})$  is a dense open set of  $\mathcal{P}(X)$  for the weak topology.

*Proof.* As  $\mathbb{R}$  is separable and metric, it is easy to show that the singleton  $\{\nu\} \in \mathcal{P}(\mathbb{R})$  is closed (see for example (Kallianpur, 1961, Thm 4.1)). As we know from Fact 1 that  $F_*$  is continuous, we conclude that  $F_*^{-1}(\{\nu\})$  is closed and  $F_*^{-1}(\{\nu\})^c$  is open.

It remains to show that it is dense in  $\mathcal{P}(X)$ . Let  $\mu$  belong to  $F_*^{-1}(\{\nu\})$ , and let  $V \subset \mathcal{P}(X)$  be an open set containing  $\mu$ . We have to show that  $F_*^{-1}(\{\nu\})^c \cap V$  is non-empty. Thanks to the definition of the weak topology, we can assume (by potentially taking a subset of  $V$ ) that

$$V = \bigcap_{i=1}^n \left\{ \tilde{\mu} \in \mathcal{P}(X) \text{ s.t. } \int_X f_i d\tilde{\mu} \in ]a_i, b_i[ \right\}$$

for some  $f_1, \dots, f_n \in C_b(X)$  and  $a_1, b_1, \dots, a_n, b_n \in \mathbb{R}$  with  $a_i < b_i$  for all  $i$ . Let  $x_1$  be any point in the support of  $\mu$ . Then  $\mu(B(x_1, \delta)) > 0$  for all  $\delta > 0$  by definition of the support. As  $F$  is non-constant, there exists  $x_2 \in X$  such that  $F(x_2)$  is not equal to  $F(x_1)$ . Let us assume that  $F(x_1) > F(x_2)$  (the proof is similar if  $F(x_2) > F(x_1)$ ). By continuity, there exists  $\epsilon > 0$  such that  $F(x) > F(x_2)$  for any  $x \in B(x_1, \epsilon)$ . Define  $m := \mu(B(x_1, \epsilon)) > 0$ . For  $t \in ]0, 1[$ , we define a new measure  $\mu_t$  as follows : for any measurable set  $A$ , we let

$$\mu_t(A) = \mu(A \setminus B(x_1, \epsilon)) + (1-t)\mu(B(x_1, \epsilon) \cap A) + tm1_{x_2 \in A}.$$

For any such  $t \in ]0, 1[$ , observe that

$$\begin{aligned} F_*(\mu_t)(]F(x_2), +\infty[) &= \mu_t(F_*^{-1}(]F(x_2), +\infty[)) \\ &= F_*(\mu)(]F(x_2), +\infty[) - t\mu(B(x_1, \epsilon)) \\ &< F_*(\mu)(]F(x_2), +\infty[), \end{aligned}$$

which shows that  $F_*(\mu) \neq F_*(\mu_t)$ , hence that  $\mu_t \in F_*^{-1}(\{\nu\})^c$ .

On the other hand, we see that  $|\int_X f_i d\mu_t - \int_X f_i d\mu| < 2tm\|f_i\|_\infty$  for  $i = 1, \dots, n$ . Since  $\mu \in V = \bigcap_{i=1}^n \{\tilde{\mu} \in \mathcal{P}(X) \text{ s.t. } \int_X f_i d\tilde{\mu} \in ]a_i, b_i[ \}$ , thus  $\mu_t \in V$  for  $t \in ]0, 1[$  small enough. This shows that  $V \cap F_*^{-1}(\{\nu\})^c$  is non-empty, and thus we conclude that  $F_*^{-1}(\{\nu\})^c$  is dense in  $\mathcal{P}(X)$ .  $\square$

As a direct corollary, we get the following statement, where *generic*, as above, means that the property is true for any random variable  $x'$  whose distribution belongs to a fixed dense open set of the space of distributions on  $\mathbb{R}^n$  :

**Corollary 1.** *Let  $F : \mathbb{R}^n \rightarrow \mathbb{R}^k$  be a non-constant continuous function represented by a neural network, and let  $x$  be a random variable on  $\mathbb{R}^n$ . For a generic random variable  $x'$  on  $\mathbb{R}^n$ , the distribution of  $F(x')$  will be different from that of  $F(x)$ .*

*Proof.*  $\mathbb{R}^n$  is a separable metric space, and if  $F$  is non-constant, so is at least one of its coordinate functions  $F_i : \mathbb{R}^n \rightarrow \mathbb{R}$ , to which Proposition 1 then applies. If the distribution of  $F_i(x')$  is different from that of  $F_i(x)$ , then the distribution of  $F(x')$  is different from that of  $F(x)$ .  $\square$

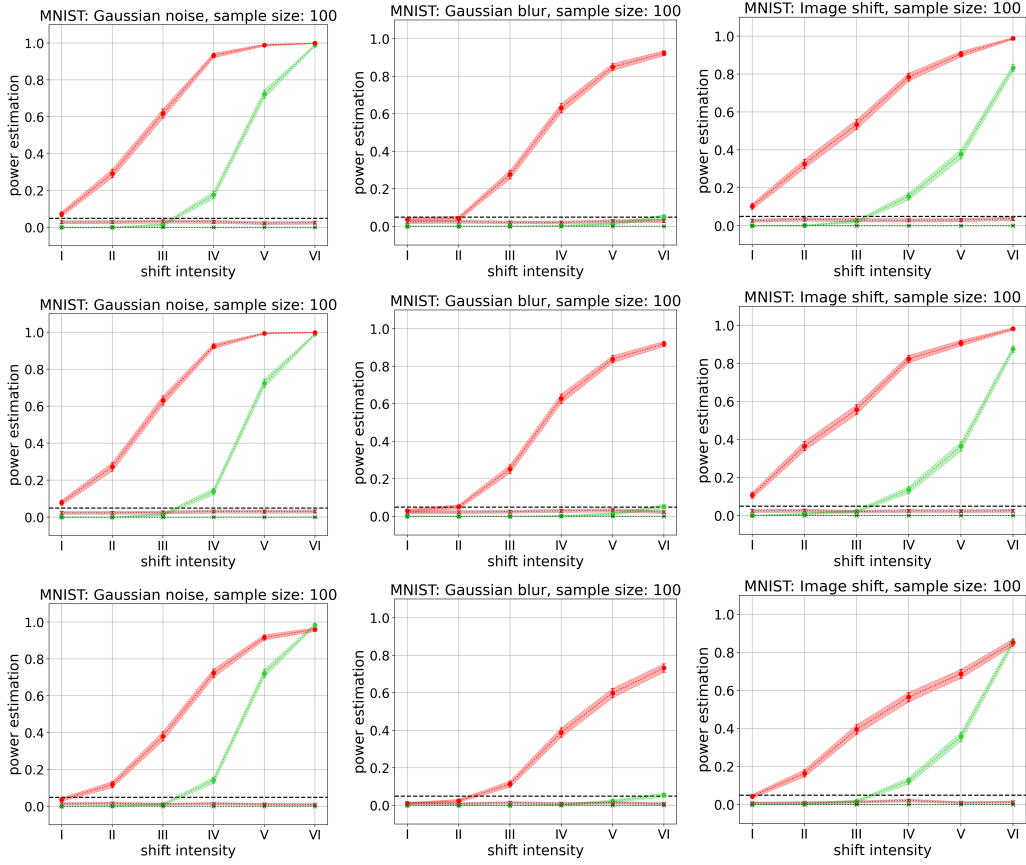


Figure 10: Power and type I error of the test with MAGDiff (red) w.r.t. the Frobenius norm, used in all other experiments, (top row), the spectral-norm (middle row) and  $\|\cdot\|_\infty$  (bottom row) and CV (green) representations w.r.t. the shift intensity for various shift types on the MNIST dataset with  $\delta = 0.5$ , sample size 100, for layer  $\ell_{-1}$ .