Latent Outlier Exposure for Anomaly Detection with Contaminated Data

Chen Qiu*[1,2], Aodong Li*[3], Marius Kloft[2], Maja Rudolph[1], Stephan Mandt[3]

[1] BOSCH [2] TECHNISCHE UNIVERSITÄT KAISERSLAUTERN [3] UCIRVINE

Motivation & Problem Setup

Anomaly Detection with Contaminated Training Data.

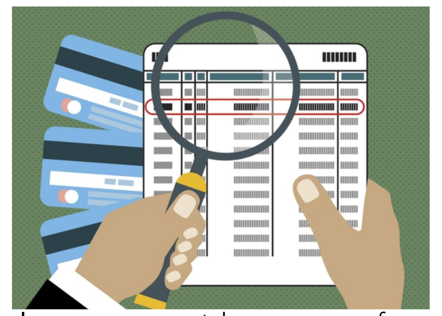
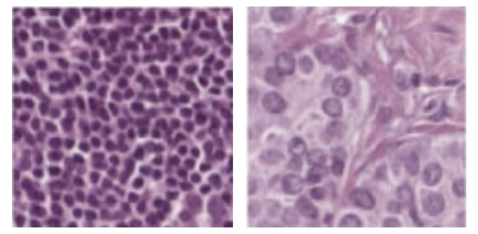
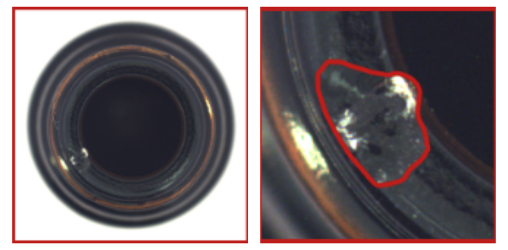


Image taken from https://kunal3836.medium.com/fraud-detection-in-payments-db6d5fc89d13

fraud transaction



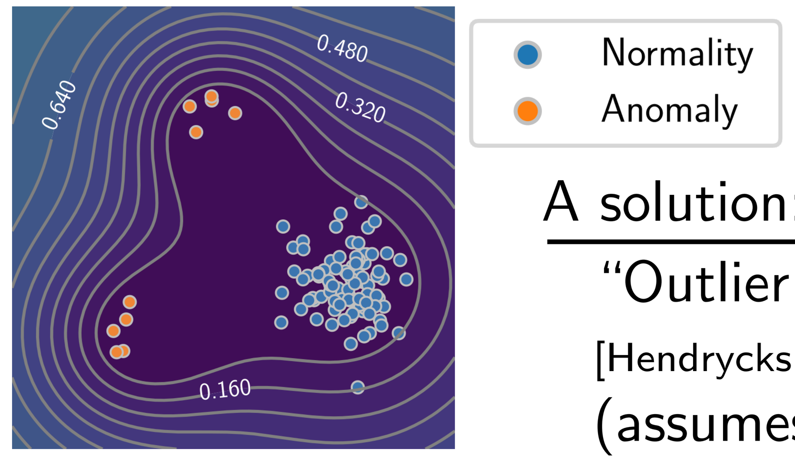
Tissue tumor



Industrial failure

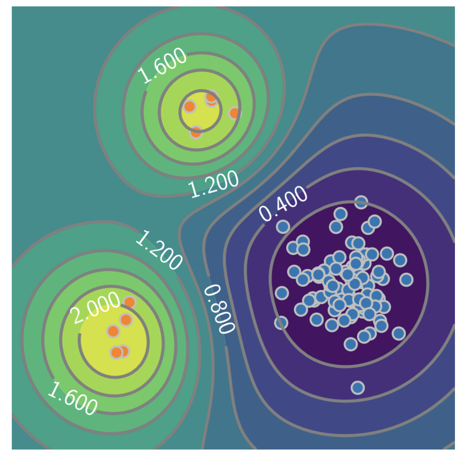
→ Common assumption: **clean** training data.
→ What if the training data contains unnoticed anomalies?

▽ Fig. Anomaly score in input space



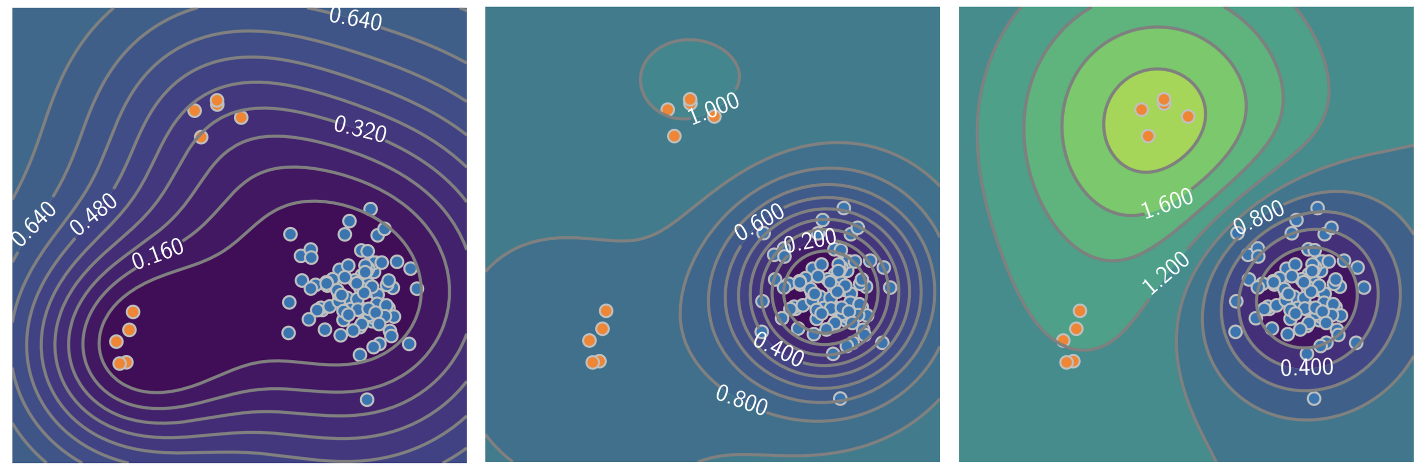
△ Incorrect normal region characterization.

A solution: exploit labels. "Outlier Exposure" [Hendrycks et al., 2018] (assumes synthetic, labeled anomalies)



△ Supervised learning characterizes boundaries well.

→ However, labels are expensive. Can we have a cheaper way?
→ **Contribution:** Unsupervised latent outlier exposure (LOE).



Refine | Soft LOE (ours) | Hard LOE (ours)

Problem Setup. Contaminated training data.
→ Training sets contain many normal samples and a few anomalies.

Method: Latent Outlier Exposure

Proposed Loss.

$$\mathcal{L}(\theta, \mathbf{y}) = \sum_{i=1}^{N}(1-y_i)\mathcal{L}_n^\theta(\mathbf{x}_i) + y_i\mathcal{L}_a^\theta(\mathbf{x}_i)$$

→ Label assignments $\mathbf{y}$ are binary variables to be optimized.
→ $\mathcal{L}_n^\theta(\mathbf{x})$: a normal loss that is designed to be minimized over normal data.
→ $\mathcal{L}_a^\theta(\mathbf{x})$: an abnormal loss that is designed to have the opposite effect.
→ E.g., for deep SVDD, $\mathcal{L}_n^\theta(\mathbf{x}) = ||f_\theta(\mathbf{x}) - \mathbf{c}||^2$ and $\mathcal{L}_a^\theta(\mathbf{x}) = 1/||f_\theta(\mathbf{x}) - \mathbf{c}||^2$.

Constrained Optimization Problem. *Hard* LOE.

$$\min_\theta \min_{\mathbf{y}\in\mathcal{Y}} \mathcal{L}(\theta, \mathbf{y}) \quad \text{s.t. } \mathcal{Y} = \left\{\mathbf{y} \in \{0,1\}^N : \sum_{i=1}^N y_i = \alpha N\right\}$$

→ α is an assumed contamination ratio.
→ Block coordinate descent (EM fashion):
  ▷ (M-step) Perform SGD on θ given current label assignments $\mathbf{y}$;
  ▷ (E-step) Rank data points by score $\mathcal{L}_n^\theta(\mathbf{x}_i) - \mathcal{L}_a^\theta(\mathbf{x}_i)$ and label top α fraction data as anomalies.

Model Extension. *Soft* LOE.

$$\min_\theta \min_{\mathbf{y}\in\mathcal{Y}'} \mathcal{L}(\theta, \mathbf{y}) \quad \text{s.t. } \mathcal{Y}' = \left\{\mathbf{y} \in \{0,0.5\}^N : \sum_{i=1}^N y_i = 0.5\alpha N\right\}$$

Anomaly Score. $S_i^{\text{test}} = \mathcal{L}_n^\theta(\mathbf{x}_i)$
→ Drop $\mathcal{L}_a^\theta(\mathbf{x}_i)$ to account for unknown anomaly types.

Experiment Setup & Findings

For various contamination ratio, compare LOE's performance with baselines.
→ One vs. the rest.
→ Corruption of training set:
  ▷ Mix abnormal samples to have an anomaly ratio of α_0.

Baselines.
→ Blind: ignore anomaly labels and train on all the data.
→ Refine: remove likely anomalies then re-train the model.

Findings. With multiple backbone models (NTL/MHRot/ICL),
→ LOE improve over the best baseline by 2.3% AUC on image data.
→ LOE significantly improves the detector based on 30 tabular datasets.
→ LOE achieves the-state-of-the-art performance on a video benchmark.

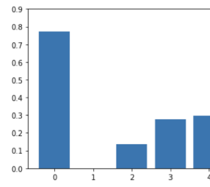
Experiments

Data.

Image.

Video.

Tabular.



Results.

Table. Image benchmark

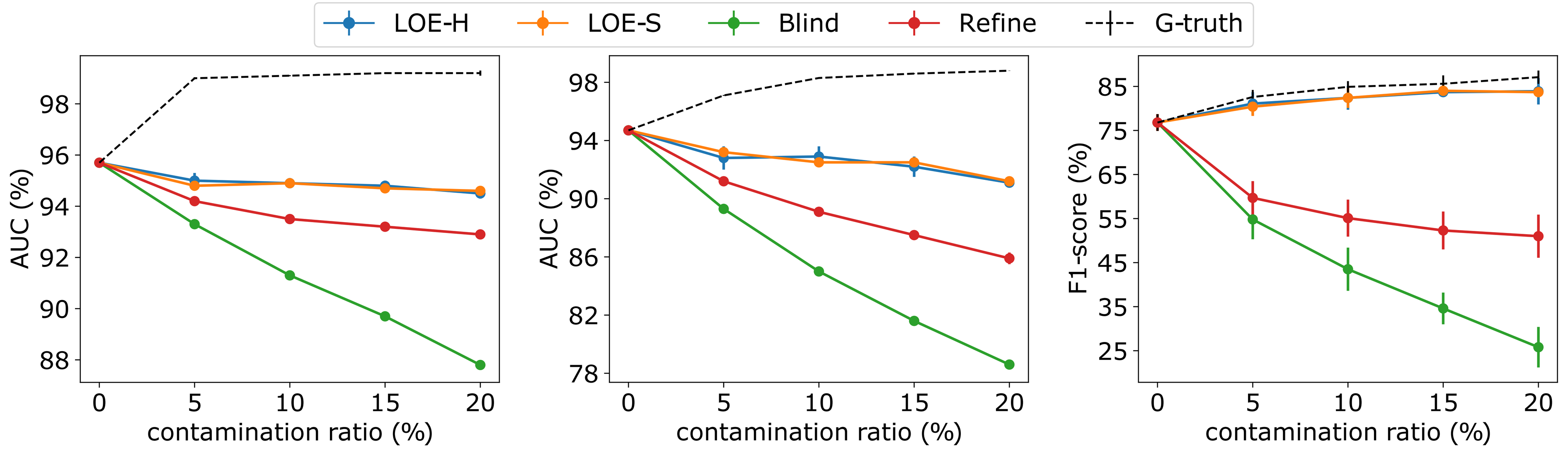
		CIFAR-10	F-MNIST
NTL	Blind	91.3±0.1 (-4.4)	85.0±0.2 (-9.7)
	Refine	93.5±0.1 (-2.2)	89.1±0.2 (-5.6)
	LOE_H (ours)	**94.9±0.2 (-0.8)**	**92.9±0.7 (-1.8)**
	LOE_S (ours)	**94.9±0.1 (-0.8)**	92.5±0.1 (-2.2)
MHRot	Blind	84.0±0.5 (-4.2)	88.8±0.1 (-4.9)
	Refine	84.4±0.1 (-3.8)	89.6±0.2 (-4.1)
	LOE_H (ours)	**86.4±0.5 (-1.8)**	**91.4±0.2 (-2.3)**
	LOE_S (ours)	86.3±0.2 (-1.9)	91.2±0.4 (-2.5)

Table. MVTec benchmark

	Detection		Segmentation	
	10%	20%	10%	20%
Blind	94.2±0.5 (-3.2)	89.4±0.3 (-8.0)	96.17±0.08 (-0.78)	95.09±0.17 (-1.86)
Refine	95.3±0.5 (-2.1)	93.2±0.3 (-4.2)	96.55±0.04 (-0.40)	96.09±0.06 (-0.86)
LOE_H (ours)	**95.9±0.9** (-1.5)	92.9±0.4 (-4.5)	95.97±0.22 (-0.98)	93.29±0.21 (-3.66)
LOE_S (ours)	95.4±0.5 (-2.0)	**93.6±0.3** (-3.8)	**96.56±0.04** (-0.39)	**96.11±0.05** (-0.84)

Table. F1-score on 30 tabular datasets ($\alpha = \alpha_0 = 10\%$)

	NTL				ICL			
	Blind	Refine	LOE_H (ours)	LOE_S (ours)	Blind	Refine	LOE_H (ours)	LOE_S (ours)
abalone	37.9±13.4	55.2±15.9	42.8±26.9	**59.3±12.0**	50.9±1.5	**54.3±2.9**	53.4±5.2	51.7±2.4
annthyroid	29.7±3.5	42.7±7.1	47.7±11.4	**50.3±4.5**	29.1±2.2	38.5±2.1	**48.7±7.6**	43.0±8.8
arrhythmia	57.6±2.5	59.1±2.1	62.1±2.8	**62.7±3.3**	53.9±0.7	60.9±2.2	62.4±1.8	**63.6±2.1**
breastw	84.0±1.8	93.1±0.9	**95.6±0.4**	95.3±0.4	92.6±1.1	93.4±1.0	**96.0±0.6**	95.7±0.6
cardio	21.8±4.9	45.2±7.9	**73.0±7.9**	57.8±5.5	50.2±4.5	56.2±3.4	**71.1±3.2**	62.2±2.7
ecoli	0.0±0.0	88.9±14.1	**100±0.0**	**100±0.0**	17.8±15.1	46.7±25.7	**75.6±4.4**	**75.6±4.4**
forest cover	20.4±4.0	56.2±4.9	61.1±34.9	**67.6±30.6**	9.2±4.5	8.0±3.6	6.8±3.6	**11.1±2.1**
glass	11.1±7.0	15.6±5.4	17.8±5.4	**20.0±8.3**	8.9±4.4	**11.1±0.0**	**11.1±7.0**	8.9±8.3
ionosphere	89.0±1.5	91.0±2.0	91.0±1.7	**91.3±2.2**	86.5±1.1	85.9±2.3	85.7±2.8	**88.6±0.6**
kdd	95.9±0.0	96.0±1.1	98.1±0.4	**98.4±0.1**	99.3±0.1	99.4±0.1	**99.5±0.0**	99.4±0.0
kddrev	98.4±0.1	98.4±0.2	89.1±1.7	**98.6±0.0**	97.9±0.5	98.4±0.4	**98.8±0.1**	98.2±0.4
letter	36.4±3.6	44.4±3.1	25.4±10.0	**45.6±10.6**	43.0±2.5	51.2±3.7	**54.4±5.6**	47.2±4.9
lympho	53.3±12.5	60.0±8.2	60.0±13.3	**73.3±22.6**	43.3±8.2	60.0±8.2	80.0±12.5	**83.3±10.5**
mammogra.	5.5±2.8	2.6±1.7	3.3±1.6	**13.5±3.8**	8.8±1.9	11.4±1.9	34.0±20.2	**42.8±17.6**
mnist tabular	78.6±0.5	**80.3±1.1**	71.8±1.8	76.3±2.1	72.1±1.0	80.7±0.7	**86.0±0.4**	79.2±0.9
mulcross	45.5±9.6	**58.2±3.5**	**58.2±6.2**	50.1±8.9	70.4±13.4	94.4±6.3	**100±0.0**	99.9±0.1
musk	21.0±3.3	98.8±0.4	**100±0.0**	**100±0.0**	6.2±3.0	**100±0.0**	**100±0.0**	**100±0.0**
optdigits	0.2±0.3	1.5±0.3	41.7±45.9	**59.1±48.2**	0.8±0.5	**1.3±1.1**	1.2±1.0	0.9±0.5
pendigits	5.0±2.5	32.6±10.0	79.4±4.7	**81.9±4.3**	10.3±4.6	30.1±8.5	80.3±6.1	**88.6±2.2**
pima	60.3±2.6	61.0±1.9	**61.3±2.4**	61.0±0.9	58.1±2.9	59.3±1.4	**63.0±1.0**	60.1±1.4
satellite	73.6±0.4	74.1±0.3	**74.8±0.4**	74.7±0.1	72.7±1.3	72.7±0.6	**73.6±0.2**	73.2±0.6
satimage	26.8±1.5	86.8±4.0	90.7±1.1	**91.0±0.7**	7.3±0.6	85.1±1.4	91.3±1.1	**91.5±0.9**
seismic	11.9±1.8	11.5±1.0	**18.1±0.7**	17.1±0.6	14.9±1.4	17.3±2.1	23.6±2.8	**24.2±1.4**
shuttle	97.0±0.3	97.0±0.2	**97.1±0.2**	97.0±0.2	96.6±0.2	96.7±0.1	96.9±0.1	**97.0±0.2**
speech	6.9±1.2	8.2±2.1	43.3±5.6	**50.8±2.5**	0.3±0.7	1.6±1.0	**2.0±0.7**	0.7±0.8
thyroid	43.4±5.5	55.1±4.2	**82.4±2.7**	**82.4±2.3**	45.8±7.3	71.6±2.4	**83.2±2.9**	80.9±2.5
vertebral	22.0±4.5	21.3±4.5	22.7±11.0	**25.3±4.0**	8.9±3.1	8.9±4.2	7.8±4.2	**10.0±2.7**
vowels	36.0±1.8	50.4±8.8	**62.8±9.5**	48.4±6.6	42.1±9.0	60.4±7.9	**81.6±2.9**	74.4±8.0
wbc	25.7±12.3	45.7±15.5	**76.2±6.0**	69.5±3.8	50.5±5.7	50.5±2.3	**61.0±4.7**	**61.0±1.9**
wine	24.0±18.5	66.0±12.0	90.0±0.0	**92.0±4.0**	4.0±4.9	10.0±8.9	98.0±4.0	**100±0.0**

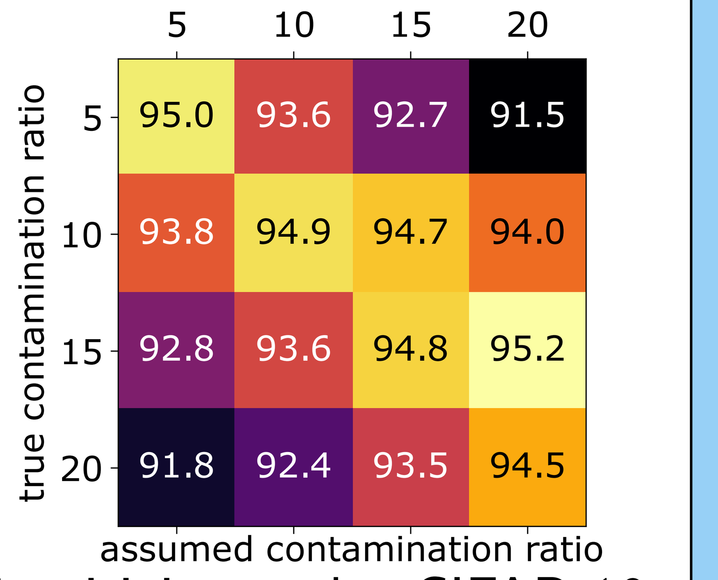


CIFAR-10 | FMNIST | Thyroid

Table. UCSD Peds1 video benchmark

Method	Contamination Ratio		
	10%	20%	30%*
(Tudor Ionescu et al., 2017)	-	-	68.4
(Liu et al., 2018)	-	-	69.0
(Del Giorno et al., 2016)	-	-	59.6
(Sugiyama & Borgwardt, 2013)	55.0	56.0	56.3
(Pang et al., 2020)	68.0	70.0	**71.7**
Blind	85.2±1.0	76.0±2.7	66.6±2.6
Refine	82.7±1.5	74.9±2.4	69.3±0.7
LOE_H (ours)	82.3±1.6	59.6±3.8	56.8±9.5
LOE_S (ours)	**86.8±1.2**	**79.2±1.3**	**71.5±2.4**

*Default setup in (Pang et al., 2020), corresponding to $\alpha_0 \approx 30\%$.



Sensitivity study: CIFAR-10