



Reliably Detecting Model Failures in Deployment Without Labels


Viet Nguyen^{1,3,†}, Changjian Shui^{2,†}, Vijay Giri⁴,
Siddharth Arya^{1,3}, Amol Verma^{1,5}, Fahad Razak^{1,5}, Rahul G. Krishnan^{1,3}

¹University of Toronto ²University of Ottawa ³Vector Institute

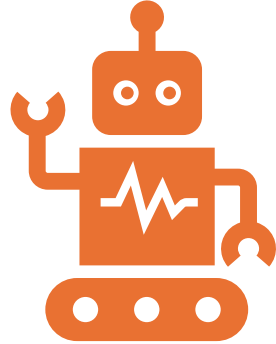
⁴University of Pennsylvania ⁵Unity Health Toronto



Motivation

- Current ML models assume that training and deployment data are independent and identically distributed (IID).
 - In practice: this assumption fails due to **distribution shifts** during model deployment.
 - e.g. temporal shifts due to long-term trends, seasonality, source-target mismatch...
 - We call this scenario **post-deployment deterioration (PDD)** and monitor its occurrence without relying on deployment labels.
- 

This work



We propose **Disagreement-Driven Deterioration Monitoring (D3M)**, a monitoring protocol leveraging model disagreement.



D3M does not require labels for deployment queries.

Maximum disagreement rate

- Train f with ERM on ID **labeled** data D^n .
- For an **unlabeled** query D^m , the maximum disagreement rate ϕ for f is the highest achievable disagreement rate on D^m between f and models that agree with f on D^n .

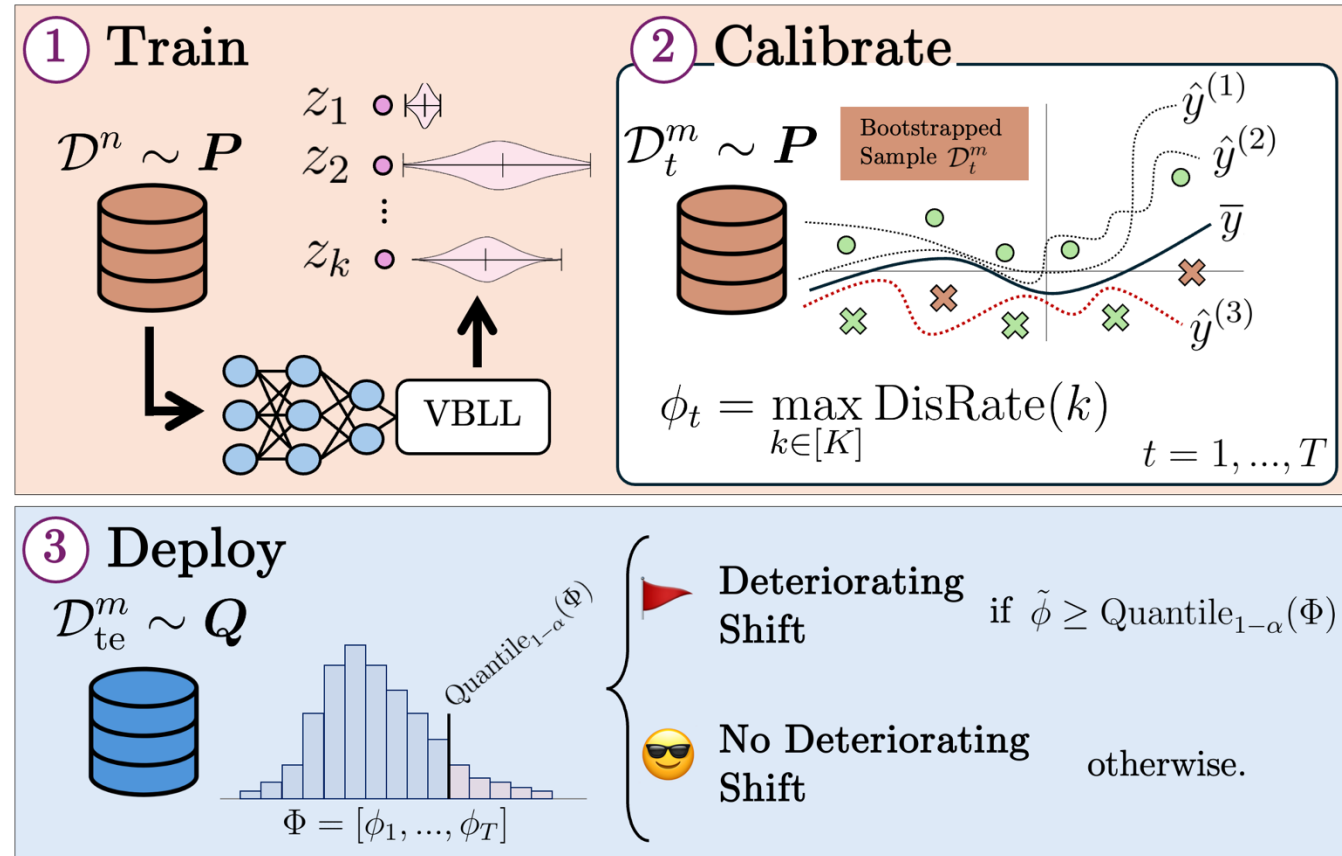
Competing hypotheses disagree better on OOD data where f underperforms than ID data.

Core mechanism

1. Collect ID maximum disagreement rates on bootstrapped validation sets into Φ .
2. For some query, compute its maximum disagreement rate ϕ .
3. If ϕ could not have come from Φ , PDD will occur!

Implementation

- f is the mean model of a (pretrained) feature extractor with a Variational Bayesian Last Layers (VBLL) output.
- Maximum disagreement rate computation:
 - Sample K hypotheses, compute their disagreement rates with respect to f .
 - Record the maximum value found!



Results:

UCI,
CIFAR-10.1,
Camelyon17

	UCI Heart Disease			CIFAR 10.1			Camelyon 17		
	10	20	50	10	20	50	10	20	50
BBSD	.13±.03	.22±.04	.46±.05	.07±.03	.05±.02	.12±.03	.16±.04	.38±.05	.87±.03
Rel. Mahalanobis	.11±.03	.36±.05	.66±.05	.05±.02	.03±.03	.04±.02	.16±.04	.40±.05	.89±.03
Deep Ensemble	.13±.03	.32±.05	.64±.05	.33±.05	.52±.05	.68±.05	.14±.03	.26±.04	.82±.04
CTST	.15±.04	.51±.05	.98±.01	.03±.02	.04±.02	.04±.02	.11±.03	.59±.05	.59±.05
MMD-D	.09±.03	.12±.03	.27±.04	.24±.04	.10±.03	.05±.02	.42±.05	.62±.05	.69±.05
H-Div	.15±.04	.26±.04	.37±.05	.02±.01	.05±.02	.04±.02	.03±.02	.07±.03	.23±.04
Detectron	.24±.04	.57±.05	.82±.04	.37±.05	.54±.05	.83±.04	.97±.02	1.0±.00	.96±.02
D3M (Ours)	.38±.19	.25±.28	.69±.33	.40±.10	.45±.10	.74±.12	.89±.20	.93±.05	.99±.02

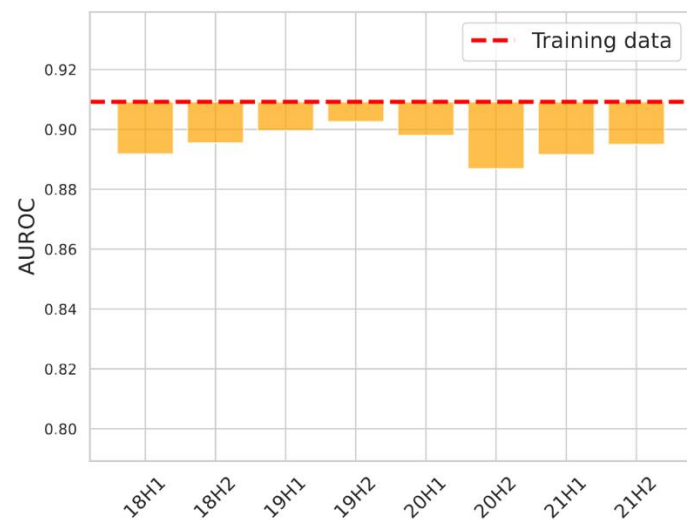
Table 2: True positive rates (TPR) comparison across datasets and query sizes. As models do experience deterioration, the higher TPR the better. **Bold** indicates best in column. We report the means and standard deviations of TPRs obtained from 10 independently seeded runs.

	UCI Heart Disease		CIFAR 10.1		Camelyon 17	
	100	200	100	200	100	200
D3M (Ours)	.93±.10	.99±.01	.91±.11	.99±.01	1.0±.00	1.0±.00

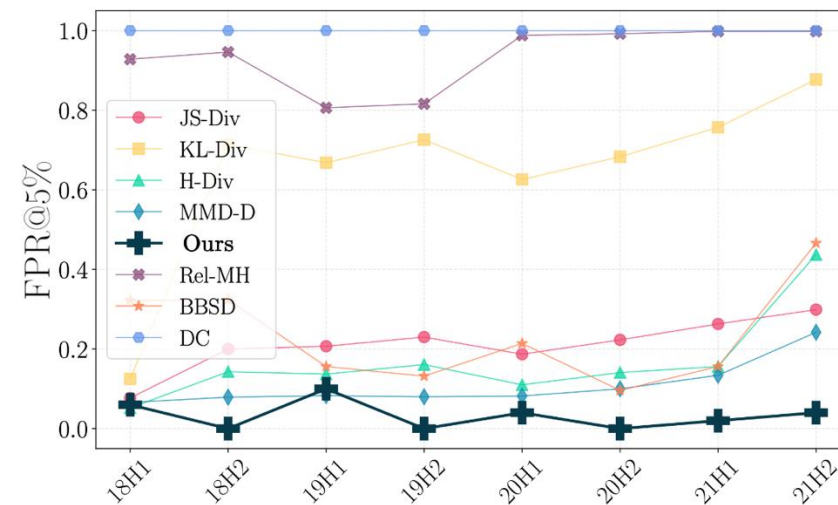
Table 7: True positive rates (TPR) for D3M across datasets and test sizes 100, 200.

Results:

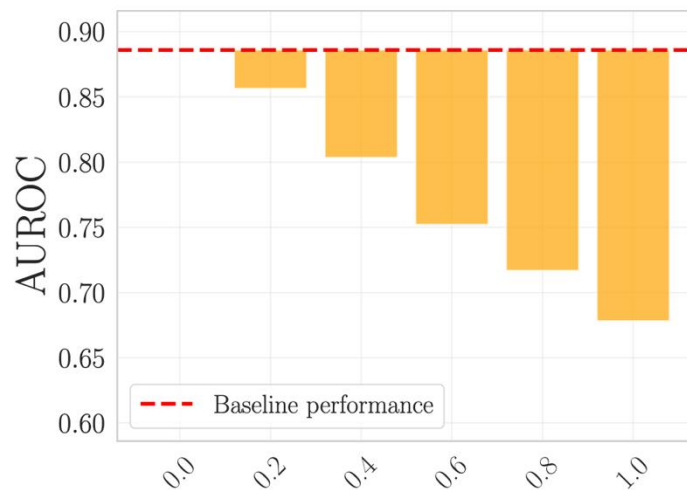
GEMINI



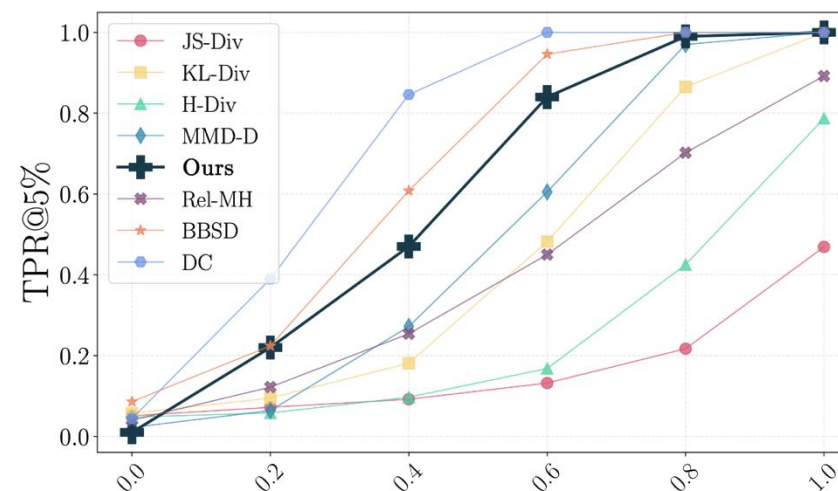
(a) Non deteriorating shift in GEMINI



(b) Performance monitoring



(a) Deteriorating shifts in GEMINI



(b) Deterioration monitoring

Thank you for your
attention!

Please visit us at our poster session!



Viet Nguyen[†]
University of Toronto



Changjian Shui[†]
University of Ottawa



Vijay Giri
University of
Pennsylvania



Siddharth Arya
University of Toronto



Amol Verma
Unity Health Toronto



Fahad Razak
Unity Health Toronto



Rahul G. Krishnan
University of Toronto