Out-of-Distribution Aware Classification for Tabular Data Supplementary Results

In the main paper, we presented results averaged across all datasets. In this document, we provide detailed results for each individual dataset, including the standard errors. The reported results are as follows:

- **OOD-Aware Classification, Test Settings I**: Detailed results for test settings I are presented for each dataset in the following tables: Adult in Table 1, Compas in Table 2, Cover in Table 3, Dilbert in Table 4, GMSC in Table 5, Heloc in Table 6, and Jannis in Table 7.
- **OOD-Aware Classification, Test Settings II**: Detailed results for test settings II are presented for each dataset in the following tables: Adult in Table 8, Compas in Table 9, Cover in Table 10, Dilbert in Table 11, GMSC in Table 12, Heloc in Table 13, and Jannis in Table 14.
- **Counterfactual Experiment**: Detailed results for the counterfactual experiment, including standard errors, are presented in Table 15.

Table 1. Detailed results for the Adult dataset in Test settings I.

Method	OOD class: 0		OOD class: 1	
	In	OOD	In	OOD
Pipeline	99.2 ± 0.0	99.2 ± 0.0	99.0 ± 0.0	99.0 ± 0.0
OCT	96.5 ± 0.4	96.4 ± 0.5	95.7 ± 1.0	95.4 ± 1.2
MCDD	90.8 ± 0.0	89.4 ± 0.0	72.9 ± 0.0	46.1 ± 0.0
O-GBDT	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0
DK	66.9 ± 0.0	2.5 ± 0.2	69.3 ± 0.1	20.8 ± 0.3
Multi	86.1 ± 0.0	89.1 ± 0.0	84.7 ± 0.0	88.3 ± 0.0
Incremental	66.8 ± 0.1	1.4 ± 0.6	68.8 ± 0.2	17.3 ± 1.6
Energy+	91.0 ± 0.7	89.4 ± 1.0	93.2 ± 0.4	92.4 ± 0.6
Self	95.5 ± 0.5	95.7 ± 0.5	96.6 ± 0.2	96.7 ± 0.3
Exposure	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0
POEM	70.9 ± 0.0	33.6 ± 0.0	91.0 ± 0.0	89.3 ± 0.0
WOODS	69.6 ± 2.3	21.5 ± 13.2	73.4 ± 3.1	38.0 ± 15.5
Logitnorm	68.3 ± 1.1	17.1 ± 6.9	66.3 ± 0.2	3.9 ± 2.0
VOS	68.9 ± 0.0	22.5 ± 0.0	66.3 ± 0.0	6.1 ± 0.0
ReAct	66.1 ± 0.2	3.8 ± 1.8	66.4 ± 0.3	4.8 ± 2.5
Energy	66.4 ± 0.3	4.4 ± 2.3	66.1 ± 0.2	2.3 ± 1.3
Confidence	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0
Original	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0

Table 2. Detailed results for the Compas dataset in Test settings I.

Method	OOD class: 0		OOD class: 1	
	In	OOD	In	OOD
Pipeline	100.0 ± 0.0	100.0 ± 0.0	98.4 ± 0.0	98.4 ± 0.0
OCT	96.4 ± 0.2	96.4 ± 0.2	96.2 ± 0.5	96.3 ± 0.5
MCDD	66.0 ± 0.0	2.0 ± 0.0	86.8 ± 0.0	84.4 ± 0.0
O-GBDT	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0
DK	67.5 ± 0.1	7.1 ± 0.8	68.1 ± 0.6	13.1 ± 4.4
Multi	76.7 ± 0.0	84.1 ± 0.0	76.3 ± 0.0	83.9 ± 0.0
Incremental	67.4 ± 0.1	6.7 ± 1.0	67.0 ± 0.1	4.3 ± 0.6
Energy+	88.2 ± 0.6	85.4 ± 0.9	79.5 ± 0.2	68.1 ± 0.7
Self	96.4 ± 0.3	96.7 ± 0.3	93.7 ± 0.3	94.1 ± 0.3
Exposure	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0
POEM	91.0 ± 0.0	89.5 ± 0.0	77.9 ± 0.0	63.3 ± 0.0
WOODS	68.2 ± 2.5	13.4 ± 12.4	69.2 ± 2.3	20.4 ± 12.0
Logitnorm	67.2 ± 0.9	9.9 ± 6.7	67.6 ± 1.4	12.5 ± 9.0
VOS	66.9 ± 0.0	5.9 ± 0.0	66.0 ± 0.0	0.0 ± 0.0
ReAct	67.1 ± 0.8	10.0 ± 5.2	69.8 ± 2.3	24.7 ± 11.9
Energy	66.7 ± 0.6	8.9 ± 3.0	69.7 ± 2.3	23.0 ± 12.4
Confidence	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0
Original	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0

Table 3. Detailed results for the Cover dataset in Test settings I. For this dataset, all samples from class 4 were detected as in-distribution by the OOD oracle, and this class was excluded from the experiment. DK could not be applied to this dataset due to huge memory requirement.

Method	OOD class: 0		OOD class: 1		OOD class: 2		OOD class: 3		OOD class: 5		OOD class: 6	
	In	OOD	In	OOD	In	OOD	In	OOD	In	OOD	In	OOD
Pipeline	86.3 ± 0.2	97.6 ± 0.1	86.4 ± 0.4	96.7 ± 0.3	81.5 ± 0.4	98.6 ± 0.0	80.8 ± 0.5	97.3 ± 0.4	83.4 ± 0.3	96.1 ± 0.2	78.8 ± 0.5	95.8 ± 0.2
OCT	79.6 ± 0.4	93.8 ± 0.5	80.7 ± 0.6	93.1 ± 0.5	71.2 ± 0.5	95.8 ± 0.2	73.2 ± 0.5	94.5 ± 0.4	76.9 ± 0.4	93.5 ± 0.3	70.6 ± 0.7	92.7 ± 0.5
MCDD	75.3 ± 0.1	12.8 ± 0.4	71.6 ± 0.6	26.4 ± 1.8	67.9 ± 0.3	34.0 ± 0.7	72.0 ± 0.9	23.7 ± 1.8	76.6 ± 0.2	11.0 ± 0.9	73.4 ± 0.4	15.3 ± 2.6
O-GBDT	71.5 ± 0.1	1.0 ± 0.1	66.7 ± 0.5	9.9 ± 1.1	63.1 ± 0.2	0.0 ± 0.0	67.4 ± 1.2	0.0 ± 0.0	73.7 ± 0.2	0.0 ± 0.0	71.2 ± 0.3	0.1 ± 0.0
DK	-	-	-	-	-	-	-	-	-	-	-	-
Multi	27.8 ± 0.3	2.0 ± 0.4	24.6 ± 0.5	32.1 ± 1.0	26.8 ± 0.3	0.8 ± 0.1	23.6 ± 0.9	0.1 ± 0.1	26.6 ± 0.4	1.1 ± 0.1	26.5 ± 0.2	0.4 ± 0.1
Incremental	5.6 ± 0.4	0.0 ± 0.0	10.2 ± 0.3	0.0 ± 0.0	6.7 ± 0.1	0.0 ± 0.0	6.9 ± 0.2	0.0 ± 0.0	6.3 ± 0.1	0.0 ± 0.0	4.0 ± 0.2	0.0 ± 0.0
Energy+	76.0 ± 0.5	81.5 ± 1.7	72.7 ± 0.7	74.2 ± 1.6	70.6 ± 0.6	94.5 ± 0.3	69.4 ± 1.2	83.8 ± 1.3	77.1 ± 0.4	87.8 ± 0.8	71.1 ± 0.9	84.2 ± 1.5
Self	71.4 ± 0.4	46.5 ± 1.9	71.8 ± 0.5	68.0 ± 3.4	58.2 ± 0.3	0.7 ± 0.1	64.8 ± 1.2	1.0 ± 0.3	72.8 ± 0.2	1.9 ± 0.3	71.2 ± 0.4	72.1 ± 2.7
Exposure	73.8 ± 0.4	35.4 ± 2.5	69.4 ± 0.5	40.0 ± 2.4	63.1 ± 0.3	7.2 ± 0.7	67.0 ± 1.2	4.5 ± 1.9	75.4 ± 0.2	4.4 ± 1.0	71.9 ± 0.5	42.1 ± 3.4
POEM	62.7 ± 0.2	0.2 ± 0.1	60.3 ± 0.9	22.1 ± 1.5	56.8 ± 0.1	51.3 ± 2.2	62.3 ± 0.8	2.9 ± 0.8	69.4 ± 0.3	19.5 ± 1.6	64.5 ± 0.4	0.0 ± 0.0
WOODS	66.1 ± 0.4	0.6 ± 0.3	62.4 ± 0.8	14.6 ± 3.2	58.3 ± 0.5	0.0 ± 0.0	62.8 ± 1.1	0.0 ± 0.0	71.3 ± 0.2	0.0 ± 0.0	69.5 ± 0.4	0.4 ± 0.2
Logitnorm	69.9 ± 0.9	1.6 ± 0.9	64.7 ± 1.2	4.4 ± 0.8	61.2 ± 0.5	3.0 ± 1.3	64.3 ± 1.7	1.1 ± 0.6	73.4 ± 0.9	1.1 ± 0.6	70.1 ± 0.4	0.3 ± 0.2
VOS	68.2 ± 0.3	0.0 ± 0.0	61.5 ± 0.8	0.1 ± 0.0	59.2 ± 0.4	25.1 ± 3.5	66.3 ± 0.0	0.0 ± 0.0	70.0 ± 0.1	0.4 ± 0.1	68.7 ± 0.3	0.0 ± 0.0
ReAct	67.7 ± 0.9	4.9 ± 2.4	64.9 ± 0.7	19.6 ± 3.1	55.3 ± 1.4	3.4 ± 1.6	59.2 ± 1.8	0.6 ± 0.5	70.2 ± 0.8	0.2 ± 0.1	66.9 ± 0.9	7.3 ± 3.2
Energy	71.3 ± 0.3	3.0 ± 1.3	67.4 ± 0.7	20.6 ± 3.3	62.5 ± 0.3	3.5 ± 1.7	65.9 ± 1.0	0.1 ± 0.1	74.4 ± 0.2	0.4 ± 0.2	71.4 ± 0.4	1.7 ± 0.9
Confidence	72.2 ± 0.2	0.3 ± 0.1	66.3 ± 0.7	4.2 ± 0.8	63.2 ± 0.3	0.3 ± 0.0	66.8 ± 1.0	0.0 ± 0.0	75.2 ± 0.3	0.1 ± 0.0	72.2 ± 0.4	0.1 ± 0.0
Original	71.8 ± 0.3	0.0 ± 0.0	66.3 ± 0.7	0.0 ± 0.0	62.8 ± 0.4	0.0 ± 0.0	66.2 ± 1.0	0.0 ± 0.0	74.9 ± 0.2	0.0 ± 0.0	72.1 ± 0.3	0.0 ± 0.0

Table 4. Detailed results for the Dilbert dataset in Test settings I. For this dataset, all samples from classes 1, 2, and 4 were detected as in-distribution by the OOD oracle, and these classes were excluded from the experiment. DK could not be applied to this dataset due to huge memory requirement.

Method	OOD class: 0		OOD class: 3	
	In	OOD	In	OOD
Pipeline	95.2 ± 0.6	98.8 ± 0.1	96.1 ± 0.2	98.9 ± 0.1
OCT	78.1 ± 1.1	68.9 ± 2.7	84.8 ± 0.9	91.1 ± 0.6
MCDD	74.0 ± 0.3	40.8 ± 2.6	76.7 ± 0.6	39.7 ± 3.7
O-GBDT	75.9 ± 0.8	59.6 ± 2.0	75.1 ± 0.4	69.2 ± 1.3
DK	-	-	-	-
Multi	51.0 ± 1.2	68.5 ± 2.0	59.2 ± 0.6	97.8 ± 0.1
Incremental	39.6 ± 1.8	0.0 ± 0.0	35.6 ± 1.6	0.0 ± 0.0
Energy+	78.0 ± 0.9	62.7 ± 2.8	78.4 ± 0.6	75.0 ± 1.6
Self	64.8 ± 4.0	12.8 ± 3.9	79.4 ± 0.6	67.5 ± 2.0
Exposure	81.5 ± 0.9	64.0 ± 2.9	84.7 ± 0.6	80.4 ± 1.5
POEM	61.0 ± 0.8	0.0 ± 0.0	73.9 ± 0.3	0.2 ± 0.1
WOODS	58.0 ± 1.2	2.7 ± 2.5	60.4 ± 0.7	3.3 ± 1.5
Logitnorm	72.4 ± 0.4	4.5 ± 1.7	74.6 ± 0.7	10.3 ± 2.6
VOS	70.1 ± 0.4	0.0 ± 0.0	71.7 ± 0.1	0.0 ± 0.0
ReAct	73.6 ± 1.0	15.8 ± 4.4	72.9 ± 0.8	13.3 ± 3.8
Energy	73.8 ± 0.8	17.6 ± 4.9	73.3 ± 0.4	6.7 ± 2.7
Confidence	72.2 ± 0.9	27.2 ± 3.6	71.7 ± 2.9	19.8 ± 3.1
Original	72.4 ± 0.6	0.0 ± 0.0	73.4 ± 0.3	0.0 ± 0.0

Table 5. Detailed results for the GMSC dataset in Test settings I.

Method	OOD class: 0		OOD class: 1	
	In	OOD	In	OOD
Pipeline	98.9 ± 0.0	99.0 ± 0.0	98.7 ± 0.0	98.8 ± 0.0
OCT	95.1 ± 1.2	94.7 ± 1.4	89.1 ± 0.9	86.9 ± 1.5
MCDD	66.8 ± 0.0	1.2 ± 0.0	82.9 ± 0.0	75.4 ± 0.0
O-GBDT	67.7 ± 0.1	8.9 ± 0.5	66.8 ± 0.0	3.1 ± 0.2
DK	66.8 ± 0.0	1.0 ± 0.2	67.0 ± 0.0	3.6 ± 0.4
Multi	94.6 ± 0.0	95.1 ± 0.0	95.1 ± 0.0	95.6 ± 0.0
Incremental	66.9 ± 0.1	1.7 ± 0.5	67.7 ± 0.1	8.8 ± 1.1
Energy+	91.1 ± 1.4	89.4 ± 2.1	79.9 ± 0.7	68.7 ± 1.7
Self	74.3 ± 18.6	86.9 ± 5.3	54.0 ± 21.0	77.7 ± 5.0
Exposure	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0
POEM	82.2 ± 0.0	73.8 ± 0.0	80.3 ± 0.0	69.7 ± 0.0
WOODS	71.9 ± 3.8	28.6 ± 17.5	70.3 ± 1.7	29.6 ± 9.1
Logitnorm	66.6 ± 0.3	6.0 ± 2.4	68.9 ± 1.1	21.8 ± 7.4
VOS	66.2 ± 0.0	3.0 ± 0.0	65.4 ± 0.0	0.0 ± 0.0
ReAct	67.5 ± 0.7	13.0 ± 4.8	67.1 ± 0.4	10.1 ± 3.0
Energy	67.4 ± 0.7	12.1 ± 4.5	68.4 ± 0.9	17.0 ± 5.7
Confidence	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0
Original	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0

Table 6. Detailed results for the Heloc dataset in Test settings I.

Method	OOD class: 0		OOD class: 1	
	In	OOD	In	OOD
Pipeline	98.8 ± 0.0	98.8 ± 0.0	99.1 ± 0.0	99.1 ± 0.0
OCT	88.9 ± 1.4	86.6 ± 2.3	89.4 ± 0.9	87.6 ± 1.4
MCDD	75.1 ± 0.0	53.9 ± 0.0	72.7 ± 0.0	43.2 ± 0.0
O-GBDT	67.8 ± 0.1	10.3 ± 0.9	67.2 ± 0.3	5.8 ± 3.5
DK	67.0 ± 0.1	3.0 ± 0.5	66.7 ± 0.0	0.0 ± 0.0
Multi	95.0 ± 0.0	95.5 ± 0.0	96.4 ± 0.0	96.6 ± 0.0
Incremental	67.1 ± 0.1	3.4 ± 0.6	66.7 ± 0.0	0.0 ± 0.0
Energy+	81.6 ± 0.4	72.7 ± 0.9	80.4 ± 0.8	69.4 ± 1.9
Self	91.9 ± 0.7	91.3 ± 1.0	85.8 ± 1.1	81.6 ± 2.4
Exposure	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0
POEM	78.9 ± 0.0	65.6 ± 0.0	72.4 ± 0.0	44.7 ± 0.0
WOODS	67.2 ± 1.7	9.8 ± 9.8	69.2 ± 1.5	22.1 ± 10.1
Logitnorm	67.6 ± 1.3	14.1 ± 8.0	67.8 ± 0.4	11.6 ± 2.7
VOS	66.6 ± 0.0	8.2 ± 0.0	67.1 ± 0.0	9.8 ± 0.0
ReAct	67.9 ± 0.8	15.9 ± 4.8	67.0 ± 0.9	12.8 ± 6.5
Energy	67.9 ± 0.8	17.6 ± 5.4	67.8 ± 0.8	17.0 ± 6.1
Confidence	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0
Original	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0

Table 7. Detailed results for the Jannis dataset in Test settings I. For this dataset, all samples from class 0 were detected as in-distribution by the OOD oracle, and this class was excluded from the experiment. DK could not be applied to this dataset due to huge memory requirement.

Method	OOD class: 1		OOD class: 2		OOD class: 3	
	In	OOD	In	OOD	In	OOD
Pipeline	63.3 ± 1.1	98.8 ± 0.1	52.2 ± 0.7	99.2 ± 0.1	51.8 ± 0.2	99.1 ± 0.0
OCT	52.2 ± 2.4	87.9 ± 1.0	41.9 ± 2.0	88.1 ± 1.0	44.9 ± 0.6	91.7 ± 0.5
MCDD	42.0 ± 0.9	12.8 ± 0.3	29.2 ± 0.7	3.5 ± 0.5	25.2 ± 0.4	2.1 ± 0.2
O-GBDT	42.8 ± 1.0	1.5 ± 0.2	43.2 ± 1.5	1.2 ± 0.2	42.1 ± 0.2	17.4 ± 1.7
DK	-	-	-	-	-	-
Multi	41.6 ± 0.3	57.2 ± 1.2	34.4 ± 0.3	44.9 ± 1.1	38.9 ± 0.4	81.6 ± 1.2
Incremental	11.7 ± 1.0	0.0 ± 0.0	12.9 ± 0.9	1.9 ± 0.2	17.4 ± 0.7	0.1 ± 0.0
Energy+	45.8 ± 0.8	49.7 ± 1.7	40.3 ± 0.6	38.9 ± 2.3	46.4 ± 0.5	75.1 ± 2.3
Self	41.3 ± 1.9	49.8 ± 3.0	32.4 ± 2.0	31.3 ± 3.5	38.0 ± 1.5	78.9 ± 1.4
Exposure	44.9 ± 0.8	31.9 ± 2.5	37.8 ± 1.6	9.2 ± 1.4	42.4 ± 0.3	9.5 ± 1.3
POEM	44.9 ± 1.1	50.5 ± 0.4	39.9 ± 0.5	61.1 ± 0.6	39.0 ± 0.2	14.7 ± 0.5
WOODS	45.5 ± 1.0	41.7 ± 4.4	39.9 ± 0.5	50.5 ± 5.9	39.9 ± 0.4	8.5 ± 1.2
Logitnorm	41.3 ± 0.7	4.3 ± 0.6	36.5 ± 0.6	3.9 ± 1.2	40.4 ± 0.3	0.3 ± 0.1
VOS	40.9 ± 0.7	10.9 ± 0.8	38.3 ± 0.6	1.0 ± 0.2	39.8 ± 0.3	0.0 ± 0.0
ReAct	41.1 ± 0.7	7.7 ± 1.3	36.9 ± 0.5	3.2 ± 1.5	39.8 ± 0.4	2.9 ± 1.5
Energy	41.4 ± 0.7	7.3 ± 1.2	37.3 ± 0.5	3.3 ± 1.5	40.2 ± 0.2	2.6 ± 1.1
Confidence	41.7 ± 0.6	9.6 ± 0.8	36.7 ± 1.0	0.5 ± 0.1	40.5 ± 0.2	1.2 ± 0.4
Original	41.1 ± 0.7	0.0 ± 0.0	37.3 ± 0.4	0.0 ± 0.0	40.4 ± 0.2	0.0 ± 0.0

Table 8. Detailed results for the Adult dataset in Test settings II.

Method	T_1		T_2		T_3		T_4	
	Actual	OOD	Actual	OOD	Actual	OOD	Actual	OOD
Pipeline	74.8 ± 0.6	99.1 ± 0.0	74.8 ± 0.6	99.1 ± 0.0	74.8 ± 0.6	99.1 ± 0.0	74.8 ± 0.6	99.1 ± 0.0
OCT	74.1 ± 0.5	98.4 ± 0.0	74.1 ± 0.5	98.3 ± 0.0	74.2 ± 0.5	98.4 ± 0.0	74.2 ± 0.5	98.5 ± 0.1
MCDD	52.8 ± 0.0	27.3 ± 0.0	53.8 ± 0.0	33.2 ± 0.0	55.2 ± 0.0	43.3 ± 0.0	56.1 ± 0.0	42.7 ± 0.0
O-GBDT	57.6 ± 0.0	52.4 ± 0.0	57.9 ± 0.0	55.1 ± 0.0	56.8 ± 0.0	49.4 ± 0.0	78.0 ± 0.1	100.0 ± 0.0
DK	57.0 ± 0.1	56.7 ± 1.0	57.3 ± 0.2	58.1 ± 1.2	59.5 ± 0.2	67.3 ± 1.0	71.0 ± 0.3	94.8 ± 0.4
Multi	62.7 ± 0.0	90.6 ± 0.0	62.7 ± 0.0	90.6 ± 0.0	62.7 ± 0.0	90.6 ± 0.0	62.7 ± 0.0	90.6 ± 0.0
Incremental	39.6 ± 1.0	53.3 ± 2.5	40.2 ± 1.1	54.9 ± 2.5	41.1 ± 1.1	62.9 ± 2.7	48.3 ± 1.7	93.8 ± 0.6
Energy+	72.9 ± 1.0	98.6 ± 0.1	72.9 ± 1.0	98.5 ± 0.1	72.9 ± 1.0	98.6 ± 0.1	73.2 ± 1.0	99.0 ± 0.0
Self	72.8 ± 0.7	97.6 ± 0.1	72.6 ± 0.7	97.3 ± 0.1	72.3 ± 0.7	97.0 ± 0.2	73.0 ± 0.7	97.8 ± 0.1
Exposure	64.7 ± 1.7	79.3 ± 4.6	63.2 ± 2.2	73.5 ± 6.8	62.8 ± 2.8	68.9 ± 10.0	65.2 ± 4.0	66.7 ± 14.6
POEM	68.0 ± 0.0	87.2 ± 0.0	70.7 ± 0.0	92.3 ± 0.0	71.6 ± 0.0	93.8 ± 0.0	75.4 ± 0.0	99.0 ± 0.0
WOODS	54.8 ± 2.3	47.2 ± 19.3	56.8 ± 2.8	51.6 ± 21.1	58.1 ± 3.1	54.0 ± 22.0	62.0 ± 4.2	59.4 ± 24.3
Logitnorm	50.0 ± 0.3	7.4 ± 3.5	50.5 ± 0.2	8.3 ± 3.4	50.6 ± 0.2	9.5 ± 3.4	54.0 ± 0.2	6.6 ± 5.1
VOS	48.5 ± 0.0	0.1 ± 0.0	49.1 ± 0.0	0.4 ± 0.0	49.6 ± 0.0	0.9 ± 0.0	52.8 ± 0.0	0.1 ± 0.0
ReAct	50.3 ± 0.4	11.8 ± 4.6	50.6 ± 0.5	11.4 ± 4.2	50.7 ± 0.5	11.3 ± 3.6	52.7 ± 1.0	4.1 ± 2.1
Energy	50.9 ± 0.6	13.0 ± 6.6	51.5 ± 0.6	13.8 ± 7.3	51.6 ± 0.8	15.4 ± 8.1	54.7 ± 0.4	9.4 ± 7.6
Confidence	50.2 ± 0.2	2.8 ± 0.3	50.6 ± 0.3	2.4 ± 0.3	50.5 ± 0.3	2.4 ± 0.5	54.0 ± 0.4	0.6 ± 0.2
Original	50.1 ± 0.2	0.0 ± 0.0	50.6 ± 0.2	0.0 ± 0.0	50.6 ± 0.2	0.0 ± 0.0	54.5 ± 0.3	0.0 ± 0.0

Table 9. Detailed results for the Compas dataset in Test settings II.

Method	T_1		T_2		T_3		T_4	
	Actual	OOD	Actual	OOD	Actual	OOD	Actual	OOD
Pipeline	67.6 ± 0.7	98.7 ± 0.0	67.6 ± 0.7	98.7 ± 0.0	67.6 ± 0.7	98.7 ± 0.0	67.6 ± 0.7	98.7 ± 0.0
OCT	67.3 ± 0.7	97.9 ± 0.1	67.3 ± 0.7	98.0 ± 0.1	67.3 ± 0.7	98.0 ± 0.1	67.3 ± 0.7	98.0 ± 0.1
MCDD	54.1 ± 0.0	57.5 ± 0.3	55.2 ± 0.1	61.8 ± 0.1	57.1 ± 0.0	69.0 ± 0.4	58.5 ± 0.3	73.5 ± 0.4
O-GBDT	52.9 ± 0.1	63.0 ± 0.0	52.0 ± 0.1	58.8 ± 0.0	51.2 ± 0.1	54.7 ± 0.0	69.1 ± 0.1	100.0 ± 0.0
DK	51.8 ± 0.8	54.6 ± 4.8	53.0 ± 1.0	60.9 ± 4.9	55.1 ± 1.0	69.8 ± 3.9	63.7 ± 1.6	91.7 ± 3.0
Multi	51.0 ± 0.0	85.5 ± 0.0	51.0 ± 0.0	85.5 ± 0.0	51.0 ± 0.0	85.5 ± 0.0	51.0 ± 0.0	85.5 ± 0.0
Incremental	40.4 ± 1.8	38.1 ± 2.5	40.6 ± 1.9	41.1 ± 2.4	41.9 ± 1.9	50.9 ± 2.7	49.0 ± 2.5	83.8 ± 3.2
Energy+	66.4 ± 0.9	97.7 ± 0.2	66.6 ± 0.8	97.9 ± 0.1	66.7 ± 0.8	98.2 ± 0.1	67.1 ± 0.8	98.5 ± 0.0
Self	66.3 ± 0.7	97.3 ± 0.4	66.3 ± 0.7	97.3 ± 0.4	66.4 ± 0.7	97.4 ± 0.4	66.5 ± 0.7	97.5 ± 0.4
Exposure	49.3 ± 0.8	35.3 ± 10.5	49.3 ± 0.9	36.0 ± 10.7	49.3 ± 0.9	33.3 ± 10.9	50.0 ± 1.3	33.9 ± 12.4
POEM	56.0 ± 0.0	89.2 ± 0.0	57.6 ± 0.0	92.5 ± 0.0	58.8 ± 0.0	94.8 ± 0.0	61.4 ± 0.0	98.7 ± 0.0
WOODS	55.8 ± 0.6	84.5 ± 0.7	58.3 ± 0.6	90.3 ± 0.4	59.7 ± 0.5	93.3 ± 0.2	63.0 ± 0.5	98.3 ± 0.1
Logitnorm	45.7 ± 0.7	16.4 ± 5.2	45.8 ± 0.7	16.9 ± 5.5	46.3 ± 0.9	17.8 ± 6.9	46.7 ± 0.8	18.1 ± 7.1
VOS	45.0 ± 0.0	21.7 ± 0.0	45.5 ± 0.0	26.5 ± 0.0	46.4 ± 0.0	33.7 ± 0.0	48.7 ± 0.0	49.4 ± 0.0
ReAct	45.8 ± 0.5	9.3 ± 4.5	45.6 ± 0.4	9.1 ± 3.1	46.1 ± 0.4	10.3 ± 2.6	46.0 ± 0.4	9.3 ± 3.3
Energy	45.7 ± 0.4	7.4 ± 3.6	45.5 ± 0.3	7.1 ± 2.7	45.9 ± 0.3	7.7 ± 3.0	45.8 ± 0.4	6.7 ± 3.3
Confidence	43.9 ± 0.6	5.7 ± 1.4	44.1 ± 0.6	4.7 ± 1.4	44.7 ± 0.6	4.3 ± 1.4	45.4 ± 0.6	4.4 ± 1.6
Original	45.5 ± 0.3	0.0 ± 0.0	45.4 ± 0.3	0.0 ± 0.0	45.8 ± 0.3	0.0 ± 0.0	45.8 ± 0.3	0.0 ± 0.0

Table 10. Detailed results for the Cover dataset in Test settings II. DK could not be applied to this dataset due to huge memory requirement.

Method	T_1		T_2		T_3		T_4	
	Actual	OOD	Actual	OOD	Actual	OOD	Actual	OOD
Pipeline	66.1 ± 0.3	99.0 ± 0.0	66.1 ± 0.3	99.0 ± 0.0	66.1 ± 0.3	99.0 ± 0.0	66.1 ± 0.3	99.0 ± 0.0
OCT	62.0 ± 0.3	98.3 ± 0.1	62.0 ± 0.3	98.3 ± 0.1	62.1 ± 0.3	98.4 ± 0.1	62.1 ± 0.3	98.4 ± 0.1
MCDD	66.1 ± 0.4	35.3 ± 1.9	66.7 ± 0.4	37.8 ± 2.0	68.5 ± 0.4	51.3 ± 3.3	75.5 ± 0.5	68.7 ± 7.7
O-GBDT	50.0 ± 0.1	69.1 ± 0.5	50.9 ± 0.1	78.5 ± 0.2	58.6 ± 0.1	95.9 ± 0.1	67.2 ± 0.1	100.0 ± 0.0
DK	-	-	-	-	-	-	-	-
Multi	17.7 ± 0.1	38.7 ± 0.5	19.1 ± 0.1	61.8 ± 0.4	25.8 ± 0.2	98.3 ± 0.0	27.5 ± 0.2	100.0 ± 0.0
Incremental	7.3 ± 0.1	12.1 ± 0.4	7.5 ± 0.1	13.4 ± 0.4	8.4 ± 0.1	49.6 ± 0.4	9.2 ± 0.1	99.9 ± 0.0
Energy+	66.9 ± 0.5	99.0 ± 0.0	67.0 ± 0.5	99.0 ± 0.0	67.0 ± 0.5	99.0 ± 0.0	67.0 ± 0.5	99.0 ± 0.0
Self	62.7 ± 0.5	99.8 ± 0.0	62.9 ± 0.5	99.8 ± 0.0	62.9 ± 0.5	99.9 ± 0.0	63.0 ± 0.5	99.9 ± 0.0
Exposure	71.2 ± 0.4	98.9 ± 0.0	71.4 ± 0.4	99.0 ± 0.0	71.5 ± 0.4	99.0 ± 0.0	71.5 ± 0.4	99.0 ± 0.0
POEM	51.2 ± 0.1	91.3 ± 0.3	51.8 ± 0.1	93.1 ± 0.2	53.9 ± 0.1	97.1 ± 0.0	57.3 ± 0.2	99.0 ± 0.0
WOODS	43.4 ± 0.7	2.3 ± 0.6	43.8 ± 0.7	2.1 ± 0.6	44.1 ± 0.6	1.3 ± 0.2	51.2 ± 0.7	0.0 ± 0.0
Logitnorm	52.4 ± 0.8	0.9 ± 0.3	53.0 ± 0.8	0.7 ± 0.3	54.3 ± 0.9	0.5 ± 0.3	57.3 ± 0.9	0.2 ± 0.2
VOS	35.7 ± 0.4	5.9 ± 0.2	34.3 ± 0.3	5.7 ± 0.2	31.0 ± 0.2	2.6 ± 0.1	34.0 ± 0.4	1.4 ± 0.0
ReAct	53.9 ± 0.9	6.0 ± 1.6	54.3 ± 1.0	6.3 ± 1.7	55.0 ± 1.1	7.5 ± 2.2	57.0 ± 0.9	7.1 ± 3.7
Energy	55.3 ± 0.7	5.8 ± 2.2	55.7 ± 0.7	6.1 ± 2.4	56.5 ± 0.8	7.6 ± 3.2	60.1 ± 0.8	7.3 ± 4.8
Confidence	51.5 ± 1.1	0.4 ± 0.1	51.8 ± 1.1	0.3 ± 0.1	52.8 ± 1.3	0.1 ± 0.0	57.6 ± 1.2	0.0 ± 0.0
Original	55.3 ± 0.7	0.0 ± 0.0	55.7 ± 0.7	0.0 ± 0.0	56.4 ± 0.8	0.0 ± 0.0	60.1 ± 0.7	0.0 ± 0.0

Table 11. Detailed results for the Dilbert dataset in Test settings II. DK could not be applied to this dataset due to huge memory requirement.

Method	T_1		T_2		T_3		T_4	
	Actual	OOD	Actual	OOD	Actual	OOD	Actual	OOD
Pipeline	89.9 ± 3.5	99.1 ± 0.0	89.9 ± 3.5	99.1 ± 0.0	89.9 ± 3.5	99.1 ± 0.0	89.9 ± 3.5	99.1 ± 0.0
OCT	87.3 ± 0.6	96.3 ± 0.2	87.1 ± 0.6	96.1 ± 0.2	88.3 ± 0.6	97.5 ± 0.3	88.3 ± 0.6	97.5 ± 0.3
MCDD	69.4 ± 0.2	31.3 ± 1.7	68.9 ± 0.2	29.2 ± 1.6	72.0 ± 0.2	28.1 ± 2.5	76.3 ± 0.7	30.1 ± 4.4
O-GBDT	81.0 ± 0.2	91.6 ± 0.2	81.3 ± 0.2	92.0 ± 0.2	85.2 ± 0.1	97.7 ± 0.1	86.9 ± 0.1	99.8 ± 0.0
DK	-	-	-	-	-	-	-	-
Multi	53.6 ± 0.2	98.3 ± 0.0	53.9 ± 0.1	98.9 ± 0.0	54.0 ± 0.1	99.2 ± 0.0	54.0 ± 0.1	99.2 ± 0.0
Incremental	27.8 ± 1.5	0.2 ± 0.0	27.7 ± 1.5	0.2 ± 0.0	28.0 ± 1.5	3.7 ± 0.2	40.5 ± 2.0	96.0 ± 4.0
Energy+	87.1 ± 0.4	96.8 ± 0.1	87.1 ± 0.4	96.8 ± 0.1	88.4 ± 0.4	98.3 ± 0.1	88.4 ± 0.4	98.3 ± 0.1
Self	76.9 ± 5.2	92.9 ± 1.6	76.1 ± 5.1	91.7 ± 1.5	80.0 ± 5.4	96.8 ± 1.8	80.0 ± 5.4	96.8 ± 1.8
Exposure	90.4 ± 0.2	95.3 ± 0.1	90.2 ± 0.3	95.0 ± 0.1	92.8 ± 0.2	98.0 ± 0.1	92.4 ± 0.4	97.1 ± 0.9
POEM	61.5 ± 0.3	44.8 ± 0.3	61.8 ± 0.3	48.9 ± 0.8	69.4 ± 0.3	77.2 ± 0.3	79.3 ± 0.3	99.0 ± 0.0
WOODS	62.9 ± 0.7	59.6 ± 5.2	62.8 ± 0.7	60.1 ± 5.2	72.4 ± 1.0	80.4 ± 7.2	74.7 ± 0.9	83.1 ± 7.4
Logitnorm	67.0 ± 0.2	13.0 ± 1.8	66.9 ± 0.2	11.9 ± 1.7	70.2 ± 0.3	10.2 ± 1.6	75.6 ± 0.6	2.5 ± 1.1
VOS	62.7 ± 0.4	3.2 ± 0.4	62.8 ± 0.3	2.1 ± 0.3	66.4 ± 0.5	0.4 ± 0.1	72.1 ± 0.4	0.0 ± 0.0
ReAct	62.7 ± 2.5	9.0 ± 1.1	62.6 ± 2.5	7.8 ± 1.1	65.6 ± 2.6	5.3 ± 1.3	72.2 ± 2.9	2.9 ± 2.4
Energy	64.1 ± 2.5	9.5 ± 1.4	64.0 ± 2.5	8.0 ± 1.4	67.0 ± 2.7	5.3 ± 1.8	72.8 ± 2.9	1.1 ± 0.8
Confidence	64.1 ± 0.5	15.7 ± 1.6	63.9 ± 0.5	14.1 ± 1.5	65.8 ± 0.6	13.8 ± 1.6	71.8 ± 0.8	5.7 ± 1.0
Original	63.7 ± 2.5	0.0 ± 0.0	63.8 ± 2.5	0.0 ± 0.0	67.2 ± 2.7	0.0 ± 0.0	73.4 ± 3.0	0.0 ± 0.0

Table 12. Detailed results for the GMSC dataset in Test settings II.

Method	T_1		T_2		T_3		T_4	
	Actual	OOD	Actual	OOD	Actual	OOD	Actual	OOD
Pipeline	58.2 ± 0.7	99.1 ± 0.0	58.2 ± 0.7	99.1 ± 0.0	58.2 ± 0.7	99.1 ± 0.0	58.2 ± 0.7	99.1 ± 0.0
OCT	60.3 ± 0.5	99.3 ± 0.1	60.4 ± 0.5	99.3 ± 0.1	60.4 ± 0.5	99.4 ± 0.1	60.4 ± 0.5	99.4 ± 0.1
MCDD	46.0 ± 0.0	40.2 ± 0.0	46.2 ± 0.0	42.8 ± 0.0	48.8 ± 0.0	66.4 ± 0.0	49.9 ± 0.0	70.0 ± 0.0
O-GBDT	52.0 ± 0.2	79.4 ± 0.4	51.5 ± 0.2	77.6 ± 0.5	48.4 ± 0.2	66.7 ± 0.8	62.8 ± 0.1	100.0 ± 0.0
DK	48.4 ± 0.8	57.7 ± 1.4	48.8 ± 0.8	63.3 ± 1.1	50.2 ± 0.9	77.2 ± 0.7	60.5 ± 1.5	100.0 ± 0.0
Multi	49.3 ± 0.0	98.0 ± 0.0	49.4 ± 0.0	98.1 ± 0.0	49.5 ± 0.0	98.3 ± 0.0	49.5 ± 0.0	98.3 ± 0.0
Incremental	35.6 ± 2.8	48.3 ± 1.3	36.3 ± 2.9	54.4 ± 1.5	37.7 ± 3.2	76.0 ± 0.3	43.6 ± 4.0	100.0 ± 0.0
Energy+	60.8 ± 1.6	99.0 ± 0.0	60.9 ± 1.6	99.1 ± 0.0	60.9 ± 1.6	99.1 ± 0.0	60.9 ± 1.6	99.1 ± 0.0
Self	58.8 ± 0.5	97.4 ± 0.1	59.4 ± 0.7	98.0 ± 0.2	58.9 ± 0.7	97.5 ± 0.3	60.1 ± 0.8	98.5 ± 0.3
Exposure	60.3 ± 0.7	95.0 ± 0.4	60.9 ± 0.7	95.5 ± 0.4	61.5 ± 0.8	95.6 ± 0.6	64.7 ± 0.7	99.0 ± 0.0
POEM	55.1 ± 0.0	91.0 ± 0.0	56.8 ± 0.0	94.1 ± 0.0	58.8 ± 0.0	96.5 ± 0.0	61.4 ± 0.0	99.1 ± 0.0
WOODS	46.0 ± 0.5	13.3 ± 13.3	46.5 ± 0.5	14.6 ± 14.6	47.4 ± 1.2	18.5 ± 18.5	48.5 ± 1.8	19.8 ± 19.8
Logitnorm	44.3 ± 1.0	9.8 ± 2.8	44.0 ± 1.0	10.7 ± 3.1	42.6 ± 1.3	13.7 ± 3.8	42.8 ± 1.6	7.3 ± 3.6
VOS	45.3 ± 0.0	5.2 ± 0.0	45.6 ± 0.0	4.4 ± 0.0	45.5 ± 0.0	4.3 ± 0.0	47.4 ± 0.0	0.1 ± 0.0
ReAct	42.8 ± 1.3	9.5 ± 5.8	42.5 ± 1.3	10.4 ± 6.5	41.5 ± 1.2	9.8 ± 6.5	42.2 ± 1.7	10.4 ± 9.4
Energy	41.4 ± 0.8	11.0 ± 5.5	41.1 ± 0.9	11.8 ± 5.9	40.2 ± 0.7	10.2 ± 5.2	41.2 ± 0.9	11.2 ± 6.2
Confidence	41.1 ± 0.3	2.1 ± 0.6	40.6 ± 0.3	2.4 ± 0.6	39.8 ± 0.2	1.7 ± 0.2	40.5 ± 0.2	0.7 ± 0.2
Original	41.1 ± 0.5	0.0 ± 0.0	40.7 ± 0.5	0.0 ± 0.0	40.0 ± 0.4	0.0 ± 0.0	40.4 ± 0.7	0.0 ± 0.0

Table 13. Detailed results for the Heloc dataset in Test settings II.

Method	T_1		T_2		T_3		T_4	
	Actual	OOD	Actual	OOD	Actual	OOD	Actual	OOD
Pipeline	72.6 ± 0.1	98.8 ± 0.0	72.6 ± 0.1	98.8 ± 0.0	72.6 ± 0.1	98.8 ± 0.0	72.6 ± 0.1	98.8 ± 0.0
OCT	70.5 ± 0.2	97.2 ± 0.1	70.7 ± 0.3	97.5 ± 0.2	70.8 ± 0.3	97.6 ± 0.2	70.7 ± 0.3	97.5 ± 0.2
MCDD	52.5 ± 0.0	46.0 ± 0.0	52.5 ± 0.0	46.8 ± 0.0	57.9 ± 0.0	72.0 ± 0.0	60.0 ± 0.0	78.0 ± 0.0
O-GBDT	65.2 ± 0.3	67.5 ± 0.7	65.3 ± 0.3	68.2 ± 0.6	65.4 ± 0.3	68.5 ± 0.7	80.6 ± 0.1	99.9 ± 0.0
DK	51.5 ± 0.4	14.9 ± 0.7	51.8 ± 0.4	16.1 ± 0.8	59.3 ± 0.4	65.5 ± 1.7	73.1 ± 0.3	100.0 ± 0.0
Multi	69.0 ± 0.0	97.1 ± 0.0	69.0 ± 0.0	97.2 ± 0.0	69.1 ± 0.0	97.3 ± 0.0	69.1 ± 0.0	97.3 ± 0.0
Incremental	26.0 ± 2.8	10.6 ± 0.2	26.1 ± 2.8	11.8 ± 0.2	30.4 ± 2.8	57.9 ± 0.2	39.0 ± 3.3	100.0 ± 0.0
Energy+	70.6 ± 0.2	97.3 ± 0.1	71.1 ± 0.3	98.0 ± 0.0	71.3 ± 0.3	98.3 ± 0.1	71.3 ± 0.3	98.3 ± 0.1
Self	70.8 ± 0.3	97.1 ± 0.1	71.2 ± 0.3	97.7 ± 0.1	71.5 ± 0.3	98.1 ± 0.0	71.5 ± 0.3	98.1 ± 0.0
Exposure	55.2 ± 0.6	44.3 ± 5.6	55.5 ± 0.6	45.4 ± 5.6	55.0 ± 1.0	40.1 ± 6.0	54.4 ± 0.5	19.4 ± 8.8
POEM	57.9 ± 0.0	68.2 ± 0.0	57.8 ± 0.0	68.0 ± 0.0	67.3 ± 0.0	92.4 ± 0.0	71.4 ± 0.0	98.8 ± 0.0
WOODS	30.3 ± 4.4	55.2 ± 2.7	30.4 ± 4.4	55.7 ± 3.0	37.1 ± 4.8	89.6 ± 0.9	40.7 ± 5.0	98.9 ± 0.0
Logitnorm	49.3 ± 1.9	20.0 ± 3.8	49.4 ± 1.9	19.4 ± 4.0	52.0 ± 2.1	30.7 ± 8.3	54.9 ± 1.6	37.9 ± 16.0
VOS	49.5 ± 0.0	11.5 ± 0.0	49.7 ± 0.0	14.3 ± 0.0	50.8 ± 0.0	17.3 ± 0.0	53.8 ± 0.0	2.3 ± 0.0
ReAct	50.5 ± 0.4	7.3 ± 0.5	50.6 ± 0.3	6.9 ± 0.8	52.2 ± 0.4	7.5 ± 0.6	54.0 ± 0.6	3.1 ± 2.7
Energy	50.9 ± 0.2	8.7 ± 1.0	51.1 ± 0.2	8.6 ± 1.3	52.8 ± 0.2	9.1 ± 2.1	54.8 ± 0.3	5.7 ± 5.1
Confidence	50.0 ± 0.3	3.8 ± 0.2	50.2 ± 0.3	4.0 ± 0.2	51.0 ± 0.3	2.5 ± 0.2	53.9 ± 0.4	1.8 ± 0.9
Original	50.7 ± 0.2	0.0 ± 0.0	50.9 ± 0.2	0.0 ± 0.0	52.5 ± 0.2	0.0 ± 0.0	55.0 ± 0.1	0.0 ± 0.0

Table 14. Detailed results for the Jannis dataset in Test settings II. DK could not be applied to this dataset due to huge memory requirement.

Method	T_1		T_2		T_3		T_4	
	Actual	OOD	Actual	OOD	Actual	OOD	Actual	OOD
Pipeline	50.8 ± 0.2	99.0 ± 0.0	50.8 ± 0.2	99.0 ± 0.0	50.8 ± 0.2	99.0 ± 0.0	50.8 ± 0.2	99.0 ± 0.0
OCT	45.0 ± 0.4	93.7 ± 0.6	45.0 ± 0.4	93.8 ± 0.6	45.0 ± 0.4	93.8 ± 0.6	45.0 ± 0.4	93.8 ± 0.6
MCDD	31.2 ± 0.1	6.6 ± 0.4	31.3 ± 0.1	7.2 ± 0.5	34.2 ± 0.2	37.5 ± 1.3	33.2 ± 0.2	28.4 ± 1.8
O-GBDT	49.4 ± 0.1	90.1 ± 0.2	51.2 ± 0.1	95.2 ± 0.1	53.0 ± 0.2	99.7 ± 0.0	53.1 ± 0.2	100.0 ± 0.0
DK	-	-	-	-	-	-	-	-
Multi	36.8 ± 0.2	88.2 ± 0.4	38.9 ± 0.1	95.2 ± 0.2	40.4 ± 0.1	99.5 ± 0.0	40.4 ± 0.1	99.5 ± 0.0
Incremental	12.1 ± 0.7	0.3 ± 0.0	12.2 ± 0.7	0.5 ± 0.0	15.4 ± 0.8	60.5 ± 0.4	18.9 ± 1.0	99.9 ± 0.0
Energy+	47.2 ± 0.7	97.2 ± 0.2	47.9 ± 0.6	98.8 ± 0.0	48.0 ± 0.6	98.9 ± 0.0	48.0 ± 0.6	98.9 ± 0.0
Self	37.8 ± 2.2	92.1 ± 1.7	38.0 ± 2.2	92.7 ± 1.7	38.2 ± 2.2	93.0 ± 1.6	38.2 ± 2.2	93.0 ± 1.6
Exposure	42.9 ± 0.5	66.3 ± 2.3	44.7 ± 0.7	74.0 ± 2.4	48.3 ± 0.9	83.8 ± 3.0	49.1 ± 0.8	85.0 ± 3.4
POEM	29.8 ± 0.7	19.0 ± 0.3	30.1 ± 0.7	23.2 ± 0.3	39.0 ± 1.0	87.8 ± 0.0	43.5 ± 1.1	99.0 ± 0.0
WOODS	34.3 ± 0.2	11.9 ± 1.3	34.4 ± 0.3	14.2 ± 1.6	41.2 ± 1.1	54.1 ± 8.0	46.3 ± 1.3	63.4 ± 9.7
Logitnorm	34.4 ± 0.2	2.8 ± 0.1	34.4 ± 0.2	2.7 ± 0.1	35.6 ± 0.3	0.8 ± 0.1	37.4 ± 0.3	0.4 ± 0.1
VOS	31.9 ± 0.1	5.4 ± 0.1	31.9 ± 0.1	5.7 ± 0.1	32.0 ± 0.1	4.5 ± 0.2	32.9 ± 0.2	2.6 ± 0.3
ReAct	33.8 ± 0.2	4.6 ± 0.4	33.9 ± 0.2	4.7 ± 0.5	35.3 ± 0.3	6.1 ± 1.6	36.9 ± 0.4	5.0 ± 1.9
Energy	34.4 ± 0.1	4.5 ± 0.3	34.5 ± 0.1	4.7 ± 0.4	35.9 ± 0.2	4.3 ± 0.7	38.0 ± 0.3	2.9 ± 0.8
Confidence	34.8 ± 0.1	4.6 ± 0.5	34.9 ± 0.1	4.7 ± 0.6	36.1 ± 0.3	5.8 ± 1.9	38.0 ± 0.4	5.7 ± 1.7
Original	34.4 ± 0.1	0.0 ± 0.0	34.5 ± 0.1	0.0 ± 0.0	35.8 ± 0.2	0.0 ± 0.0	38.0 ± 0.3	0.0 ± 0.0

Table 15. Detailed counterfactual experiment results including standard errors.

	cf alg.	classifier	success rate ↑	valid rate ↑	numerical cost ↓	categorical cost ↓
Adult	GD	Original	100.0 ± 0.0	90.8 ± 0.8	5.0 ± 0.2	22.9 ± 0.1
		DK	100.0 ± 0.0	84.3 ± 0.5	4.9 ± 0.1	0.9 ± 0.2
		OCT	99.8 ± 0.1	99.4 ± 0.2	7.5 ± 0.4	12.2 ± 2.8
	GS	Original	100.0 ± 0.0	74.8 ± 1.1	4.3 ± 0.1	26.7 ± 1.4
		DK	100.0 ± 0.0	88.8 ± 5.0	4.1 ± 0.1	26.9 ± 0.8
		OCT	100.0 ± 0.0	100.0 ± 0.0	4.5 ± 0.1	29.1 ± 1.1
	CCHVAE	Original	100.0 ± 0.0	99.8 ± 0.1	16.7 ± 0.1	19.4 ± 1.4
		DK	100.0 ± 0.0	99.2 ± 0.3	16.9 ± 0.2	22.1 ± 1.7
		OCT	100.0 ± 0.0	100.0 ± 0.0	17.1 ± 0.1	21.8 ± 1.8
	Revise	Original	99.9 ± 0.1	98.2 ± 0.5	16.1 ± 0.5	13.6 ± 1.1
		DK	100.0 ± 0.0	99.2 ± 0.2	15.9 ± 0.2	15.7 ± 2.1
		OCT	99.9 ± 0.1	99.9 ± 0.1	16.0 ± 0.7	21.9 ± 1.6
Compas	GD	Original	100.0 ± 0.0	83.7 ± 1.2	21.4 ± 1.4	55.1 ± 0.4
		DK	100.0 ± 0.0	90.1 ± 1.1	20.7 ± 0.8	4.9 ± 4.2
		OCT	99.7 ± 0.2	99.4 ± 0.4	20.3 ± 0.8	31.0 ± 2.0
	GS	Original	100.0 ± 0.0	77.5 ± 2.1	10.6 ± 0.6	61.4 ± 1.6
		DK	100.0 ± 0.0	79.2 ± 4.7	12.4 ± 0.5	67.3 ± 2.9
		OCT	100.0 ± 0.0	99.9 ± 0.1	14.3 ± 1.0	63.0 ± 2.2
	CCHVAE	Original	100.0 ± 0.0	100.0 ± 0.0	27.7 ± 0.4	21.2 ± 5.1
		DK	100.0 ± 0.0	93.7 ± 6.3	28.2 ± 0.5	31.1 ± 13.2
		OCT	100.0 ± 0.0	100.0 ± 0.0	26.5 ± 0.2	14.1 ± 3.0
	Revise	Original	93.8 ± 6.0	93.8 ± 6.0	28.8 ± 0.2	9.1 ± 1.2
		DK	99.9 ± 0.1	99.9 ± 0.1	33.4 ± 2.4	14.4 ± 4.6
		OCT	99.9 ± 0.1	99.9 ± 0.1	32.0 ± 1.9	10.0 ± 1.2
GMSC	GD	Original	100.0 ± 0.0	88.0 ± 3.3	18.1 ± 0.4	-
		DK	99.9 ± 0.1	97.8 ± 0.9	18.0 ± 0.5	-
		OCT	90.4 ± 2.9	90.4 ± 2.9	18.5 ± 0.7	-
	GS	Original	100.0 ± 0.0	67.0 ± 2.9	12.4 ± 0.4	-
		DK	99.6 ± 0.1	78.3 ± 11.4	15.1 ± 0.7	-
		OCT	99.4 ± 0.0	99.4 ± 0.1	16.0 ± 0.5	-
	CCHVAE	Original	100.0 ± 0.0	100.0 ± 0.0	15.6 ± 0.2	-
		DK	100.0 ± 0.0	99.9 ± 0.1	15.6 ± 0.3	-
		OCT	100.0 ± 0.0	100.0 ± 0.0	15.8 ± 0.2	-
	Revise	Original	100.0 ± 0.0	100.0 ± 0.0	18.6 ± 0.3	-
		DK	100.0 ± 0.0	100.0 ± 0.0	18.9 ± 0.2	-
		OCT	99.2 ± 0.3	99.2 ± 0.3	18.2 ± 0.3	-
Heloc	GD	Original	100.0 ± 0.0	93.6 ± 2.0	11.8 ± 0.1	-
		DK	100.0 ± 0.0	92.2 ± 1.7	11.3 ± 0.3	-
		OCT	99.6 ± 0.3	99.5 ± 0.2	12.9 ± 0.4	-
	GS	Original	100.0 ± 0.0	96.6 ± 0.5	12.5 ± 0.3	-
		DK	100.0 ± 0.0	96.4 ± 1.2	12.9 ± 0.4	-
		OCT	99.8 ± 0.1	99.4 ± 0.1	12.4 ± 0.8	-
	CCHVAE	Original	100.0 ± 0.0	100.0 ± 0.0	22.9 ± 0.2	-
		DK	100.0 ± 0.0	100.0 ± 0.0	22.4 ± 0.1	-
		OCT	100.0 ± 0.0	100.0 ± 0.0	22.7 ± 0.3	-
	Revise	Original	100.0 ± 0.0	100.0 ± 0.0	23.1 ± 0.2	-
		DK	97.4 ± 2.2	97.4 ± 2.2	21.7 ± 0.3	-
		OCT	96.9 ± 2.6	96.9 ± 2.6	22.5 ± 0.3	-