ICML Workshop
Artificial
Intelligence &
Human-Computer
Interaction

Human-in-the-Loop Out-of-Distribution Detection with False Positive Rate Control

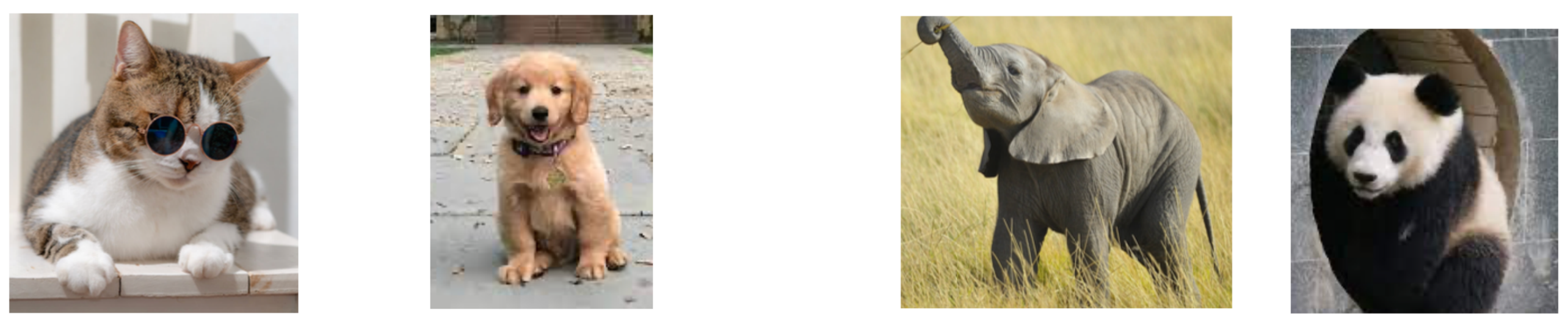
Harit Vishwakarma*, Heguang Lin* and Ramya Korlakai Vinayak

University of Wisconsin-Madison, WI, USA

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

ML models need to be robust to OOD samples after deployment.

ID: positive, OOD: Negative



In Distribution (ID) Out of Distribution (OOD)

ID: Normal and Alzheimer brain scans OOD : Brain scans for other diseases

OOD data is not known during training and hard to foresee.

Existing methods

Design scoring functions g to quantify OOD uncertainty.

Find OOD classification threshold λ to maximize TPR.

Often lead to high False Positive Rate (FPR).

Cannot adapt to shifts in the OOD data after deployment.

Our method

Adapts threshold λ as it receives OOD samples.

Ensures FPR remains below a user-given threshold α.

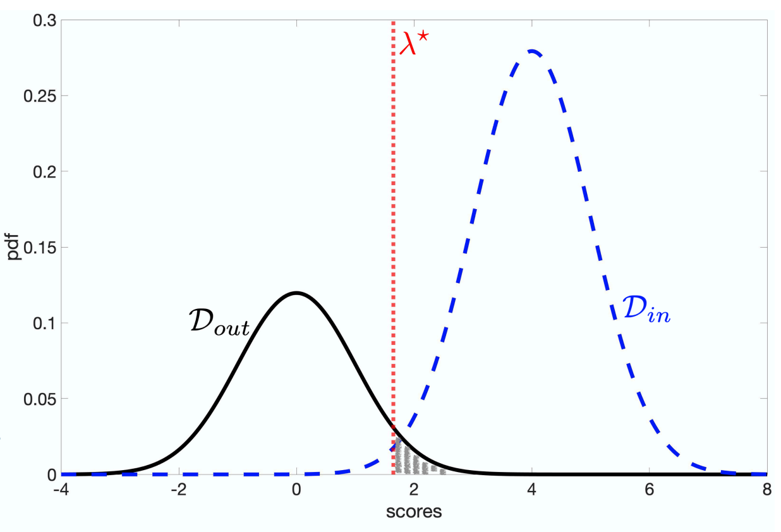
Can adapt to shifts in OOD data.

Setting

ID and OOD samples come from some mixed distribution, $\mathscr{D} = (1-\gamma)\mathscr{D}_{id} + \gamma\mathscr{D}_{ood}$ that could potentially change over time.

A pre-trained model on ID data is given and samples come one at a time, i.e. post deployment, we are in the online setting.

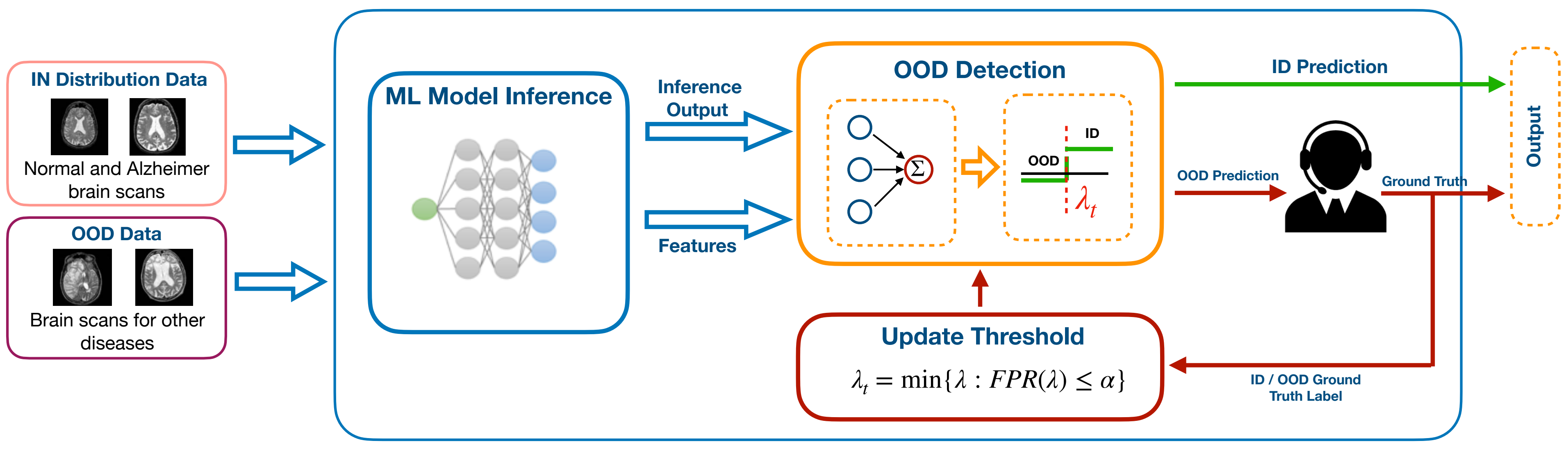
We use existing methods (e.g. Energy score) that provide a score to quantify the uncertainty of a given sample being OOD.



$\mathrm{FPR}(\lambda^\star) = \alpha \longleftrightarrow \mathrm{CDF}_{\mathcal{D}_{out}}(\lambda^\star) = 1-\alpha$

Method

1. Get sample x_t , compute the score $s_t = g(x_t)$.
2. If $s_t \geq \hat{\lambda}_{t-1}$ predict ID else predict OOD.
3. If $s_t \leq \hat{\lambda}_{t-1}$ predict OOD and query true label from Oracle w.p. p .
4. Update the threshold by solving the optimization,
   $\hat{\lambda}_t = \underset{\lambda\in\Lambda}{\arg\min}\ \lambda \text{ s.t. } \widehat{FPR}(\lambda, t) + \psi(t,\delta) \leq \alpha$
   $\psi(t,\delta)$ is our confidence interval on FPR estimate that is valid for all $\lambda \in \Lambda$ and at all time $t \geq 1$.

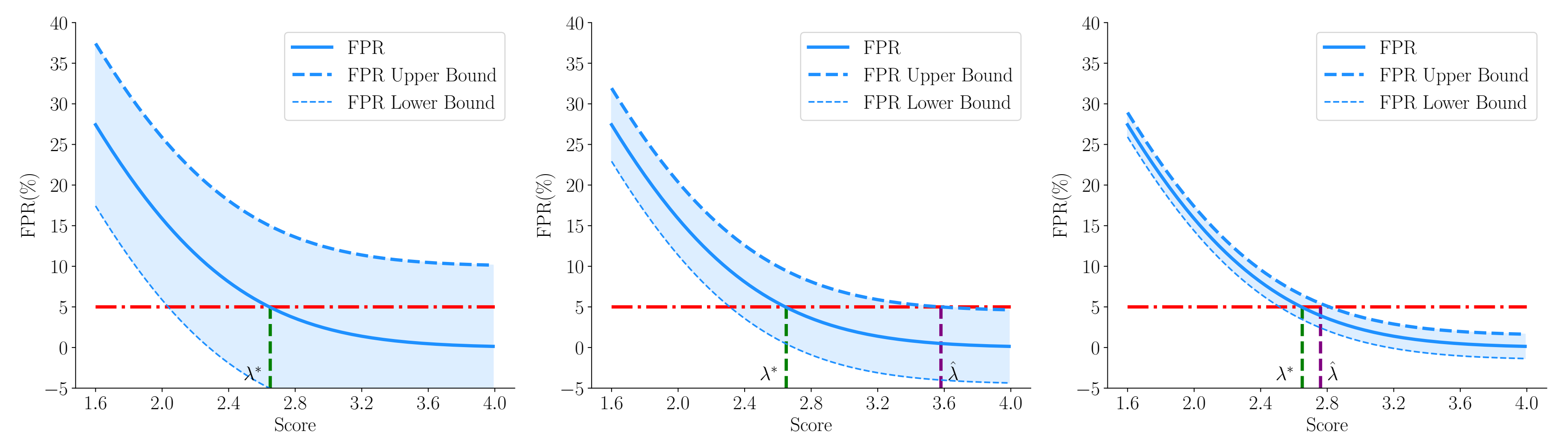


Theoretical Results

1. Anytime valid confidence interval $\psi(t,\delta)$ on FPR estimate at any $\lambda \in \Lambda$. $N_t^{(o)}$ is the count of OOD samples at time t and $|\Lambda| < \infty$

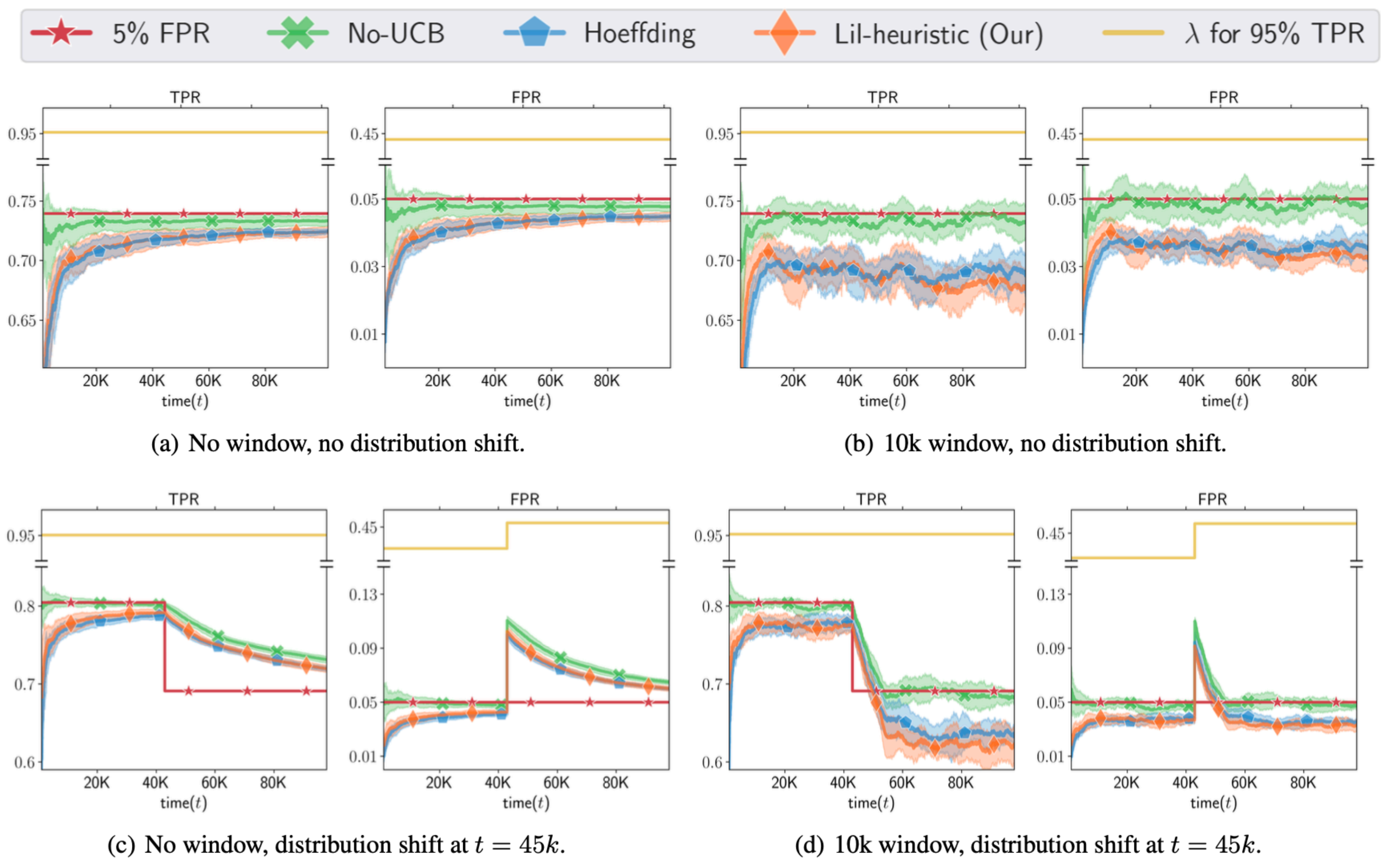
$$\psi(t,\delta) = \sqrt{\frac{3c_t}{N_t^{(o)}}\left[2\log\log\left(\frac{3c_t N_t^{(o)}}{2}\right) + \log\left(\frac{2|\Lambda|}{\delta}\right)\right]},$$

2. $FPR(\hat{\lambda}_t) \leq \alpha$ for all $t \geq 1$.
3. Based on $\psi(t,\delta)$ we provide bounds on time points at which the system reaches the feasible and optimal regimes.



(a) No feasible solution, in the beginning (b) Feasible solution, after sometime (c) Near optimal solution, eventually

Experiments



(a) No window, no distribution shift. (b) 10k window, no distribution shift.

(c) No window, distribution shift at $t = 45k$. (d) 10k window, distribution shift at $t = 45k$.

Experiments on the KNN method with cifar10 as the ID dataset.

We run experiments on synthetic and real datasets to validate our claims.

Across all the settings we find that our method always respects the FPR constraint while optimizing for TPR.

On the other hand not using any confidence interval on FPR estimate results in violations of FPR constraint.

Further, we also observe that with appropriate window size our method adapts well when the distributions shift.

References

(1) Balsubramani, A. *Sharp finite-time iterated logarithm martingale concentration, arXiv 1405.2639,* 2015.
(2) Howard, S. R. and Ramdas, A. *Sequential estimation of quantiles with applications to A/B testing and best-arm identification.* Bernoulli, 28(3):1704 – 1728, 2022.
(3) Yang, J., et al. *Openood: Benchmarking generalized out-of-distribution detection*, 2022.

* Equal Contributions
Correspondence: hvishwakarma@cs.wisc.edu

AI & HCI Workshop at the 40th International Conference on Machine Learning (ICML),2023, Honolulu, Hawaii, USA.