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

Harit Vishwakarma HVISHWAKARMA@CS.WISC.EDU
University of Wisconsin-Madison

Heguang Lin HGLIN@SEAS.UPENN.EDU
University of Pennsylvania

Ramya Korlakai Vinayak RAMYA@ECE.WISC.EDU
University of Wisconsin-Madison

Abstract

Robustness to Out-of-Distribution (OOD) samples is crucial for the successful deployment of machine learning models in the open world. Since it is not possible to have a priori access to a variety of OOD data before deployment, several recent works have focused on designing scoring functions to quantify OOD uncertainty. These methods often find a threshold that achieves 95% true positive rate (TPR) on the In-Distribution (ID) data used for training and use this threshold for detecting OOD samples. However, this can lead to very high FPR as seen in a comprehensive evaluation of the Open-OOD benchmark, the FPR can range between 60 to 96% on several ID and OOD dataset combinations. In contrast, practical systems deal with a variety of OOD samples on the fly and critical applications, e.g., medical diagnosis, demand guaranteed control of the false positive rate (FPR). To meet these challenges, we propose a mathematically grounded framework for human-in-the-loop OOD detection, wherein expert feedback is used to update the threshold. This allows the system to adapt to variations in the OOD data while adhering to the quality constraints. We propose an algorithm that uses any time-valid confidence intervals based on the Law of Iterated Logarithm (LIL). Our theoretical results show that the system meets FPR constraint while minimizing the human feedback for points that are in-distribution. Another key feature of the system is that it can work with any existing post-hoc OOD uncertainty-quantification methods. We evaluate our system empirically on a mixture of benchmark OOD datasets in image classification tasks on CIFAR-10 and CIFAR-100 as in distribution datasets and show that our method can maintain FPR at most 5% while maximizing TPR.

Keywords: Out-of-Distribution Detection, False Positive Rate Control, Law of Iterated Logarithms

1. Introduction

Deploying machine learning (ML) models in the open world makes them subject to out-of-distributions (OOD) inputs — in the classification setup OOD data points are those that do not belong to any of the classes in the training data. The modern ML models, in particular deep neural networks, can fail silently with high confidence on OOD points Nguyen et al. (2015); Amodei et al. (2016) rather than flagging them as OOD and asking for human intervention as they are not designed to do so. Such failures can have serious consequences in high-risk applications e.g. medical diagnosis, autonomous driving, etc. For a successful deployment of an ML model in the open world, we need mechanisms that ensure robustness to the OOD inputs.

The importance of this problem has prompted the development of several solutions. Many of these works have addressed this problem by proposing post-hoc methods for OOD detection Liang

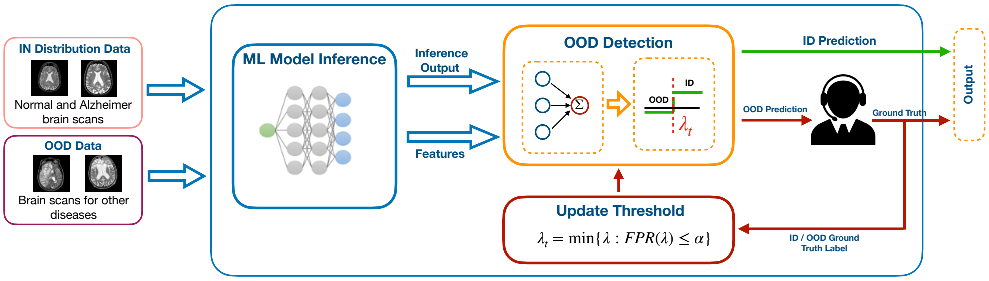


Figure 1: Illustration of OOD detection with human-in-the-loop with FPR control. In this example, the ID data is of brain scans of normal people and those with Alzheimer's disease. The OOD data could be anything other than these, e.g. brain scans of patients with some other diseases.

et al. (2017); Lee et al. (2018); Liu et al. (2020); Ming et al. (2022). Broadly, these works offer methods to quantify a *score* that can be used to decide OOD vs ID label for a given point. Many of these methods are based on the distance between data points or a model's confidence score in prediction, for a detailed survey of literature in the area of generalized OOD detection, see Yang et al. (2021b). However, many of these works are largely limited to static settings where the ID data, which is available in a plenty for training and validating the ML system is used to set a threshold on the scores used for OOD detection Liang et al. (2017); Liu et al. (2020); Ming et al. (2022). This is usually done by setting a threshold that achieves a certain level of true positive rate (TPR), e.g., TPR of 95%. However, this can lead to a very high false positive rate (FPR,) e.g., ranging between 60% to 90% on several benchmarked ID and OOD dataset combinations Yang et al. (2022). Furthermore, even if the ID data distribution remains the same after deployment, the OOD data could vary, resulting in highly fluctuating FPR. Thus, having a small and fixed amount of OOD data collected a priori to validate the FPR at a given threshold would not help in guaranteeing FPR.

In critical applications, the consequences of classifying an OOD point as ID (false positive) could be more catastrophic than classifying an ID point as OOD (false negative), e.g. in medical diagnosis, when in doubt it is better to classify a brain scan as OOD and defer the decision to human experts rather than for the ML model to give it a disease label or classify as normal assuming it to be ID. Therefore, it is crucial to guarantee that the false positive rate (FPR) is below a certain acceptable rate, e.g., FPR below 5%. Since the availability of exact type of OOD data that the system can encounter during deployment is rare, it is crucial to design systems that can adapt to the OOD data while controlling the FPR during deployment.

Goal: Develop human-in-the-loop out-of-distribution detection system that has guaranteed false positive rate control while minimizing the amount of human intervention needed.

In this work we present a framework for human-in-the-loop Out-of-Distribution (OOD) detection, ensuring strict control over the false positive rate (FPR) while adapting to diverse OOD data.

Our Contributions: Toward this goal, we make the following contributions:

1. **Human-in-the-loop OOD detection framework**: We propose a novel mathematically grounded framework that incorporates expert human feedback to adaptively update the OOD detection threshold, ensuring robustness to variations in OOD data encountered after deployment. Our framework can be used with any scoring function.

2. **Guaranteed FPR control**: Our approach leverages mathematically grounded confidence intervals based on the Law of Iterated Logarithm to meet false positive rate (FPR) constraints while minimizing the need for human feedback on in-distribution points. For stationary settings, we provide theoretical guarantees for our proposed framework and algorithm on controlling FPR at the desired level at all times and also provide a bound on the time taken to reach a given level of optimality. Using the insight from this analysis, we also propose an approach for settings with change points that reduces the duration of violation of FPR control before adapting to the change.
3. **Empirical validation on benchmark datasets**: We evaluate our framework through extensive simulations both in stationary and distribution shift settings. Through experiments on benchmark OOD datasets in image classification tasks, we demonstrate the practical effectiveness of our proposed approaches.

The paper is structured as follows: Section 2 presents the framework in detail, while Section 5 provides theoretical guarantees on False Positive Rate (FPR) control. In Section 7, we conduct a comprehensive empirical evaluation of the proposed system. The proof details and extensive experimental results are in Appendix.

2. Human-in-the-Loop OOD Detection

In this section, we discuss our proposed system in detail. Recall that we are motivated by two facts. First, the type of OOD samples the system will encounter after deployment are often not available during development, hence we need to build OOD detection systems that can adapt to various kinds of OOD data that it encounters on-the-fly after deployment. Second, in many critical applications, the cost of false positives i.e. misclassifying an OOD point as ID can have more severe consequences than misclassifying an ID point as OOD e.g., in medical diagnosis of brain scans, when in doubt it is preferable to classify a scan as OOD and seek the input of a radiologist, rather than labeling it with a disease or as normal using the machine-learned classifier.

We propose a human-in-the-loop OOD detection system (Figure 1) that can work with any ML inference model and scoring function for OOD detection. We begin by describing the problem setting and then discuss each component of our proposed system in detail. See algorithm 1 for step-by-step details.

Algorithm 1 Human in the Loop OOD Detection

```
Input: FPR threshold $\alpha$ , window size $N_w$,
sampling probability $p \in (0, 1)$ Scoring function $g : \mathcal{X} \mapsto \mathbb{R}$,
$S_0 = \Phi$, $\hat{\lambda}_0 = \infty$
for $t = 1, 2, \ldots$ do
    Receive data point $x_t$ ; $s_t = g(x_t)$
    if $s_t \leq \hat{\lambda}_{t-1}$ then $l_t = 1$ else $l_t \sim \texttt{Bernoulli}(p)$
    if $l_t = 1$ then
        $y_t = \texttt{GetExpertLabel}(x_t)$
        $S_t = S_{t-1} \cup \{(s_t, y_t)\}$
    $\hat{\lambda}_t = \texttt{SolveForLambda}(S_t, N_w, \alpha)$
    $\hat{y}_t = \text{sign}(s_t - \hat{\lambda}_t)$
    if $l_t = 1$ then Output $y_t$ else Output $\hat{y}_t$
end for
```

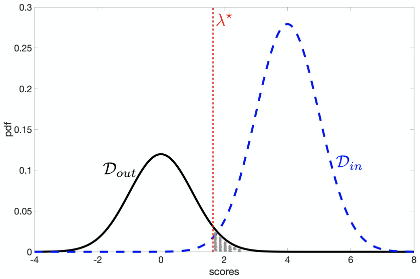
Data stream: Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the feature space of the data points. The OOD detection system is expected to classify an incoming data point as either "1" i.e. ID (In Distribution) or "-1" stands i.e. OOD (Out Of Distribution). In short, the label space is $\mathcal{Y} = \{-1, 1\}$. Let the distribution of ID and OOD data be denoted by $\mathcal{D}_{id}$ and $\mathcal{D}_{ood}$ respectively. Let $x_t \in \mathcal{X}$ denote the sample received at the time t. Let $y_t \in \{-1, 1\}$ denote the true label for x_t with respect to ID or OOD classification. We assume x_t are independent and drawn according to the following mixture model, $x_t \sim (1 - \gamma)\mathcal{D}_{id} + \gamma\mathcal{D}_{ood}$, where $\gamma \in (0, 1)$ is the fraction of OOD in the mixed stream. Note that $\mathcal{D}_{id}$, $\mathcal{D}_{ood}$ and γ are *unknown*.

Scoring function for OOD detection: After receiving data point x_t, the system uses a given scoring function, $g : \mathcal{X} \mapsto \mathcal{S} \subseteq \mathbb{R}$, to compute a real-valued score quantifying the uncertainty of the point being ID or OOD. There are several scoring functions that have been developed recently e.g. energy-based scores Liu et al. (2020), Mahalanobis distance-based scores Lee et al. (2018) etc. Some of these scoring functions are quite accurate in providing scores to classify points as OOD or ID and designing more accurate functions is an active area of research. Our system can use any of these post-hoc OOD uncertainty quantification functions. We emphasize that our aim is not to design a new OOD uncertainty quantification method (scoring function), instead, we propose a system in which any such g can be plugged in and it can control the FPR.

Denote the score computed for point x_t as $s_t = g(x_t)$. To be consistent across various scoring functions, let a higher score indicate ID and a lower score indicate OOD points. After computing the uncertainty score s_t the system needs to decide whether x_t is OOD or ID, which is done using a threshold-based classifier parameterized with $\lambda \in \Lambda \subseteq \mathbb{R}$: $h_\lambda(g(x)) = \text{sign}(g(x) - \lambda)$. Here we assume $\Lambda = (\Lambda_{\min}, \Lambda_{\max})$ is a subset of $\mathbb{R}$. The threshold-based prediction is common in the OOD detection literature Liu et al. (2020); Lee et al. (2018). Since OOD data is usually not available during development, a common practice is to find a threshold $\hat{\lambda}$ that correctly classifies at least 95% of the ID data used for training/validation of the ML system as ID, i.e., $\hat{\lambda}$ is chosen for achieving 95% TPR. While simple, a drawback of this approach is that it can result in an exceedingly high FPR, as demonstrated by a thorough examination conducted in the Open-OOD benchmark, where the FPR can range from 60% to 96% on various combinations of ID and OOD datasets. In contrast, real-world systems must handle a diverse range of OOD samples in real time, and for critical applications such as medical diagnosis, it is imperative to ensure control over the FPR. The population level FPR and TPR for any $\lambda \in \Lambda$ are defined as follows,

$$\text{FPR}(\lambda) = \mathbb{E}_{x \sim \mathcal{D}_{ood}}[\mathbf{1}\{g(x) > \lambda\}] \quad \text{and} \quad \text{TPR}(\lambda) = \mathbb{E}_{x \sim \mathcal{D}_{id}}[\mathbf{1}\{g(x) > \lambda\}]. \tag{1}$$

Note that the cumulative distribution function (CDF) of $\mathcal{D}_{ood}$, $\text{CDF}_{\mathcal{D}_{ood}}(\lambda) = \mathbb{E}_{x \sim \mathcal{D}_{ood}}[\mathbf{1}\{g(x) \leq \lambda\}]$. Therefore, $\text{FPR}(\lambda) = 1 - \text{CDF}_{\mathcal{D}_{ood}}(\lambda)$. Similarly, $\text{TPR}(\lambda) = 1 - \text{CDF}_{\mathcal{D}_{id}}(\lambda)$. Since the CDF of any distribution is a monotonic function, both the FPR and TPR are monotonic in λ.



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

Figure 2: Optimal $\lambda^\star$ for the optimization problem (P1) with $\alpha = 0.05$ and $x_t \overset{i.i.d}{\sim} 0.7\ \mathcal{D}_{in} + 0.3\ \mathcal{D}_{out}$, where the scores of $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ are distributed as $\mathcal{N}(4, 1)$ and $\mathcal{N}(0, 1)$ respectively.

Expert feedback and importance sampling: In our proposed system, we choose λ adaptively using human feedback so that the FPR is maintained below the user-specified rate of α. One can of course achieve this trivially by setting $\lambda_t = \Lambda_{\max}$, i.e., always getting human feedback. This would of course be extremely expensive and defeat the purpose of having a machine learned classification model. Therefore, in addition to controlling the FPR, we want to minimize the human feedback solicited by the system. This is equivalent to maximizing the true positive rate. That is, $\lambda_t := \arg\ \max_\lambda \text{TPR}(\lambda)$ subject to $\text{FPR}(\lambda) \leq \alpha$. Since the TPR is monotonic in λ, this can be re-written as,

$$\boxed{\lambda_t^\star := \underset{\lambda\in\Lambda}{\arg\text{minimize}}\ \lambda, \quad \text{subject to} \quad \text{FPR}(\lambda) \le \alpha. \qquad \text{(P1)}}$$

Optimal threshold, denoted by $\lambda^\star$, is the smallest λ such that $\text{FPR}(\lambda^\star) = \alpha = 1 - \text{CDF}_{\mathcal{D}_{out}}(\lambda^\star)$ (see Figure 2).

When the distribution of the OOD points, $\mathcal{D}_{ood}$, is not changing, if we had access to the true FPR, then $\lambda_t^\star = \lambda^\star$. Note that, γ, the mixture ratio, or the distribution of the ID points $\mathcal{D}_{id}$ changing does not affect the value of the optimal threshold. As we do not have access to the true FPR and TPR values, we cannot solve the optimization problem (P1). Instead, we have to estimate the threshold at time t, $\lambda_t^\star$, using the observations made up to time t. Thus, at each time point our goal is to find $\hat{\lambda}_t \in \Lambda$ such that the FPR when using $\hat{\lambda}_t$ as the threshold in h_λ, denoted by $\text{FPR}(\hat{\lambda}_t)$, is at most α.

Ideally, we want to avoid human feedback for points with a score greater than $\hat{\lambda}_t$, i.e., those points that are determined as ID by the system. However, in order to have an unbiased estimate of the FPR and also to allow for potential change in the distribution of OOD samples and therefore change in true FPR, we allow for human feedback with a small probability p for points predicted as in-distribution by the system to be able to detect a change.

Algorithm 2 SolveOptForLambda

```
Input: FPR threshold $\alpha$ , $S_t$
  $\hat{\lambda}_t := \arg\min_{\lambda\in\Lambda} \lambda$  s.t.  $\widehat{\text{FPR}}(\lambda, t) + \psi(t, \delta) \le \alpha$
  Output $\hat{\lambda}_t$
```

FPR estimation and adapting the threshold: At each time t, we observe $x_t \overset{i.i.d}{\sim} (1-\gamma)\mathcal{D}_{id} + \gamma\mathcal{D}_{ood}$, and $s_t = g(x_t)$ is the corresponding score. If $s_t \le \hat{\lambda}_{t-1}$, then it is considered an OOD point and hence gets a human label for it and we get to know whether it is in fact OOD or ID. If $s_t > \hat{\lambda}_{t-1}$, then it is considered an ID point and hence gets a human label only with probability p. So, we get to know whether it is truly ID or not with probability p. Now we have to update the threshold, $\hat{\lambda}_t$, such that the $\text{FPR}(\hat{\lambda}_t) \le \alpha$ for all t, while trying to maximize $\text{TPR}(\hat{\lambda}_t)$. Our approach is based on constructing an unbiased estimator of $\text{FPR}(\lambda)$ using the OOD samples received till time t and in conjunction with confidence intervals for $\text{FPR}(\lambda)$ at all thresholds $\lambda \in \Lambda$ that is valid for all times simultaneously. Together, at each time t, these give us a reliable upper bound on the true $\text{FPR}(\lambda)$ for all λ enabling us to find the smallest λ such that the upper bound on $\text{FPR}(\lambda)$ is at most α. Let $S_t^{(o)} = \{s_1^{(o)}, \dots s_{N_t^{(o)}}^{(o)}\}$ denote the scores of the points that have been truly identified as OOD from human feedback and $I_t^{(o)}$ be the corresponding time points. We estimate the FPR as follows,

$$\widehat{\text{FPR}}(\lambda, t) = \frac{1}{N_t^{(o)}} \sum_{u \in I_t^{(o)}} Z_u(\lambda), \ \text{ where } Z_u(\lambda) := \begin{cases} \mathbf{1}(s_u^{(o)} > \lambda), & \text{if } s_u^{(o)} \le \hat{\lambda}_{u-1} \\ \frac{1}{p}\mathbf{1}(s_u^{(o)} > \lambda), & \text{w.p. } p \text{ if } s_u^{(o)} > \hat{\lambda}_{u-1} \\ 0, & \text{w.p. } 1-p \text{ if } s_u^{(o)} > \hat{\lambda}_{u-1} \end{cases}. \tag{2}$$

Finding threshold using a UCB on FPR: We use this estimated FPR with an upper confidence bound (UCB) to replace the unknown true FPR in the optimization problem (P1) to obtain the following optimization problem (P2),

$$\hat{\lambda}_t := \arg \underset{\lambda \in \Lambda}{\text{minimize}} \;\; \lambda \;\; \text{subject to} \;\; \widehat{\text{FPR}}(\lambda, t) + \psi(t, \delta) \leq \alpha, \tag{P2}$$

where the term $\psi(t, \delta)$ is a time-varying upper confidence which is simultaneously valid for all λ for all time with probability at least $1 - \delta$ for any given $\delta \in (0, 1)$. The minimization problem can be solved in many ways. We use a binary search procedure where we search over a grid on $[\Lambda_{\min}, \Lambda_{\max}]$ with grid-size ν. The procedure 2 searches for a smallest λ such that $\widehat{\text{FPR}}(\lambda, t) + \psi(t, \delta) \leq \alpha$. It uses eq. (4) to compute the empirical FPR at various thresholds and the confidence interval $\psi(t, \delta)$ given in eq. (3). Details of the binary search procedure are in the Appendix.

Upper confidence bound (UCB): At each time point, the algorithm estimates the FPR using a finite number of samples at all thresholds. We need confidence intervals that are valid for all thresholds at all time points to ensure the algorithm has reliable upper bounds on the FPR. In particular, we use the Law of iterated logarithm(LIL) Khinchine (1924) based confidence bounds that are known to be tight. In our setting, due to the importance sampling, the samples become conditionally dependent. This dependence prevents direct application of known results like Howard and Ramdas (2022). In section 5 we build upon the LIL bounds for martingales Balsubramani (2015) and derive a confidence interval bound that is valid in our setting (see equation (3)),

$$\psi(t, \delta) := \sqrt{\frac{3c_t}{N_t^{(o)}} \left[2 \log \log \left(\frac{3c_t N_t^{(o)}}{2} \right) + \log \left(\frac{2}{\delta} \left(\frac{\Lambda_{\max} - \Lambda_{\min}}{\nu} \right) \right) \right]}, \tag{3}$$

where $c_t = 1 - \beta_t + \frac{\beta_t}{p^2}$, $\beta_t = \frac{N_t^{(o,p)}}{N_t^{(o)}}$ and $N_t^{(o,p)}$ is the number of points sampled using importance sampling until time t and $\nu \in (0, 1)$ is a discretization parameter set by the user.

Handling distribution shift: One of the motivations for the system is to be able to adapt to the variations of the OOD data. As long as $\mathcal{D}_{ood}$ does not change, changes in the $\mathcal{D}_{id}$ or the mixing ratio γ do not affect the true FPR. However, the true FPR does get affected when $\mathcal{D}_{ood}$ changes. When there is a change in $\mathcal{D}_{ood}$, estimating the FPR using all the acquired samples so far can lead to inaccurate estimates as the current estimate is highly influenced by scores that are far behind in time from the previous D_{ood}. This can lead to inaccurate estimation of λ and erroneous predictions for the current time. To overcome this challenge, we propose a sliding window-based approach where the user can set a window size $N_w > 0$ and the system will only estimate the FPR and the confidence intervals using the most recent N_w OOD samples. This will allow the system to adapt the threshold that is well aligned with the new distribution(s) of OOD samples. Next, we provide theoretical guarantees for the proposed algorithm when $\mathcal{D}_{ood}$ does not change over time. We refer the reader to the appendix for theoretical analysis and experimental results.

3. Conclusion

We presented a mathematically grounded framework for human-in-the-loop Out-of-Distribution (OOD) detection. By incorporating expert feedback and utilizing confidence intervals based on the Law of Iterated Logarithm (LIL), our approach maintains control over false positive rates (FPR) while maximizing true positive rates (TPR). The empirical evaluations on synthetic data and image classification tasks demonstrate the effectiveness of our method in maintaining FPR at or below 5% while achieving high TPR. Our work gives a promising solution for addressing the challenge of robustness to OOD samples in real-world applications.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016.

Fabrizio Angiulli and Fabio Fassetti. Detecting distance-based outliers in streams of data. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 811–820, 2007. ISBN 9781595938039.

Akshay Balsubramani. Sharp finite-time iterated-logarithm martingale concentration, 2015.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826*, 2023.

Guilherme O. Campos, Arthur Zimek, Jörg Sander, Ricardo J. Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E. Houle. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Min. Knowl. Discov.*, 30(4):891–927, 2016. ISSN 1384-5810. doi: 10.1007/s10618-015-0444-8. URL `https://doi.org/10.1007/s10618-015-0444-8`.

Chengliang Chai, Lei Cao, Guoliang Li, Jian Li, Yuyu Luo, and Samuel Madden. Human-in-the-loop outlier detection. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 19–33, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367356. doi: 10.1145/3318464.3389772. URL `https://doi.org/10.1145/3318464.3389772`.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), jul 2009. ISSN 0360-0300. doi: 10.1145/1541880.1541882. URL `https://doi.org/10.1145/1541880.1541882`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

Steven R. Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28(3):1704 – 1728, 2022. doi: 10.3150/21-BEJ1388. URL `https://doi.org/10.3150/21-BEJ1388`.

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Md Rakibul Islam, Shubhomoy Das, Janardhan Rao Doppa, and Sriraam Natarajan. Glad: Glocalized anomaly detection via human-in-the-loop learning. *arXiv preprint arXiv:1810.01403*, 2018.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits, 2013.

Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park, Edgar Dobriban, Oleg Sokolsky, and Insup Lee. idecode: In-distribution equivariance for conformal out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7104–7114, 2022.

Aleksandr Khinchine. Über einen satz der wahrscheinlichkeitsrechnung. *Fundamenta Mathematicae*, 6:9–20, 1924.

Andrey Kolmogorov. Über das gesetz des iterierten logarithmus. *Mathematische Annalen*, 101: 126–135, 1929.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Gustaf Kylberg. *Kylberg texture dataset v. 1.0*. Centre for Image Analysis, Swedish University of Agricultural Sciences and . . . , 2011.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. Cider: Exploiting hyperspherical embeddings for out-of-distribution detection. *arXiv preprint arXiv:2203.04450*, 2022.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436. IEEE Computer Society, 2015.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

Nikolai Smirnov. Approximate laws of distribution of random variables from empirical data. *Uspekhi Matematicheskikh Nauk*, 10:179–206, 1944.

S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. Online outlier detection in sensor data using non-parametric models. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, VLDB '06, page 187–198. VLDB Endowment, 2006.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? *arXiv preprint arXiv:2110.06207*, 2021.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022.

Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021a.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021b.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection, 2022.

Yang Zhang, Nirvana Meratnia, and Paul J.M. Havinga. Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine. *Ad Hoc Networks*, 11(3): 1062–1074, 2013.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Appendix

The appendix is organized as follows. We begin with discussing the related works in section 4 and then provide the technical details and proofs of our theoretical results in sections 5 and 6. The notations are summarized the notation in Table 1. Further we provide details of experiments and insights from them in sections 7 and 8.

4. Related Works

Out-Of-Distribution detection: The problem of OOD detection has been addressed in many recent works where the main contributions have been methods to quantify a score (uncertainty) which gives a better separation of OOD and ID data points. Liang et al. Liang et al. (2017) proposed ODIN, which uses temperature scaling to separate the softmax score distributions between ID and OOD images. Liu et al. Liu et al. (2020) proposed a framework using energy score to perform OOD detection on pre-trained neural classifiers. Lee et al. Lee et al. (2018), Sehwag et al. Sehwag et al. (2021), and Ming et al. Ming et al. (2022) proposed Mahalanobis distance-based scores to detect OOD samples. While these methods perform well, the evaluation setup is rather static and does not reflect the real-world deployment scenario, wherein the system has to adapt to new and evolving OOD data. In our work, we are proposing a simple and extensible system for online OOD detection. Moreover, the system can also adapt by getting ground truth labels from humans on selected points.

Online anomaly detection: There is rich literature on anomaly (or outlier) detection in offline settings Chandola et al. (2009); Campos et al. (2016); Chalapathy and Chawla (2019). However, our setting is akin to the online anomaly (outlier) detection – wherein the system receives samples one at a time and it has to figure out the outliers or anomalous behavior within a given window of time. Some of the notable works along this line are Subramaniam et al. (2006); Angiulli and Fassetti (2007); Zhang et al. (2013). The methods proposed are unsupervised and perform density or distance-based detection.

Outlier detection with human in the Loop: The notion of an outlier may not always be based on statistical rarity and might need input from humans to learn the notion of an outlier in the application of interest. Some of the recent works Chai et al. (2020); Islam et al. (2018) have given methods for outlier detection in offline settings leveraging human inputs. The focus has been on minimizing human effort by figuring out some candidate outliers and designing good questions and context for getting human inputs.

While there are a number of works on outlier or OOD detection, the main focus has been on designing methods (scoring functions) to distinguish inliers vs outliers mostly in the offline setting. Our work is rather complementary – we consider a deployed OOD system that can work with any scoring function and propose ways for online adaptation of this system based on human inputs.

5. Theoretical Guarantees

In this section, we provide theoretical analysis and guarantees for Algorithm 1 when the distributions are fixed. Also, assume that the scores $g(x)$ have sub-Gaussian tails. Here we assume that $\mathcal{D}_{ood}$ is not changing. We provide anytime valid confidence intervals on the FPR at all thresholds which are used in the optimization problem (P2), using which we can guarantee that the FPR is always controlled. We also provide a bound on the time taken to reach feasibility, i.e., for the constraint in Equation (P2) to be feasible. Furthermore, we also provide the bound on time taken to reach η-Optimality which is defined as follows,

Definition 1 *(η-Optimality) For any η, the system is said to be operating in the η-Optimal regime after some time point T_η, if FPR(λ^*) − FPR$(\hat{\lambda}_t) \leq \eta$ for all $t \geq T_\eta$.*

Note that the values of η for which η-Optimality is achievable depend on the continuity of the CDF. In particular, it is achievable for any $\eta \geq \text{FPR}(\lambda^*) - \lim_{\epsilon \to 0^+} \text{FPR}(\lambda^* + \epsilon)$.

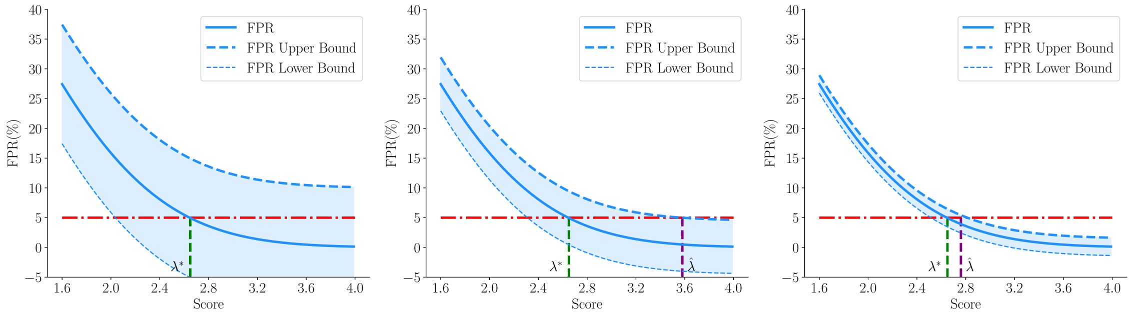
Theorem 2 *Let $\alpha, \delta \in (0,1)$. Let $x_t \overset{i.i.d}{\sim} (1-\gamma)\mathcal{D}_{id} + \gamma\mathcal{D}_{ood}$ and let $c_t = 1 - \beta_t + \frac{\beta_t}{p^2}$, $\beta_t = \frac{N_t^{(o,p)}}{N_t^{(o)}}$ where $N_t^{(o,p)}$ is the number of OOD points sampled using importance sampling until time t and $N_t^{(o)}$ is the total number of OOD points observed till time t. Let $n_0 = \min\{u : c_u N_u^{(o)} \geq 173\log(\frac{8}{\delta})\}$ and t_0 be such that $N_{t_0}^{(o)} \geq n_0$. If Algorithm 1 uses the optimization problem* (P2) *to find the thresholds with the upper confidence term $\psi(N_t^{(o)}, \delta/2)$ given by equation* (3)*, then with probability at least $1-\delta$,*

1. ***Controlled FPR**: For all $t \geq t_0$, FPR$(\hat{\lambda}_t) \leq \alpha$.*
2. ***Time to reach feasibility** Let $T_f = \frac{2C_1}{\gamma\alpha^2}\log\left(\frac{4C_2}{\delta}\log(\frac{C_3}{\alpha})\right) + \frac{1}{\gamma^2}\log(\frac{4}{\delta})$, then for any $t \geq \max(t_0, T_f)$ the algorithm will find a feasible threshold, $\hat{\lambda}_t$ such that $\widehat{FPR}(\hat{\lambda}_t) + \psi(N_t^{(o)}) \leq \alpha$.*
3. ***Time to reach $\eta-$Optimality** Let $T_{opt} = \frac{8C_1}{\gamma\eta^2}\log\left(\frac{4C_2}{\delta}\log(\frac{2C_3}{\alpha})\right) + \frac{1}{\gamma^2}\log(\frac{4}{\delta})$ and $\widehat{FPR}(\hat{\lambda}_{T_{opt}}) \in [\alpha - \eta/2, \alpha]$, then for any $t \geq \max(t_0, T_{opt})$,$\hat{\lambda}_t$ satisfy the η-Optimality condition in definition 1.*

Lemma 3 *$\widehat{FPR}(\lambda, t)$ as defined in equation* (4) *is an unbiased estimate of the true FPR(λ), i.e., $\mathbb{E}[\widehat{FPR}(\lambda, t)] =$ FPR(λ).*

The above theorem establishes key properties of Algorithm 1 and provides insights into its behavior and performance guarantees. We now discuss each property in detail, along with their implications.

Controlled false positive rate: The first property of Theorem 9 ensures that the Algorithm 1 effectively controls the False Positive Rate (FPR) throughout its operation. Specifically, it guarantees that for all time steps $t \geq t_0$, the FPR of the estimated threshold obtained by the algorithm will be less than or equal to a predetermined threshold α. This property is crucial in applications where accurately controlling the rate of false positives is essential. By limiting the FPR to a predefined threshold, Algorithm 1 provides a reliable mechanism for distinguishing between in-distribution and out-of-distribution samples, reducing the likelihood of erroneous classifications.



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

Figure 3: Illustration of the confidence intervals on FPR and their effect on threshold estimation. We expect as the system receives more OOD samples the confidence intervals will shrink and lead to a better threshold.

Time to reach feasibility: The second property of Theorem 9 concerns the time it takes for Algorithm 1 to find a feasible threshold. It provides conditions under which the algorithm is guaranteed to discover a suitable threshold, $\hat{\lambda}_t$, that has FPR at most α. It is contingent upon the feasible time T_f, the time step at which a sufficient number of observations $N^{(o)}_{T_f}$ is obtained so that the confidence interval $\psi(t, \delta/2) \leq \alpha$.

Time to reach η-Optimality: The third property provides a bound on the time T_{opt} after which the Algorithm 1 achieves the η-Optimal regime. This regime implies that the algorithm operates in a state where the difference between the FPR of the true optimal threshold, FPR(λ^*), and the FPR of the estimated threshold FPR$(\hat{\lambda}_t)$, is within the range η. The theorem says that, if the estimated FPR at time step T_{opt}, denoted as $\widehat{\text{FPR}}(\hat{\lambda}_{T_{opt}})$, is within the range $[\text{FPR}(\lambda^*) - \eta/2, \alpha]$ and the confidence interval $\psi(T_{opt}, \delta/2) \leq \eta/2$. Then for all time points after T_{opt} the algorithm will find a $\hat{\lambda}_t$ that satisfies the $\eta-$ Optimality condition. T_{opt} is defined in terms of the time point when the number of acquired OOD samples $N^{(o)}_{T_{opt}}$ becomes at least $\frac{4C_1}{\eta^2}\log\left(\frac{2C_2}{\delta}\log(\frac{2C_3}{\eta})\right)$. We require these many samples in order to guarantee the confidence intervals $\psi(t, \delta)$ are sufficiently small (of the order of η in this case) so that when the empirical estimate of FPR is very close to α we know that the algorithm will return a threshold satisfying $\eta-$Optimality.

If γ is not changing, it is very easy to bound T_f and T_{opt}. When t is large enough, with a high probability γ fraction of what is observed is going to be OOD. So, for T_f, N_f will concentrate around γT_f. And similarly for T_{opt}, N_{opt} while also accounting for importance sampling.

The details of the proof of the statements in the main theorem are provided in the appendix. Here we provide the key results and outline the key ideas of the proof.

Proof outline and discussion: The main technical challenge is to obtain accurate confidence intervals $\psi(t, \delta)$ that are simultaneously valid with high probability for the FPR estimates at all time points and all thresholds. Fortunately, there is a rich line of work that provide tight confidence intervals valid for all times based on the Law of Iterated Logarithm (LIL) Khinchine (1924); Kolmogorov (1929); Smirnov (1944). Non-asymptotic versions of LIL have been proved in various settings e.g. multi-armed bandits Jamieson et al. (2013), quantile estimation Howard and Ramdas

(2022). Roughly speaking, these bounds provide confidence intervals that are $\mathcal{O}(\sqrt{\log\log(t)/t}$ and are known to be tight. However, most of these works assume i.i.d samples. In our setting, our treatment of observing the human feedback is dependent on whether the score is above or below $\hat{\lambda}_{t-1}$ which itself is estimated using all the past data which creates dependence. Though this way of sampling saves unnecessary queries for expert labels, it breaks the independence between samples and prevents us from utilizing results developed for independent samples in Howard and Ramdas (2022).

We handle this by first showing that there is a martingale structure that we then exploit by using LIL results for martingales Balsubramani (2015). A limitation of Balsubramani (2015) is that it can only provide us confidence intervals valid for FPR estimate for a given threshold λ. However, we need intervals that are simultaneously valid for all λ as well. Building upon the work in Balsubramani (2015) we derive confidence intervals that are simultaneously valid for all t and finitely many thresholds. Equation (3) shows the $\psi(t, \delta)$ we obtain. Please see section 6 for detailed proofs and discussion.

Glossary

The notation is summarized in Table 1 below.

6. Proofs

At each time t, we observe $x_t \overset{i.i.d}{\sim} (1-\gamma)\mathcal{D}_{id} + \gamma\mathcal{D}_{ood}$, and $s_t = g(x_t)$ is the corresponding score. If $s_t \leq \hat{\lambda}_{t-1}$, then it is considered an OOD point and hence gets a human label for it and we get to know whether it is in fact OOD or ID. If $s_t > \hat{\lambda}_{t-1}$, then it is considered an ID point and hence gets a human label only with probability p. So, we get to know whether it is truly ID or not with probability p. Now we have to update the threshold, $\hat{\lambda}_t$, such that the FPR$(\hat{\lambda}_t) \leq \alpha$ for all t, while trying to maximize TPR$(\hat{\lambda}_t)$. Our approach is based on constructing an unbiased estimator of FPR(λ) using the OOD samples received till time t and in conjunction with confidence intervals for FPR(λ) for at all thresholds $\lambda \in \Lambda$ that is valid for all times simultaneously. Together, at each time t, these give us a reliable upper bound on the true FPR(λ) for all λ enabling us to find the smallest λ such that the upper bound on FPR(λ) is at most α. Let $S_t^{(o)} = \{s_1^{(o)}, \ldots s_{N_t^{(o)}}^{(o)}\}$ denote the scores of the points that have been truly identified as OOD points from human feedback and $I_t^{(o)}$ be the corresponding time points. We estimate the FPR as follows,

$$\widehat{\text{FPR}}(\lambda, t) = \frac{1}{N_t^{(o)}} \sum_{u \in I_t^{(o)}} Z_u(\lambda), \quad \text{where } Z_u(\lambda) := \begin{cases} \mathbf{1}(s_u^{(o)} > \lambda), & \text{if } s_u^{(o)} \leq \hat{\lambda}_{u-1} \\ \frac{1}{p}\mathbf{1}(s_u^{(o)} > \lambda), & \text{w.p. } p \text{ if } s_u^{(o)} > \hat{\lambda}_{u-1} \\ 0, & \text{w.p. } 1-p \text{ if } s_u^{(o)} > \hat{\lambda}_{u-1} \end{cases}. \tag{4}$$

Next, we show that the above estimator $\widehat{\text{FPR}}(\lambda, t)$ is indeed an unbiased of false positive rate FPR(λ).

Lemma 4 *$\widehat{FPR}(\lambda, t)$ as defined in equation* (4) *is an unbiased estimate of the true FPR(λ), i.e., $\mathbb{E}[\widehat{FPR}(\lambda, t)] = FPR(\lambda)$.*

Proof Let $i_t^{(o)}$ be the indicator variable denoting whether $s_t^{(o)}$ was sampled using importance sampling (i.e. $i_t^{(o)} = 1$) or not (i.e. $i_t^{(o)} = 0$). Denote the pair as $r_t^{(o)} = (s_t^{(o)}, i_t^{(o)})$ for brevity.

Symbol	Definition
$\mathcal{X}$	feature space.
$\mathcal{Y}$	label space, $\{+1, -1\}$, +1 for ID and -1 for OOD .
$\mathcal{D}_{id}, \mathcal{D}_{ood}$	distributions of ID and OOD points.
γ	mixing ratio of OOD and ID distributions.
λ	threshold for OOD classification.
FPR(λ)	population level false positive rate with threshold λ.
TPR(λ)	population level true positive rate with threshold λ.
$\widehat{\text{FPR}}(\lambda, t)$	empirical FPR at time t, adjusted to account for importance sampling (see eq. (4)).
λ^*	the optimal threshold for OOD classification s.t. FPR$(\lambda) \leq \alpha$ and TPR(λ) is maximized.
$\hat{\lambda}_t$	the estimated threshold at round t.
x_t, y_t	sample and the true label at time t .
g	OOD uncertainty quantification (score) function.
$s_u^{(o)}$	score of u^{th} OOD sample.
$i_u^{(o)}$	indicator variable denoting whether $s_u^{(o)}$ was importance sampled or not.
$N_t^{(o)}$	number of OOD points till time t.
$N_t^{(o,p)}$	number of OOD points sampled using importance sampling until time t.
β_t	it is equal to $N_t^{(o,p)}/N_t^{(o)}$.
p	probability for importance sampling.
δ	failure probability.
α	user given upper bound on FPR that the algorithm needs to maintain.
η	the algorithm is in $\eta-$optimality if close FPR$(\lambda^*) -$ FPR$(\hat{\lambda}_t) \leq \eta$.
$\Lambda_{\min}, \Lambda_{\max}$	the minimum and maximum scores(thresholds) considered by the algorithm.
ν	discretization parameter for the interval $[\Lambda_{\min}, \Lambda_{\max}]$ set by the user.
$\psi(t, \delta)$	LIL based confidence interval at time t.

Table 1: Glossary of variables and symbols used in this paper.

$$\begin{aligned}
\mathbb{E}_{r_t^{(o)}, r_{t-1}^{(o)}, \ldots, r_1^{(o)}}[\widehat{\text{FPR}}(\lambda, t)] &= \frac{1}{N_t^{(o)}} \sum_{u \in I_t^{(o)}} \mathbb{E}_{r_u^{(o)} | r_{u-1}^{(o)}, \ldots, r_1^{(o)}}[Z_u(\lambda)] \\
&= \frac{1}{N_t^{(o)}} \sum_{u \in I_t^{(o)}} \mathbb{E}_{r_u^{(o)} | \hat{\lambda}_{u-1}}[Z_u(\lambda)] \\
&= \frac{1}{N_t^{(o)}} \sum_{u \in I_t^{(o)}} \mathbb{E}_{(s_u^{(o)}, i_u^{(o)}) | \hat{\lambda}_{u-1}}[Z_u(\lambda)] \\
&= \frac{1}{N_t^{(o)}} \sum_{u \in I_t^{(o)}} \mathbb{E}_{s_u^{(o)} | \hat{\lambda}_{u-1}}[\mathbb{E}_{i_u^{(o)} | s_u^{(o)}, \hat{\lambda}_{u-1}}[Z_u(\lambda)]] \\
&= \frac{1}{N_t^{(o)}} \sum_{u \in I_t^{(o)}} \mathbb{E}_{s_u^{(o)} | \hat{\lambda}_{u-1}}[\mathbf{1}(s_u^{(o)} > \lambda)]
\end{aligned}$$

$$= \frac{1}{N_t^{(o)}} \sum_{u \in I_t^{(o)}} \text{FPR}(\lambda)$$

$$= \text{FPR}(\lambda)$$

■

Having an unbiased estimator solves one part of the problem. In addition we need confidence intervals on this estimate that are valid for anytime and for the choices of λ considered. Due to the dependence between the samples we cannot directly apply similar results developed for quantile estimation in the i.i.d. setting Howard and Ramdas (2022). Fortunately, part of this problem has been addressed in Balsubramani (2015), where they provide anytime valid confidence intervals when the estimators form a martingale sequence. We restate this result in the following lemma 5 and then building upon this result, in the next lemma 6 we derive such confidence intervals for our setting.

Lemma 5 *(Balsubramani (2015)) Let $\overline{M}_t$ be a martingale and suppose $|\overline{M}_t - \overline{M}_{t-1}| \leq \rho_t$ for constants $\{\rho_t\}_{t>1}$, let $m_0 = \min_{t\geq 1} |\overline{M}_t|$. Fix any $\delta \in (0,1)$, and let $t_0 = \min\{u : \sum_{t=1}^{u} \rho_t^2 \geq 173 \log(\frac{4}{\delta})\}$ then,*

$$\mathbb{P}\bigg(\exists t \geq t_0 : |\overline{M}_t| \geq \sqrt{3\Big(\sum_{i=1}^{t} \rho_i^2\Big)\Big(2 \log\log\big(\frac{3\sum_{i=1}^{t} \rho_i^2}{m_0}\big) + \log\big(\frac{2}{\delta}\big)\Big)}\bigg) \leq \delta \tag{5}$$

Proof This lemma is a restatement of theorem 4 in Balsubramani (2015). For proof details please see Balsubramani (2015). ■

In the next lemma we show that the sums of $Z_u(\lambda)$ form a martingale sequence, allowing us to apply the results from the above lemma (5) and then we generalize it to all λ in some finite set.

Lemma 6 *(Anytime valid confidence intervals on FPR) Let $X_t^{(o)} = \{x_1^{(o)}, \ldots x_{N_t{(o)}}^{(o)}\}$ be the samples drawn from D_{ood} till round t and let $S_t^{(o)} = \{s_1^{(o)}, \ldots s_{N_t{(o)}}^{(o)}\}$ be the scores of these points, let $c_t = 1 - \beta_t + \frac{\beta_t}{p^2}$, $\beta_t = \frac{N_t^{(o,p)}}{N_t^{(o)}}$ and $N_t^{(o,p)}$ is the number of points sampled using importance sampling until time t and $\nu \in (0,1)$ is a discretization parameter set by the user. Let $\Lambda = \{\Lambda_{\min}, \Lambda_{\min} + \nu, \ldots, \Lambda_{\max}\}$. Let $n_0 = \min\{u : c_u N_u^{(o)} \geq 173 \log(\frac{4}{\delta})\}$ and t_0 be such that $N_{t_0}^{(o)} \geq n_0$. , then for any $\delta \in (0,1)$,*

$$\mathbb{P}\bigg(\exists t \geq t_0 : \sup_{\lambda \in \Lambda} \widehat{FPR}(\lambda, t) - FPR(\lambda) \geq \psi(t, \delta)\bigg) \leq \delta \tag{6}$$

for,

$$\psi(t, \delta) = \sqrt{\frac{3c_t}{N_t^{(o)}} \Big[2 \log\log\Big(\frac{3c_t N_t^{(o)}}{2}\Big) + \log\Big(\frac{2|\Lambda|}{\delta}\Big)\Big]} \tag{7}$$

Proof First, we show that we have a martingale sequence as follows,

Let $M_t(\lambda) = \sum_{u=1}^{N_t^{(o)}} Z_u(\lambda)$, and consider the centered random variables,

$$\overline{M}_t(\lambda) = M_t(\lambda) - \mathbb{E}[M_t(\lambda)] \quad \text{and} \quad \overline{Z}_t(\lambda) = Z_t(\lambda) - \text{FPR}(\lambda)$$

Let $\mathcal{F}_t$ be the $\sigma-$algebra of events till time t i.e. $(s_1^{(o)}, i_1^{(o)}), \dots, (s_{t-1}^{(o)}, i_{t-1}^{(o)}), (s_t^{(o)}, i_t^{(o)})$.

It is easy to see that $\mathbb{E}[\overline{M}_t] \leq \frac{1}{p} < \infty$ and $\overline{M}_t$ is $\mathcal{F}_t$-measurable for all $t > 1$. Further, we can see,

$$\mathbb{E}\big[\overline{M}_t(\lambda)|\mathcal{F}_{t-1}\big] = \mathbb{E}[\overline{Z}_t(\lambda) + \overline{M}_{t-1}(\lambda)|\mathcal{F}_{t-1}] = \mathbb{E}[\overline{Z}_t(\lambda)|\mathcal{F}_{t-1}] + \mathbb{E}[\overline{M}_{t-1}(\lambda)|\mathcal{F}_{t-1}] = \overline{M}_{t-1}(\lambda)$$

Since, $\mathbb{E}[\overline{Z}_t(\lambda)|\mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[\overline{M}_{t-1}(\lambda)|\mathcal{F}_{t-1}] = \overline{M}_{t-1}(\lambda)$. Thus we have that $\overline{M}_t$ is a martingale sequence. Further, we also have the following,

$$|\overline{M}_t(\lambda) - \overline{M}_{t-1}(\lambda)| \leq \begin{cases} 1 & \text{if } i_t^{(o)} = 0 \\ \frac{1}{p} & \text{if } i_t^{(o)} = 1 \end{cases}$$

Let $\beta_t \in (0,1)$ be the fraction of OOD points sampled using probability p till round t. Let $N_t^{(o)}$ be the total number of points OOD points sampled till round t and $N_t^{(o,p)}$ be the points sampled from importance sampling, then $\beta_t = \frac{N_t^{(o,p)}}{N_t^{(o)}}$.

Let $c_t = 1 - \beta_t + \frac{\beta_t}{p^2}$. We know p and the number of points sampled with importance sampling, without importance sampling we knowβ_t, c_t are at time t. Applying lemma 5 we get the following result for a given λ,

$$\mathbb{P}\bigg(\exists t \geq t_0 : \overline{M}_t(\lambda) \geq \sqrt{3\Big(c_t N_t^{(o)}\Big)\Big(2\log\log\big(3c_t N_t^{(o)}\big) + \log\big(\frac{2}{\delta}\big)\Big)}\bigg) \leq \delta \tag{8}$$

$$\mathbb{P}\bigg(\exists t \geq t_0 : \widehat{\text{FPR}}(\lambda, t) - \text{FPR}(\lambda, t) \geq \sqrt{\frac{3c_t}{N_t^{(o)}}\Big(2\log\log\big(3c_t N_t^{(o)}\big) + \log\big(\frac{2}{\delta}\big)\Big)}\bigg) \leq \delta \tag{9}$$

Doing the union bound for the failure probability over all $\lambda \in \Lambda$, (where $|\Lambda| < \infty$) gives us the result.

■

Our performance guarantees in the main theorem 9 are based on $\psi(t, \delta)$ becoming smaller than certain values. In the next lemma we derive bound on $N_t^{(o)}$ such that $\psi(t, \delta)$ is at most μ and use it in the proof of the main theorem 9.

Lemma 7 *Let $\psi(t, \delta) = \sqrt{\frac{3c_t}{N_t^{(o)}}\Big(2\log\log\Big(3c_t N_t^{(o)}\Big) + \log\Big(\frac{2|\Lambda|}{\delta}\Big)\Big)}$, and let there be a constant C_0 and time T_0, such that $\beta_t \leq C_0 p^2$ for all $t \geq T_0$ (worst case $T_0 = 1$ and $C_0 = 1/p^2$). Then $\psi(t, \delta) \leq \mu$ for any $t > T_\mu > T_0$ such that $N_{T_\mu}^{(o)} = \frac{10(C_0+1)}{\mu^2}\log\Big(\frac{|\Lambda|}{\delta}\log(\frac{5(C_0+1)}{\mu})\Big)$.*

Proof First we write a simplified form of $\psi(t,\delta)$ for all $t > T_0$ as follows,

$$\psi(t,\delta) = \sqrt{\frac{3(C_0+1)}{N_t^{(o)}}\left(2\log\log\left(3(C_0+1)N_t^{(o)}\right) + \log\left(\frac{2|\Lambda|}{\delta}\right)\right)}$$

In the above equation we used the bound on $\beta_t \leq C_0 p^2$ in the equation $c_t = 1 - \beta_t + \beta_t/p^2$ leading to $c_t \leq C_0 + 1$, Now, for brevity let $a_1 = 3(C_0+1)$ and $a_2 = 2|\Lambda|$ and rewrite $\psi(t,\delta)$ as follows,

$$\psi^2(t,\delta) = \frac{a_1}{N_t^{(o)}}\left(2\log\log\left(a_1 N_t^{(o)}\right) + \log\left(\frac{a_2}{\delta}\right)\right) \leq \frac{2a_1}{N_t^{(o)}}\left(\log\left(\frac{a_2}{\delta}\log\left(a_1 N_t^{(o)}\right)\right)\right)$$

We want to find $N_t^{(o)}$ such that $\psi^2(t,\delta) \leq \mu^2$. It is difficult to directly invert this function. To get a bound on $N_t^{(o)}$ we first assume the following form for it with unknown constants $b_1, b_2, b_3 > 0$ and then figure out the constants by simplifying $\psi^2(N_t^{(o)})$ and constraining it to be at most μ^2.

$$\text{Let } N_{T_\mu}^{(o)} = \frac{b_1}{\mu^2}\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)$$

$$\begin{aligned}
\psi^2(T_\mu,\delta) &\leq \frac{2a_1}{N_{T_\mu}^{(o)}}\log\left[\frac{a_2}{\delta}\log(a_1 N_{T_\mu}^{(o)})\right] \\
&= \frac{2a_1\mu^2}{b_1\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)}\log\left[\frac{a_2}{\delta}\log\left\{\frac{a_1 b_1}{\mu^2}\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)\right\}\right] \\
&\overset{a}{\leq} \frac{2a_1\mu^2}{b_1\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)}\log\left[\frac{a_2}{\delta}\log\left\{\frac{a_1 b_1}{\mu^2}\log\left(\frac{a_2}{b_3\delta}\frac{b_2}{\mu}\right)\right\}\right] \\
&= \frac{2a_1\mu^2}{b_1\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)}\log\left[\frac{a_2}{\delta}\log\left\{\frac{a_1 b_1}{\mu^2}\log\left(\frac{a_2 b_2}{b_3\delta\mu}\right)\right\}\right] \\
&\overset{b}{\leq} \frac{2a_1\mu^2}{b_1\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)}\log\left[\frac{a_2}{\delta}\log\left\{\frac{a_1 b_1}{\mu^2}\frac{a_2 b_2}{b_3\delta\mu}\right\}\right] \\
&= \frac{2a_1\mu^2}{b_1\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)}\log\left[\frac{a_2}{\delta}\log\left\{\frac{a_1 b_1}{\mu^2}\frac{a_2 b_2}{b_3\delta\mu}\right\}\right] \\
&= \frac{2a_1\mu^2}{b_1\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)}\log\left[\frac{a_2}{\delta}\log\left\{\frac{a_1 b_1 a_2}{b_3 b_2^2\delta}\left(\frac{b_2}{\mu}\right)^3\right\}\right] \\
&\overset{c}{\leq} \frac{2a_1\mu^2}{b_1\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)}\log\left[\frac{a_2}{\delta}\frac{a_1 b_1 a_2}{b_3 b_2^2\delta}\log\left\{\left(\frac{b_2}{\mu}\right)^3\right\}\right] \\
&= \frac{2a_1\mu^2}{b_1\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)}\log\left[\frac{a_2}{\delta}\frac{3a_1 b_1 a_2}{b_3 b_2^2\delta}\log\left(\frac{b_2}{\mu}\right)\right] \\
&= \frac{2a_1\mu^2}{b_1\log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)}\log\left[\left(\frac{a_2}{b_3\delta}\right)^2\frac{3a_1 b_1 b_3}{b_2^2}\log\left(\frac{b_2}{\mu}\right)\right]
\end{aligned}$$

$$\begin{aligned}
&\overset{d}{\leq} \frac{2a_1\mu^2 \frac{3a_1b_1b_3}{b_2^2}}{b_1 \log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)} \log\left[\left(\frac{a_2}{b_3\delta}\right)^2 \log\left(\frac{b_2}{\mu}\right)\right] \\
&\overset{e}{\leq} \frac{2a_1\mu^2 \cdot 2\frac{3a_1b_1b_3}{b_2^2}}{b_1 \log\left(\frac{a_2}{b_3\delta}\log\left(\frac{b_2}{\mu}\right)\right)} \log\left[\frac{a_2}{b_3\delta} \log\left(\frac{b_2}{\mu}\right)\right] \\
&= \frac{12\mu^2 a_1^2 b_3}{b_2^2}.
\end{aligned}$$

The inequalities a, b follow from $\log(x) \leq x$ for any $x > 0$.

The inequality c comes from $\log(ax) \leq a\log(x)$ for any $a > 2, x > 2$. We use $a = \frac{a_1b_1a_2}{b_3b_2^2\delta}$ and $x = \left(\frac{b_2}{\mu}\right)^3$, this enforces the following constraints,

$$\frac{b_2}{\mu} > 2^{1/3} \tag{10}$$

$$\frac{a_1b_1a_2}{b_3b_2^2\delta} > 2 \tag{11}$$

For d we again use $\log(ax) \leq a\log(x)$ with $a = \frac{3a_1b_1b_3}{b_2^2}$ and $x = \left(\frac{a_2}{b_3\delta}\right)^2 \log\left(\frac{b_2}{\mu}\right)$, this enforces the following constraints,

$$\frac{3a_1b_1b_3}{b_2^2} > 2 \tag{12}$$

$$\left(\frac{a_2}{b_3\delta}\right)^2 \log\left(\frac{b_2}{\mu}\right) > 2 \tag{13}$$

Lastly, e follows by using $\log(x^a y) \leq a\log(xy)$ for any $x > 0, a > 1, y > 1$. For this we use $x = \frac{a_2}{b_3\delta}$ and $y = \log(\frac{b_2}{\mu})$, leading the following constraints,

$$\log(\frac{b_2}{\mu}) > 1 \tag{14}$$

For $\psi^2(T_\mu) \leq \mu^2$, we need

$$12a_1^2b_3 \leq b_2^2 \tag{15}$$

Let $b_3 = 2, b_1 = 10a_1, b_2 = 5a_1$ then the constraints 10,11,12,13,14 and 15 are satisfied (when $|\Lambda| \geq 10$) for any $\mu \in (0,1), \delta \in (0,1)$. Thus we have,

$$\psi(T_\mu, \delta) \leq \mu \;\text{ for }\; N_{T_\mu} = \frac{10(C_0+1)}{\mu^2} \log\left(\frac{|\Lambda|}{\delta} \log(\frac{5(C_0+1)}{\mu})\right) \tag{16}$$

■

Lemma 8 *Let the data points $\{x_t\}_{t\geq1}$ be independent draws from the mixture distribution $(1-\gamma)\mathcal{D}_{id} + \gamma\mathcal{D}_{ood}$, and $N_t^{(o)}$ be the number of OOD points received till time t from this distribution, then for any $\delta \in (0,1)$ for any $t \geq T_k$ we have $N_t^{(o)} \geq k$ w.p. $1-\delta$, where T_k is given as follows,*

$$T_k = \frac{2k}{\gamma} + \frac{1}{\gamma^2}\log(\frac{1}{\delta}). \tag{17}$$

Proof We want to find t such that $N_t^{(o)} \geq k$ w.p. $1-\delta$. This is the same as finding the number of coin tosses of a coin with bias γ so that the number of heads observed is at least k. Applying Hoeffding's inequality gives us the following w.p. $1-\delta$,

$$N_t^{(o)} \geq \gamma t - \sqrt{\frac{t}{2}\log(\frac{1}{\delta})}.$$

Equating the r.h.s. above with k and solving for t will give us the desired bound on t. Note that it is enough to have an upper bound on t that satisfies the following and then use that upper bound as T_k.

$$\gamma t - \sqrt{\frac{t}{2}\log(\frac{1}{\delta})} = k.$$

To simplify the calculations, let $c = \sqrt{\frac{1}{2}\log\frac{1}{\delta}}$ and let $t = u^2$ then we have the following quadratic equation,

$$\gamma u^2 - cu - k = 0.$$

Considering the larger of the two solutions,

$$u = \frac{c + \sqrt{c^2 + 4k\gamma}}{2\gamma}.$$

Using the fact that for any $a, b \geq 0$, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$,

$$u \leq \frac{c + \sqrt{c^2} + \sqrt{4k\gamma}}{2\gamma} = \frac{2c + 2\sqrt{k\gamma}}{2\gamma} = \frac{c}{\gamma} + \sqrt{\frac{k}{\gamma}}.$$

Lastly, using $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$ we get the following upper bound on t,

$$t = u^2 \leq \frac{2c^2}{\gamma^2} + \frac{2k}{\gamma} = \frac{2k}{\gamma} + \frac{1}{\gamma^2}\log(\frac{1}{\delta}).$$

■

Theorem 9 *Let $\alpha, \delta \in (0,1)$. Let $x_t \overset{i.i.d}{\sim} (1-\gamma)\mathcal{D}_{id} + \gamma\mathcal{D}_{ood}$ and let $c_t = 1 - \beta_t + \frac{\beta_t}{p^2}$, $\beta_t = \frac{N_t^{(o,p)}}{N_t^{(o)}}$ where $N_t^{(o,p)}$ is the number of OOD points sampled using importance sampling until time t and $N_t^{(o)}$ is the total number of OOD points observed till time t. Let $n_0 = \min\{u : c_u N_u^{(o)} \geq 173\log(\frac{8}{\delta})\}$ and t_0 be such that $N_{t_0}^{(o)} \geq n_0$. If Algorithm 1 uses the optimization problem* (P2) *to find the thresholds with the upper confidence term $\psi(N_t^{(o)}, \delta/2)$ given by equation* (3)*, then with probability at least $1-\delta$,*

1. *(Controlled FPR) For all $t \geq t_0$, FPR$(\hat{\lambda}_t) \leq \alpha$.*
2. *(Time to reach feasibility) Let $T_f = \frac{2C_1}{\gamma\alpha^2}\log\left(\frac{4C_2}{\delta}\log(\frac{C_3}{\alpha})\right) + \frac{1}{\gamma^2}\log(\frac{4}{\delta})$, then for any $t \geq \max(t_0, T_f)$ the algorithm will find a feasible threshold, $\hat{\lambda}_t$ such that $\widehat{FPR}(\hat{\lambda}_t) + \psi(N_t^{(o)}) \leq \alpha$.*

3. *(Time to reach $\eta-$Optimality) Let $T_{opt} = \frac{8C_1}{\gamma\eta^2}\log\left(\frac{4C_2}{\delta}\log(\frac{2C_3}{\alpha})\right) + \frac{1}{\gamma^2}\log(\frac{4}{\delta})$ and $\widehat{FPR}(\hat{\lambda}_{T_{opt}}) \in [FPR(\lambda^*) - \eta/2, \alpha]$, then for any $t \geq \max(t_0, T_{opt})$,$\hat{\lambda}_t$ satisfy the η-Optimality condition in definition 1.*

Proof

To prove this we first obtain confidence intervals on FPR valid w.p. $1 - \delta/2$ using Lemma 6. Then applying Lemma 7 on these confidence intervals gives us the number of OOD samples that are sufficient to guarantee certain width of the confidence intervals and lastly we use Lemma 8 to bound the time point such that we observe a certain number of OOD points till that time. We do this for the second and third points separately each time invoking Lemma 8 with failure probability $\delta/4$ and then union bound over them.

Controlled FPR: This follows from Lemma 6 (with probability $1 - \frac{\delta}{2}$) and the fact the algorithm uses confidence intervals on FPR estimate that are valid for all $t \geq t_0$ for the choices of λ it considers.

Time to reach feasibility: Applying Lemma 7 with $\mu = \alpha$ gives bound on N_{T_f} with $C_1 = 10(C_0 + 1), C_2 = |\Lambda|, C_3 = 5(C_0 + 1)$. Then using Lemma 8 with $k = N_{T_f}$ gives us the desired T_f.

Time to reach η-optimality : We know, FPR$(\lambda^*) = \alpha$, and it is given that $\widehat{\text{FPR}}(\hat{\lambda}_t) \in [\text{FPR}(\lambda^*) - \eta/2, \alpha]$

$$\text{FPR}(\hat{\lambda}_t) \in [\text{FPR}(\lambda^*) - \eta/2 - \psi(t, \delta), \alpha]$$

this means FPR$(\hat{\lambda}_t) \geq$ FPR$(\lambda^*) - \eta/2 - \psi(t, \delta)$

$$\text{FPR}(\lambda^*) - \text{FPR}(\hat{\lambda}_t) \leq \eta/2 + \psi(t, \delta)$$

If $\psi(t, \delta) \leq \eta/2$ we have, FPR$(\lambda^*) -$ FPR$(\hat{\lambda}_t) \leq \eta$. Thus applying we want to find t for which $\psi(t, \delta) = \eta/2$. Applying lemma 7 with $\mu = \eta/2$ gives bound on N_{T_f} with $C_1 = 40(C_0 + 1), C_2 = |\Lambda|, C_3 = 10(C_0 + 1)$. Then using Lemma 8 with $k = N_{T_{opt}}$ gives us the desired T_{opt}. ■

This concludes the proofs of the main results. Next, we present additional experiments on synthetic and real datasets.

7. Empirical Evaluation

Methods: We evaluate our method on synthetic and real-world image classification datasets. We compare our (a) LIL confidence interval based method against (b) No-UCB: which does not use any confidence intervals (c) Hoeffding: which uses the confidence intervals from Hoeffding's inequality. We consider two variations of each method, one without using a window and the other using a window size. We expect that No-UCB will violate the FPR constraint since it does not account for the uncertainty in the estimates. While the methods that accurately account for the uncertainty using confidence intervals like LIL, and Hoeffding are expected to adhere to the FPR constraints. We note, that the confidence intervals from Hoeffding inequality are not theoretically valid for these settings but are a reasonable choice for a practitioner, and in our evaluation, we do not observe significant differences between Hoeffding and LIL-based bounds in the results. We use $\alpha = 0.05$, $\delta = 0.2$, and importance sampling probability $p = 0.2$ through all the empirical evaluations.

7.1 Searching for constants in LIL-Heuristic

The theoretical LIL bound in equation3 has constants that can be pessimistic in practice. We get around this by using a LIL-Heuristic bound which has the same form as in equation (3) but with different constants in particular we consider the form in equation LIL-Heuristic. We find the constants C_1, C_2, C_3 using a simulation on estimating the bias of a coin with different constants and picking the ones so that the observed failure probability is below 5%.

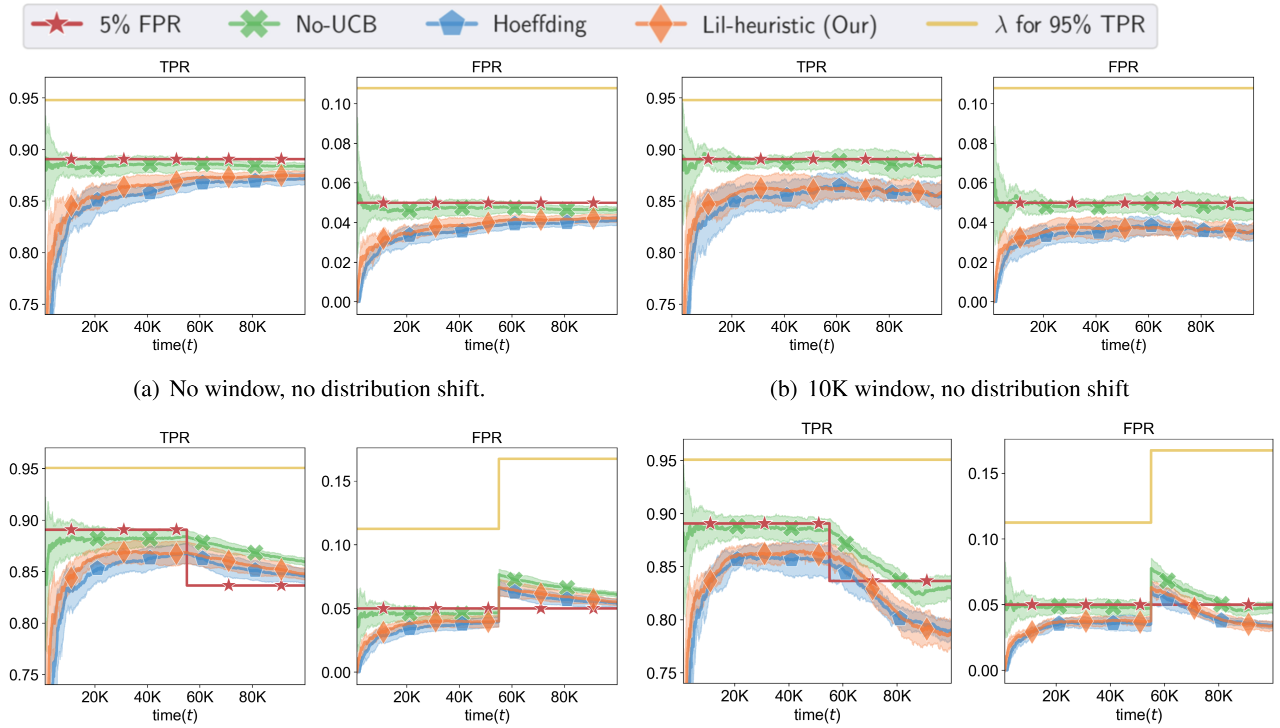
$$\tilde{\psi}(t, \delta) = C_1 \sqrt{\frac{c_t}{N_t^{(o)}} \left(\log\log\left(C_2 c_t N_t^{(o)}\right) + \log\left(\frac{C_3}{\delta}\right) \right)}. \tag{LIL-Heuristic}$$

Specifically, we keep $C_3 = 1$, and run for $\delta \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4\}$ with varying C_1 from 0.1 to 0.9 and C_2 from 1.5 to 4.75. For each choice of δ, C_1, C_2, we toss an unbiased coin (mean $p = 0.5$) for $T = 10k$ times. For each choice of $t = 1, 2, \cdots, T$, we compute the empirical mean $\hat{p}$ of the coin and define it as a failure if $p \notin [\hat{p} - \tilde{\psi}(t, \delta), \hat{p} + \tilde{\psi}(t, \delta)]$. We run this process for 100 times and compute the average failure probability for each choice of $t = 1, 2, \cdots, T$. We then pick the constant so that the observed average probability is below 5%. Throughout the paper, we use $C_1 = 0.5$ and $C_2 = 0.75$.

7.2 Simulations

Synthetic data setup: We simulate the OOD and ID scores using a mixture of two Gaussians $\mathcal{N}_{id}(\mu = 5.5, \sigma = 4)$ and $\mathcal{N}_{ood}(\mu = -6, \sigma = 4)$. We randomly draw and shuffle 100k samples with ID: OOD sample ratios of 2:1 in figure.4 and 4:1 in figure.5. We run the following simulation to better understand the performance of the system. We use $\alpha = 0.05$, $\delta = 0.2$, and importance sampling probability $p = 0.2$ through all the simulations. Through simulations, we study the behavior of methods in the following settings,

1. **Fixed distributions setting:** This is when the data distributions do not change over time. In this setting, we expect the methods will perform well even if they use all the observed samples. This is because there is no change in the distributions. The results of simulations in this setting are shown in figure 5(a). We can see that the results are consistent with our expectations.
2. **Effect of window size:** This parameter controls how many samples the system would look back. It is important for the settings with distributions shift. We study the role of different window sizes in the shift and no-shift scenarios. To simulate distribution change we change the OOD distribution to $\mathcal{N}_{ood}(\mu = -5, \sigma = 4)$ at time $t = 55k$. The results with various window sizes with a shift in OOD distribution is shown in figure 6. In the shifted settings, we observe that using a sufficiently small window size is crucial because the system can quickly react to the shift.
3. **Effect of γ:** We show that changing the mixing ratio γ of ID and OOD samples does not affect the control of FPR. We show two settings for changing γ. First, the gamma changes from 0.1 to 1 when $t = 55k$. Second, we show γ gradually changes from 0.1 to 1 starting from $t = 20k$ and ends at $t = 80k$. Specifically, γ increases 0.1 for every $6k$ sample the system receives. We see that our system is able to control FPR with the change of γ. The results are shown in figure 7 and they match with what we expected.
4. **In-Distribution shift :** We show that our system is able to control FPR when the ID distribution is shifted. We simulate the OOD and ID scores using a mixture of two Gaussians $\mathcal{N}_{id}(\mu = 5.5, \sigma = 4)$ and $\mathcal{N}_{ood}(\mu = -5, \sigma = 4)$ with $\gamma = 0.1$. To simulate distribution change we change the ID



(a) No window, no distribution shift.

(b) 10K window, no distribution shift

(c) No window, distribution shift at $t = 55k$.

(d) 10K window, distribution shift at $t = 55k$.

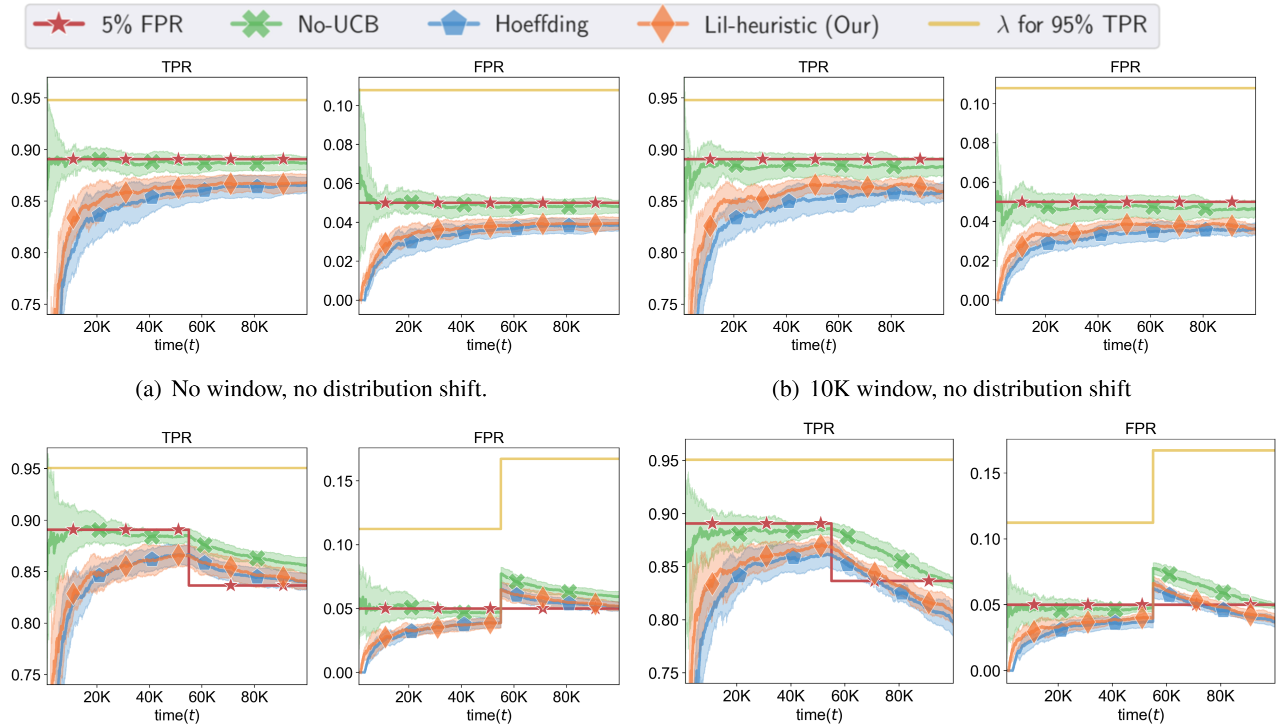
Figure 4: Experiments on Synthetic Data. The ratio of ID: OOD = 2:1. Each method is repeated 10 times. The mean and standard deviation are shown. The distribution shift starts at $t = 55k$ for figure 5(c),5(d).

distribution to $\mathcal{N}_{id}(\mu = 5, \sigma = 4)$ at time $t = 55k$. The results are shown in figure 8. We can clearly see that changing ID distribution(ID scores getting closer to the OOD scores) leads to a decrease in the TPR at the threshold with 5% FPR. Since the estimation of threshold only depends on the FPR estimates and hence only on OOD samples, changing ID distribution does not affect this estimation so the methods perform the same as in the setting of no-distribution shift but get a reduction in the TPR at FPR 5%.

Next, we present our results on real ID and OOD datasets. In the synthetic setup, we directly simulated the scores and ran experiments on those. Here we obtain scores using some of the highly effective scoring functions for OOD detection. We first provide details of ID and OOD datasets and then discuss the scoring functions we used.

7.3 Real data experiments:

We evaluate our proposed system empirically on two sets of benchmarks from OpenOOD Benchmark Yang et al. (2022). Here we show the results on CIFAR-10 ID dataset and show the results on CIFAR-100, and Imagenet-1K Deng et al. (2009) ID dataset in the appendix. CIFAR-10Krizhevsky et al. (2009). CIFAR-10 is a 10-class dataset for general object classification. We use the official testing datasets as the ID dataset. The OOD datasets for CIFAR-10 consist of CIFAR100, SVHN Netzer



(a) No window, no distribution shift.

(b) 10K window, no distribution shift

(c) No window, distribution shift at $t = 55k$.

(d) 10K window, distribution shift at $t = 55k$.

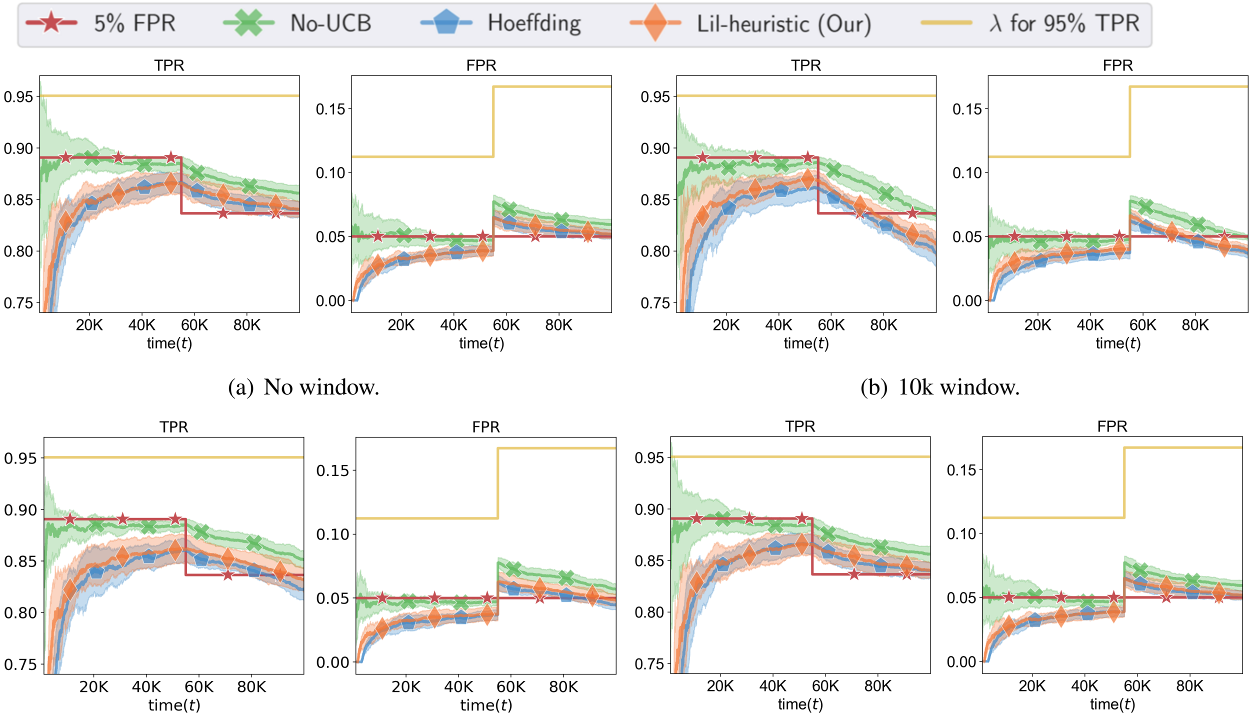
Figure 5: Experiments on Synthetic Data. The ratio of ID: OOD = 4:1. Each method is repeated 10 times. The mean and standard deviation are shown. The distribution shift starts at $t = 55k$.

et al. (2011), TinyImageNet Krizhevsky et al. (2017) (1,207 images are removed from TinyImageNet since they belong to CIFAR-10Yang et al. (2021a)), MNIST Deng (2012), Texture Kylberg (2011), Places365 Zhou et al. (2017) (1,305 images are removed due to semantic overlaps). We use a pre-train ResNet-18 with 94.3% accuracy throughout all the experiments on CIFAR-10.

Data Stream: For the non-distribution shifted setting, we combine all six OOD datasets for a joint OOD distribution. For shifted distribution setting, we sample a portion of OOD samples from three OOD datasets and then sample the shifted distribution samples from the rest of the OOD datasets. We randomly sample 90k OOD samples and 9k ID samples.

Computing OOD Scores: Accurately detecting OOD points in the online setting needs a good scoring function that separates the ID and OOD points at some threshold score λ. We leverage existing works on the construction of the scoring function. We use ODIN Liang et al. (2017), Mahalanobis Distance Lee et al. (2018), Energy Score Liu et al. (2020), SSD Sehwag et al. (2021), VIM Wang et al. (2022), and KNN Sun et al. (2022) scores for the evaluation. We use an open-source codebase, OpenOOD Yang et al. (2022), to implement all the methods. Due to space limitation, we present results for KNN Sun et al. (2022) score here. For more details on these scores and results on the rest of the scores please see the Appendix.

Discussion: As expected, in the fixed distributions setting in both synthetic and real data settings (figures 5(a), and 10(a), respectively), we see that not using a UCB leads to violation of FPR



(a) No window.

(b) 10k window.

(c) 15k window.

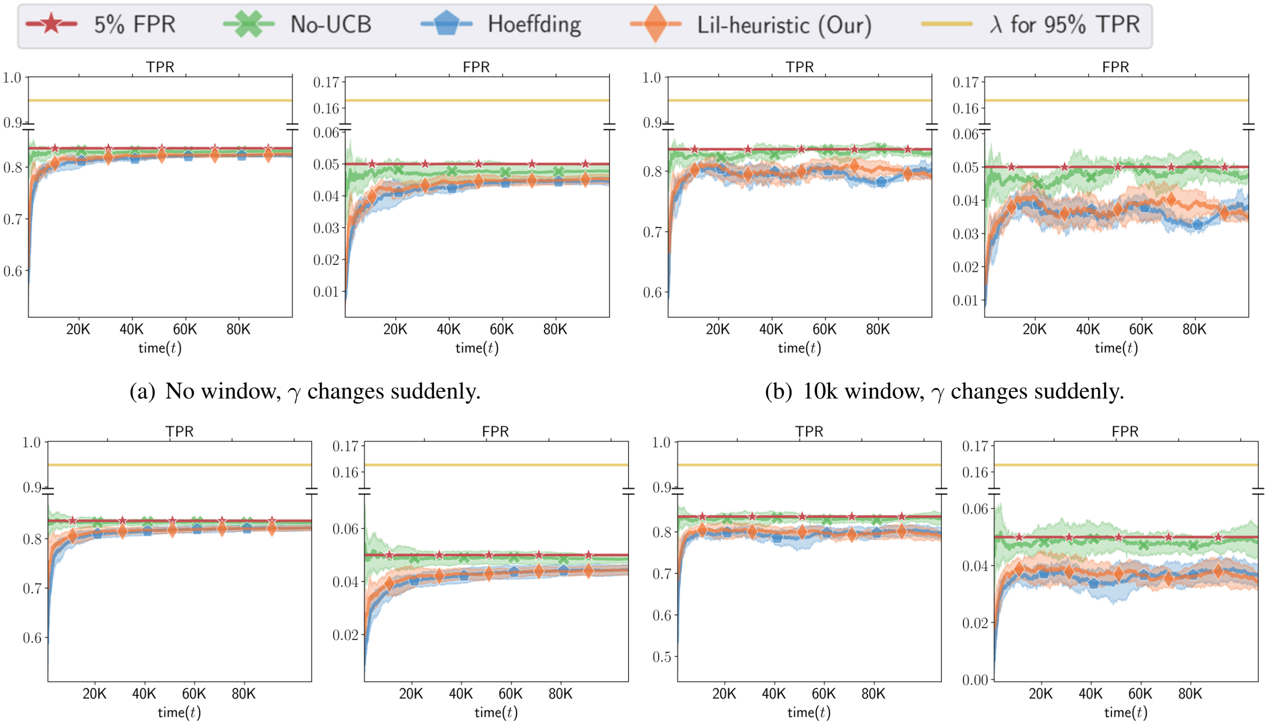
(d) 20k window.

Figure 6: Effect of using various window sizes in synthetic data experiments.

constraints and the methods with LIL-Heuristic, Hoeffding based intervals are able to maintain the FPR below the user given threshold 5%. Moreover, all the methods improve as they acquire more samples with time and eventually reach very close to the optimal solution. When we run these methods with a window size of 10k in the fixed distribution setting (figures 6(b), 10(b)) we observe similar behavior except with a bit more variance since with a fixed window the confidence intervals are not shrinking with time. Though the windowed setting is more useful when the distributions change and not so much of use in the fixed distribution case, nevertheless we show this experiment to validate our understanding of the fixed distribution setting.

Moving forward, we investigate the case where the distributions change at a specific time point. In such scenarios, we find that the windowed approach adapts more rapidly compared to the method without a window (see figures 5(c),5(d)). Using a fixed window allows the algorithm to quickly adjust to the changed distribution, whereas without a window, the adaptation process is significantly delayed.

In summary, our findings demonstrate that the choice of using a windowed approach or not depends on the nature of the data and the presence of distribution changes. The windowed approach proves advantageous in scenarios where rapid adaptation is crucial, while the non-windowed approach can still be effective, albeit potentially with longer adaptation times. The consistency between our observations in the synthetic experiments and the real-data evaluations provides strong evidence of the reliability and effectiveness of our proposed methods. These findings demonstrate the robustness of our approaches and their applicability to various practical scenarios.



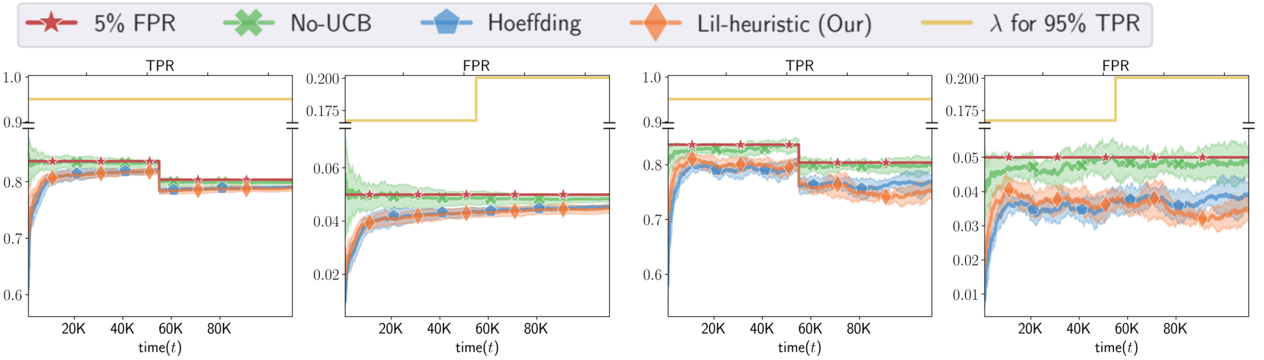
(a) No window, γ changes suddenly.

(b) 10k window, γ changes suddenly.

(c) No window, γ changes gradually.

(d) 10k window, γ changes gradually.

Figure 7: Changing the mixing ratio γ in the synthetic data.



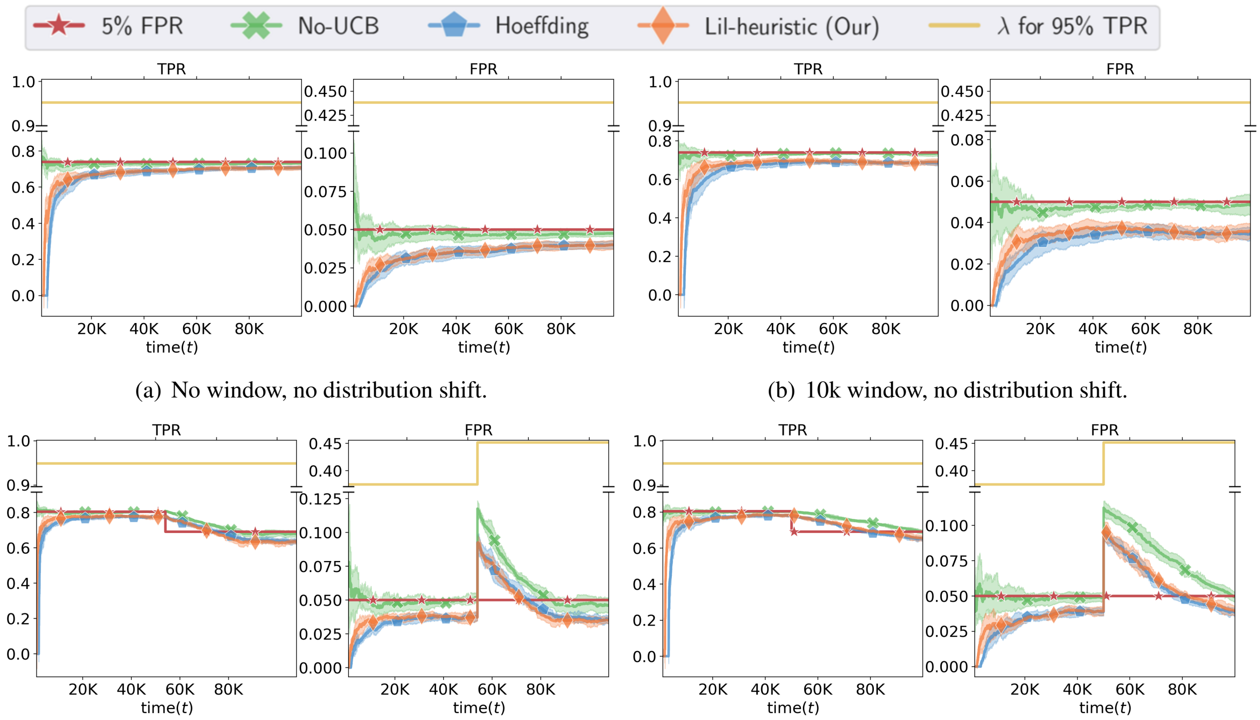
(a) No window.

(b) 10k window.

Figure 8: Changing ID distribution in synthetic data.

8. Additional Experiments and Details

In the simulations we study the effect of changing γ, using different window sizes and the case when the In-Distribution shifts. For the real data experiments, we study the performance of the methods under different settings with different scoring functions on CIFAR-10 and CIFAR-100 as In-Distribution datasets.



(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

(c) ID:OOD = 2:1 w. distribution shift & 10k window.

(d) ID:OOD = 4:1 w. distribution shift & 10k window.

Figure 9: Results on the KNN scores with Cifar-10 as ID dataset.

8.1 Additional Real OOD datasets Experiments

We run our proposed system on different OOD scoring methods. We use $\alpha = 0.05$, $\delta = 0.2$, and importance sampling probability $p = 0.2$ through all the experiments.

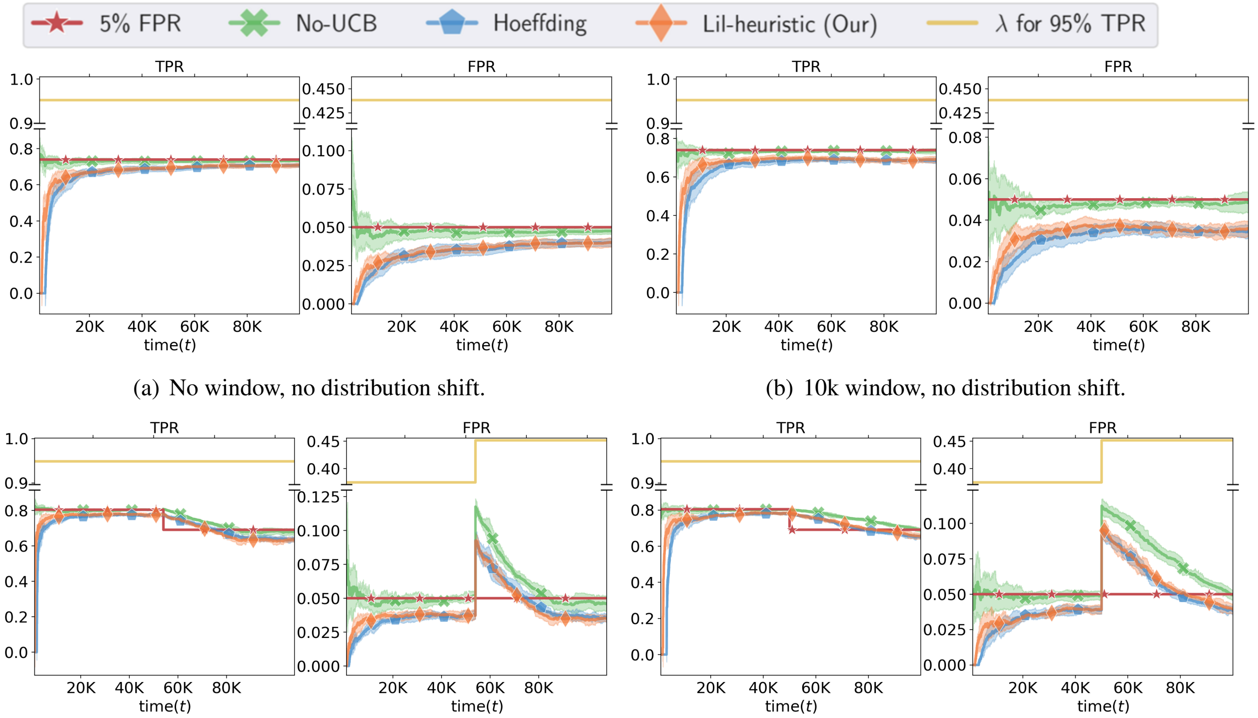
CIFAR-10 and CIFAR-100. We use CIFAR-10 or CIFAR-100 as ID datasets. We run the experiment with different window sizes and distribution shifts. In the distribution shift setting, if not specified, we use MNIST, SVHN, and Texture as the first mixture of OOD datasets, and TinyImageNet, Places365, and CIFAR-10/100 as the second mixture of OOD datasets by default. We use a pre-trained Resnet-50 model for SSD method, Resnet-34, for iDECODE method, and Resnet-18 for the rest of the methods.

Imagenet-1K. Additionally, We use Imagenet-1K Deng et al. (2009) as the ID datasets and SSB-hard Vaze et al. (2021), NINCO Bitterwolf et al. (2023), iNaturalist Huang and Li (2021), Texture, and OpenImage-O Hendrycks et al. (2021) as the OOD datasets. We use ResNet-50 for ASHDjurisic et al. (2022), GradNorm Huang et al. (2021), and, ReAct Sun et al. (2021) methods. We also apply vision transformers (SWIN-T Liu et al. (2021)) for ReAct. Each experiment is conducted in shifted and non-shifted cases. For shifted cases, we shift the OOD datasets from Far OOD (iNaturalist, Texture, and OpenImage-O) to Near OOD (SSB-hard, NINCO).

1. **ODIN**: ODIN Liang et al. (2017) takes the soft-max score from DNNs, and scales the score with temperature. A gradient-based input perturbation is also used for better performance. We choose temperature 1000 and input perturbation noise 0.0014, as discussed in Liang et al. (2017). Please see figures 20 and 21 for the results with this score.

2. **Mahalanobis Distance:** For a given point x, the Mahalanobis Distance (MDS) based score is its MD from the closest class conditional distribution. We use the MD-based score as given in Lee et al. (2018) for detecting OOD and adversarial samples. They compute the scores using representations from various layers of DNNs and combine them to get a better scoring function. We choose input perturbation noise to be 0.0014. Please see figures 14 and 15 for the results with this score.
3. **Energy Score:** This score was proposed in Liu et al. (2020) and it is well aligned with the probability density of the samples, with low energy implying ID and high energy implying OOD. We choose the temperature parameter to be 1. Please see figures 12 and 13 for the results with this score.
4. **SSD**. It is based on computing the Mahalanobis distance in the feature space of the model trained on the unlabeled in-distribution data using self-supervised learning. We use the official implementation of Sehwag et al. (2021). For CIFAR-10, we use the pre-train model they released. For CIFAR-100, We train a Resnet-50 using a contrastive self-supervised learning loss, SimCLR Chen et al. (2020). When calculating the distance-based OOD scores, we use one unsupervised clustering center as the only center for ID distribution for both CIFAR-10 and CIFAR-100. Please see figures 18 and 19 for the results with this score.
5. **Virtual-logit Match**. Virtual-logit Match (VIM) Wang et al. (2022) combines the class-agnostic score from feature space and ID class-dependent logits. Specifically, an additional logit representing the virtual OOD class is generated from the residual of the feature against the principal space and then matched with the original logits by a constant scaling. We set the dimension of the principal space $D = 100$. Please see figures 16 and 17 for the results with this score.
6. **K-Nearest-Neighborhood**. Unlike other methods that impose a strong distributional assumption of the underlying feature space, the KNN-based method Sun et al. (2022) explores the efficacy of non-parametric nearest-neighbor distance for OOD detection. The distance between the test sample and its k-nearest training IN sample will be used as the score for a threshold based OOD detection. We choose neighbor number $k = 50$. Please see figures 10 and 11 for the results with KNN scores.
7. **Activation Shaping**. ASHDjurisic et al. (2022). extremely simple post-hoc activation shaping method, ASH, where a large portion of a sample's activation at a late layer is removed, and the rest simplified or lightly adjusted. The shaping is applied at inference time, and does not require any statistics calculated from training data.
8. **GradNorm**. GradNorm Huang et al. (2021) is a simple and effective approach for detecting OOD inputs by utilizing information extracted from the gradient space. GradNorm directly employs the vector norm of gradients, backpropagated from the KL divergence between the softmax output and a uniform probability distribution. The key idea is that the magnitude of gradients is higher for in-distribution (ID) data than for OOD data, making it informative for OOD detection.
9. **Rectified Activations**. ReActSun et al. (2021) is a simple and effective technique for reducing model overconfidence in OOD data. ReAct is motivated by a novel analysis of internal activations of neural networks, which displays highly distinctive signature patterns for OOD distributions.ReAct can generalize effectively to different network architectures and different OOD detection scores.
10. **iDECODE**. iDECODe Kaur et al. (2022) leverages in-distribution equivariance for conformal OOD detection. It relies on a novel base non-conformity measure and a new aggregation method,

used in the inductive conformal anomaly detection framework, thereby guaranteeing a bounded false detection rate.



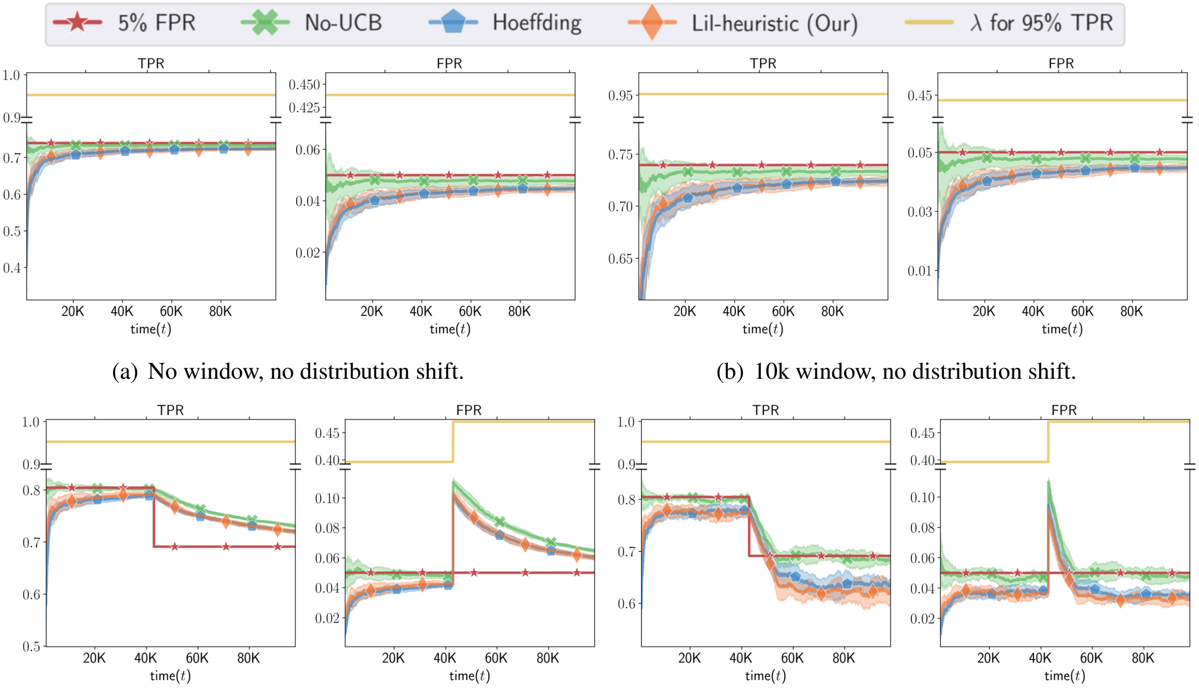
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

(c) ID:OOD = 2:1 w. distribution shift & 10k window.

(d) ID:OOD = 4:1 w. distribution shift & 10k window.

Figure 10: Results on the KNN scores with Cifar-10 as ID dataset.



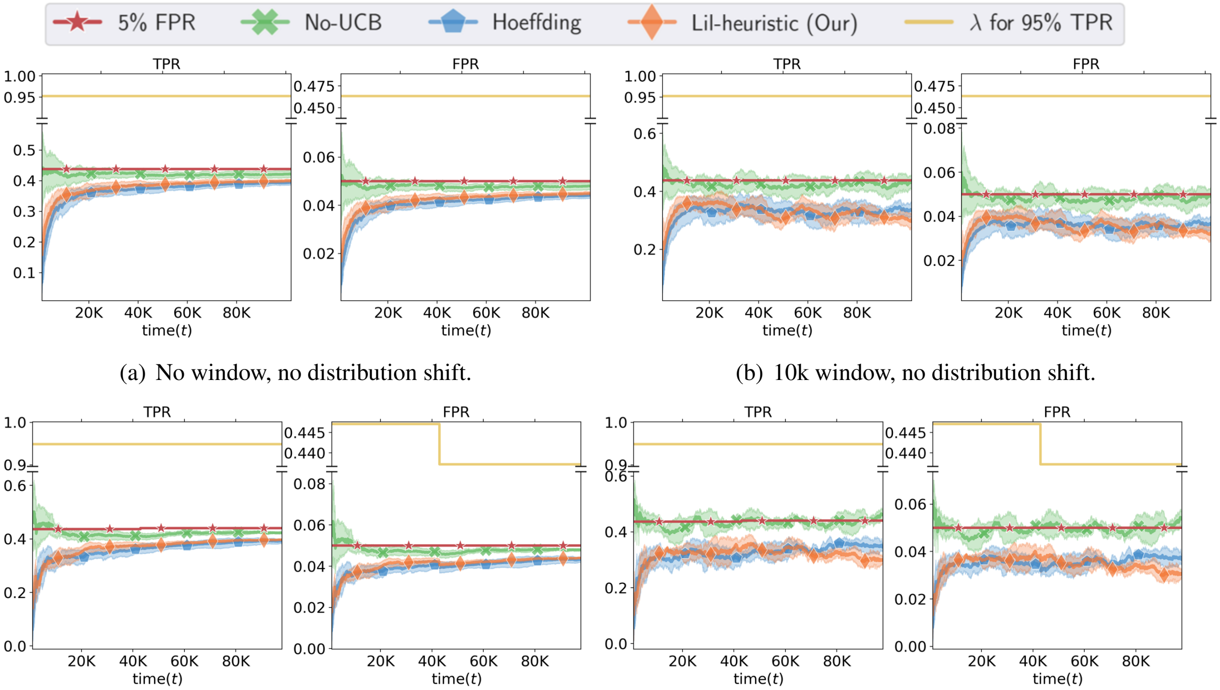
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

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

(d) 10k window, distribution shift at $t = 45k$.

Figure 11: Results on the KNN scores with Cifar-100 as the ID dataset.



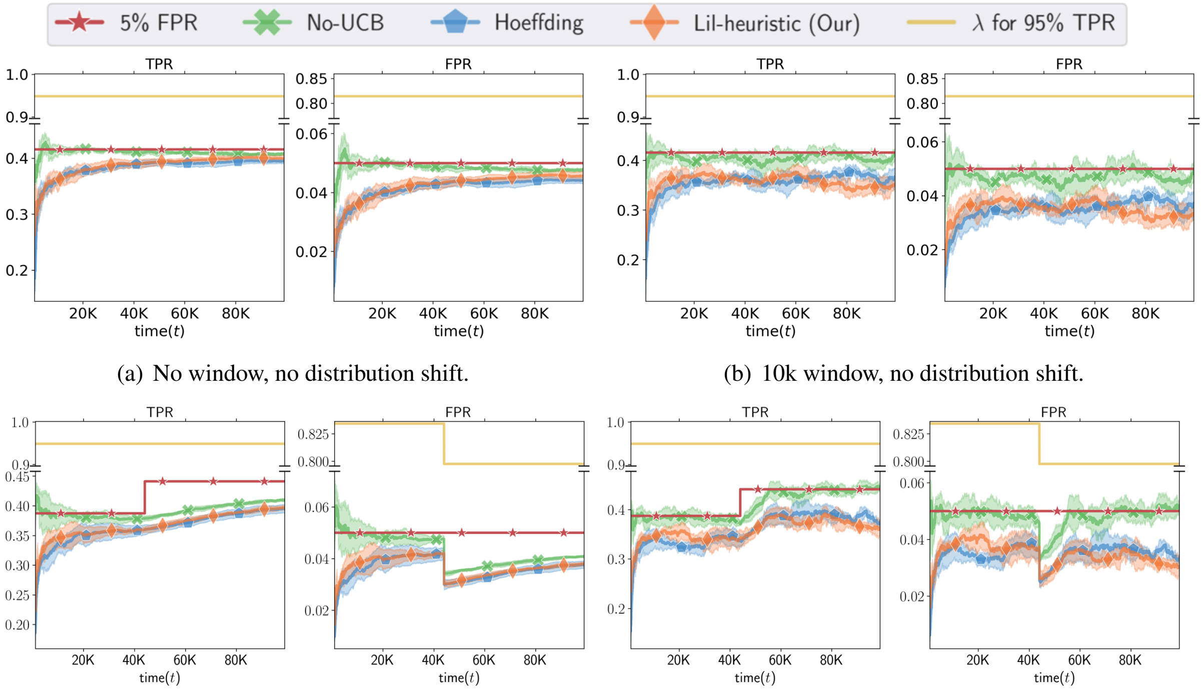
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

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

(d) 10k window, distribution shift at $t = 45k$.

Figure 12: Results on the Energy Based Score (EBO) method with Cifar-10 as the ID dataset.



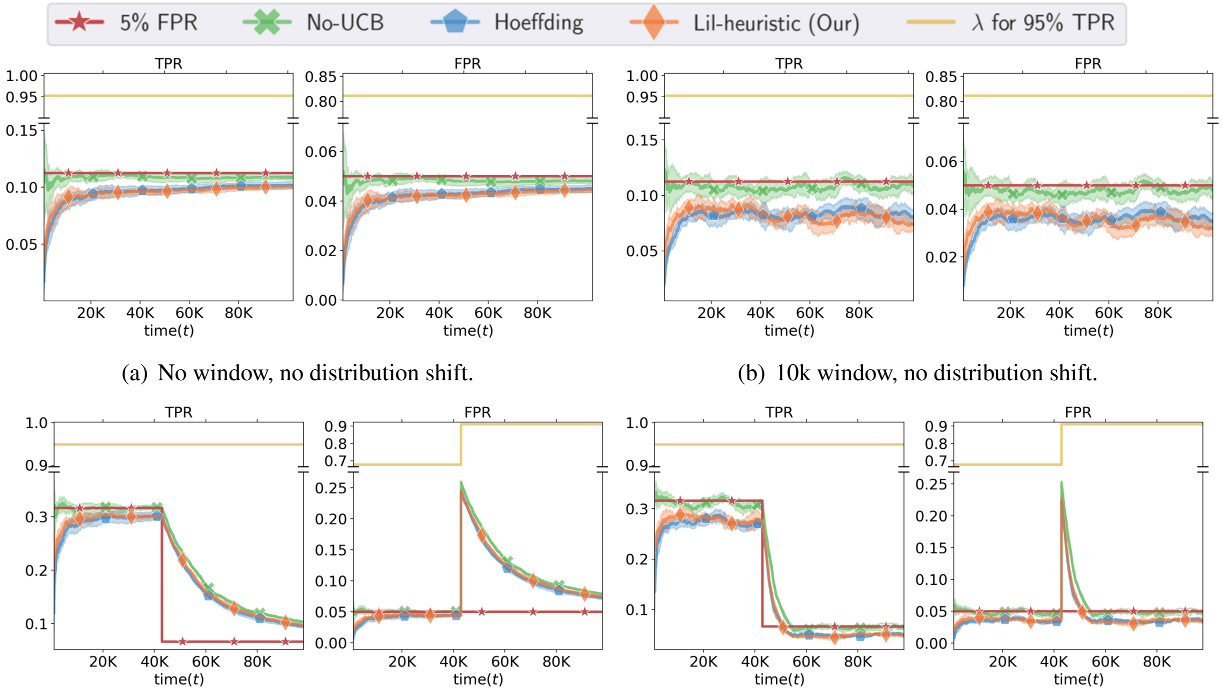
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

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

(d) 10k window, distribution shift at $t = 45k$.

Figure 13: Results on the EBO scores with Cifar-100 as the ID dataset.



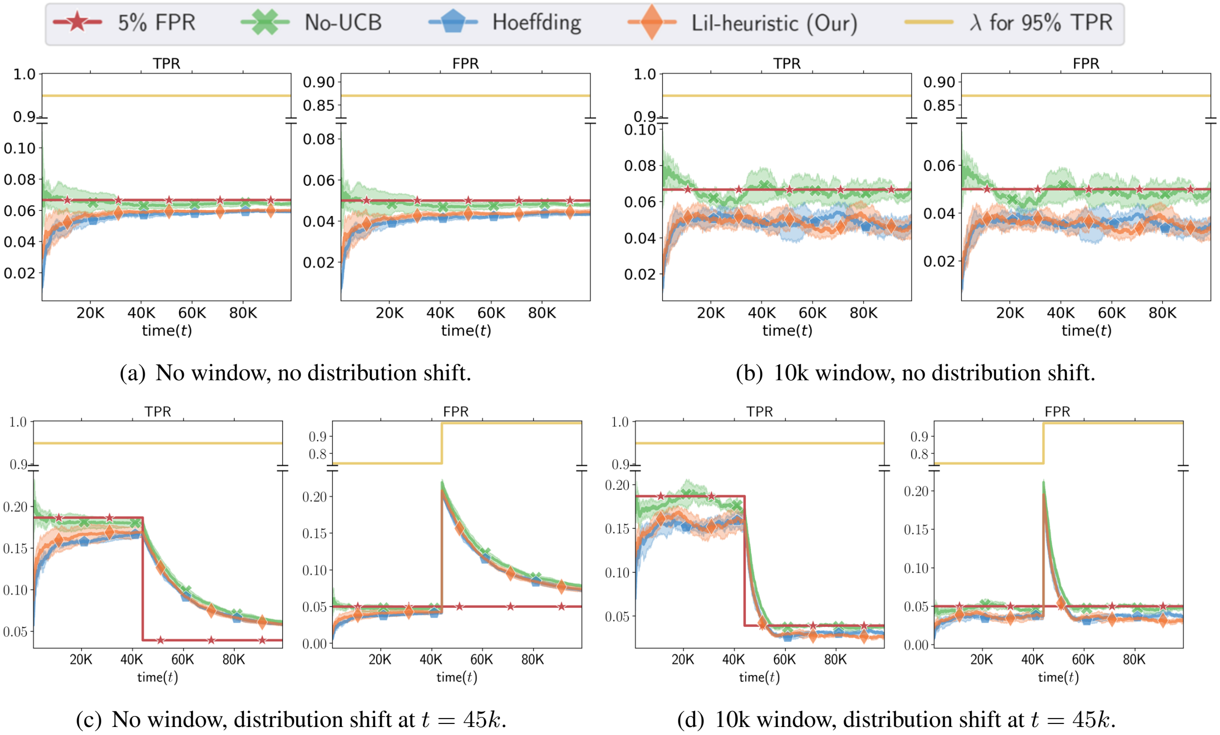
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

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

(d) 10k window, distribution shift at $t = 45k$.

Figure 14: Results on the Mahalanobis distance (MDS) method with Cifar-10 as the ID dataset.



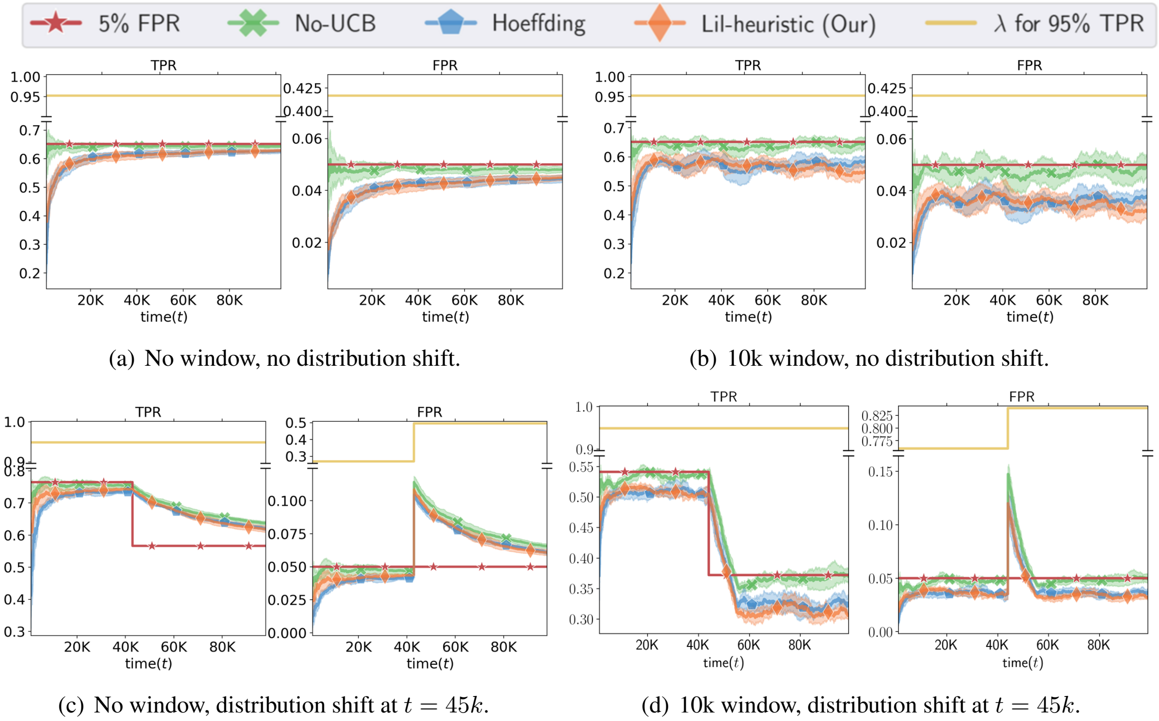
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

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

(d) 10k window, distribution shift at $t = 45k$.

Figure 15: Results on the MDS scores with Cifar-100 as the ID dataset.



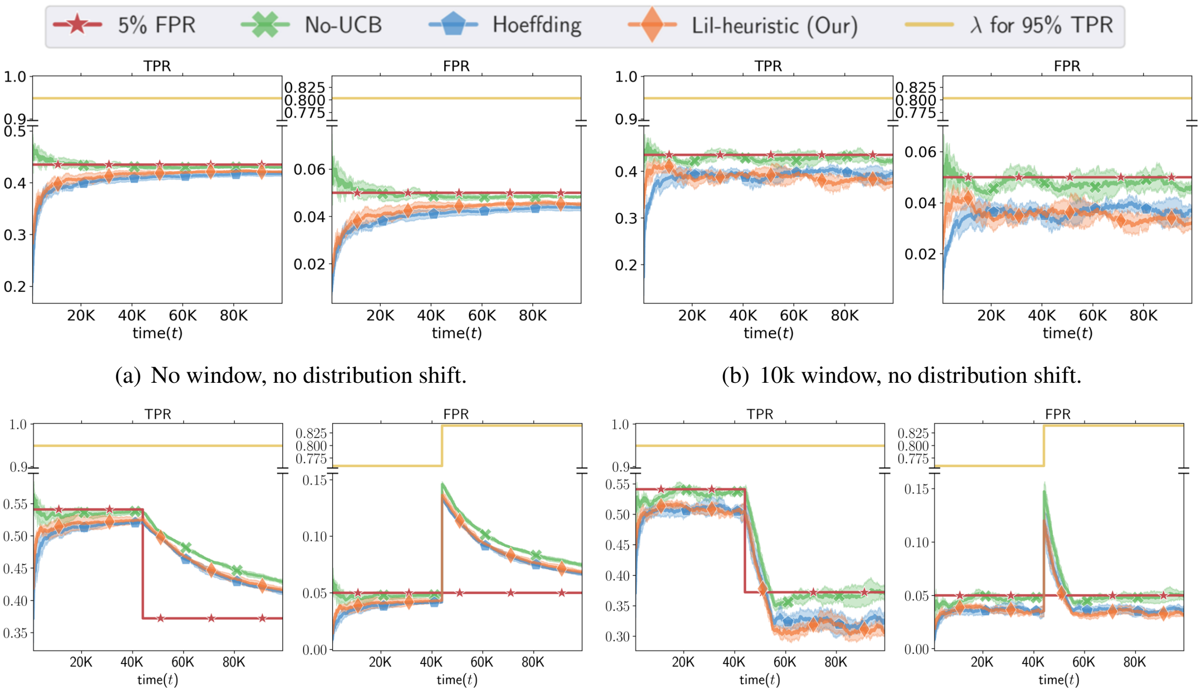
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

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

(d) 10k window, distribution shift at $t = 45k$.

Figure 16: Results on the Virtual-logit Match (VIM) method with Cifar-10 as the ID dataset.



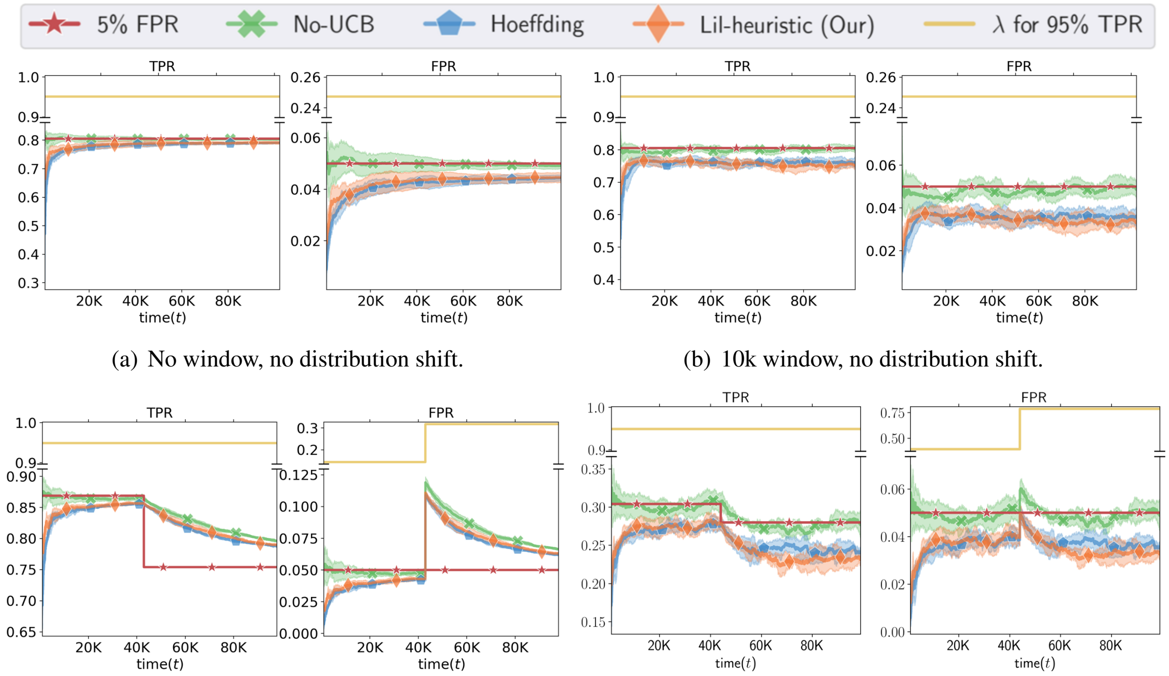
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

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

(d) 10k window, distribution shift at $t = 45k$.

Figure 17: Results on the VIM scores with Cifar-100 as the ID dataset.



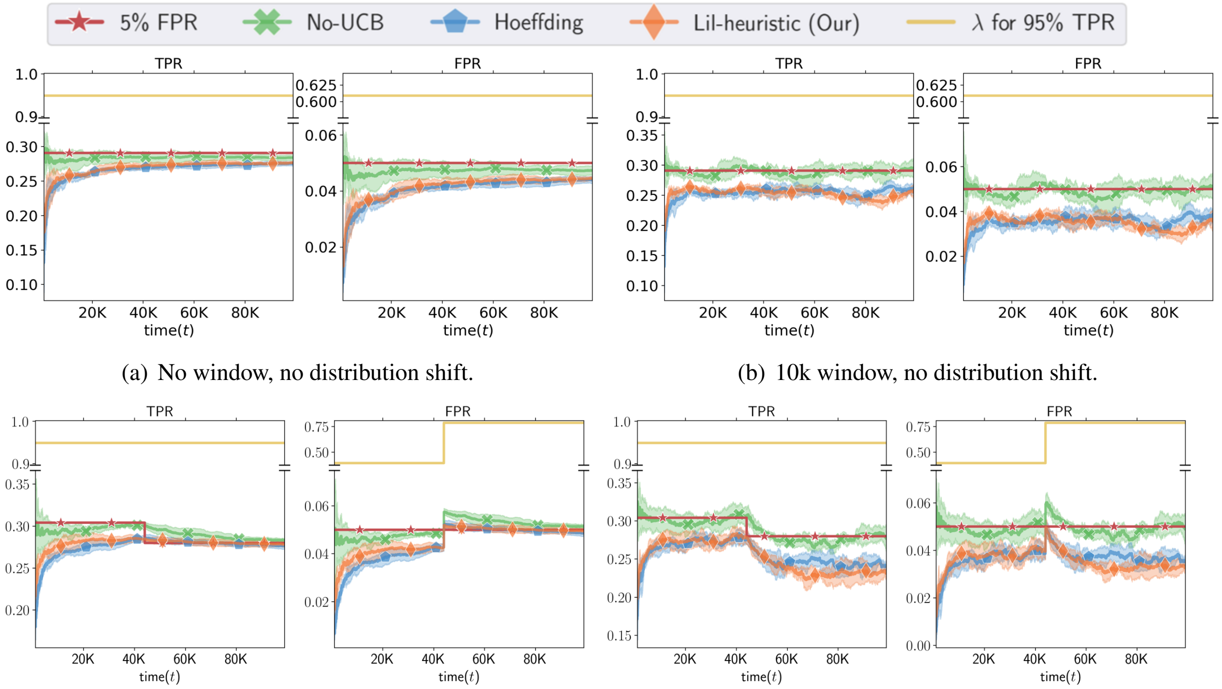
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

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

(d) 10k window, distribution shift at $t = 45k$.

Figure 18: Results on the SSD method with Cifar-10 as the ID dataset.



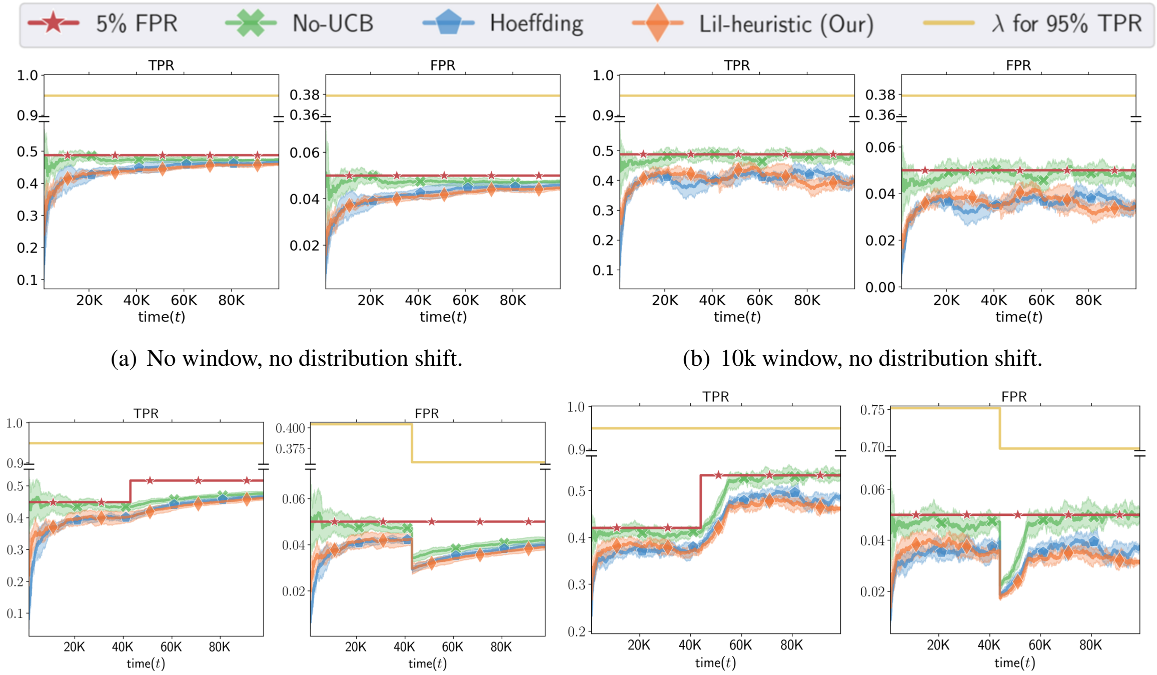
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

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

(d) 10k window, distribution shift at $t = 45k$.

Figure 19: Results on the SSD scores with Cifar-100 as the ID dataset.



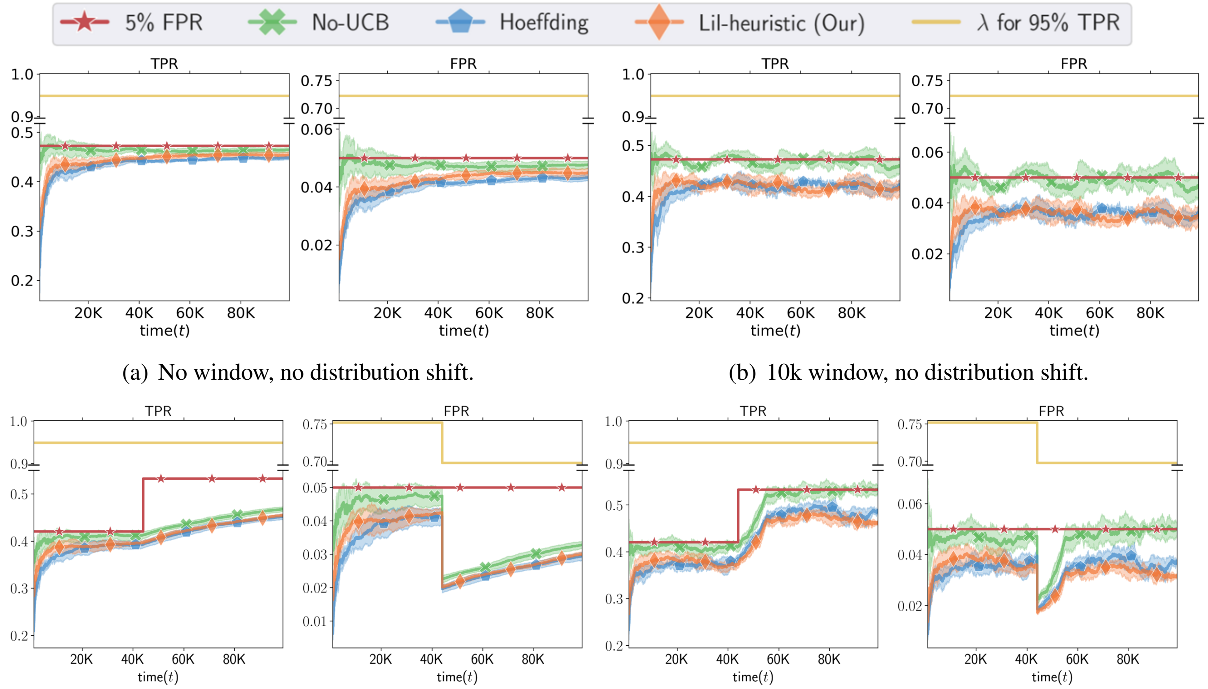
(a) No window, no distribution shift.

(b) 10k window, no distribution shift.

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

(d) 10k window, distribution shift at $t = 45k$.

Figure 20: Results on the ODIN method with Cifar-10 as the ID dataset.



(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 $t = 45k$.

Figure 21: Results on the ODIN scores with Cifar-10 as the ID dataset.

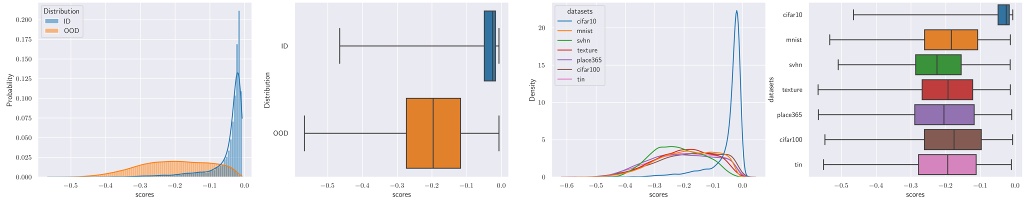


Figure 22: Scores distribution for KNN with CIFAR-10 as In-Distribution.

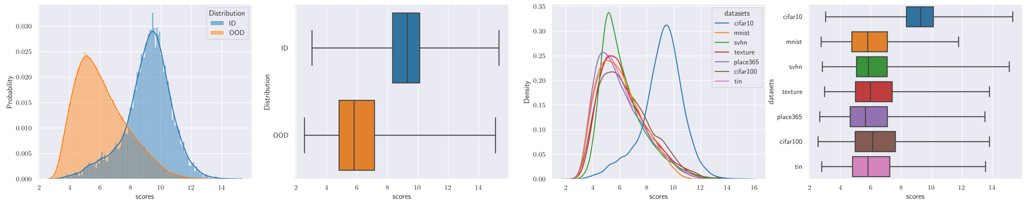


Figure 23: Scores distribution for EBO with CIFAR-10 as In-Distribution.

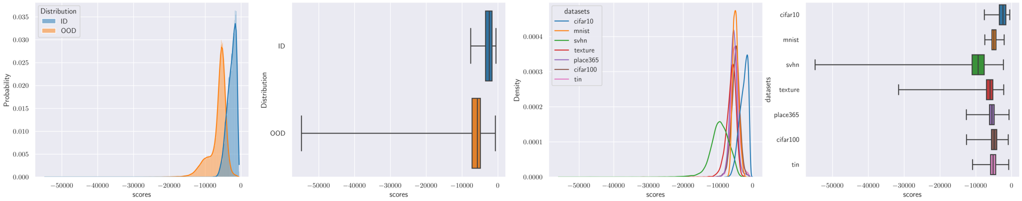


Figure 24: Scores distribution for SSD with CIFAR-10 as In-Distribution.

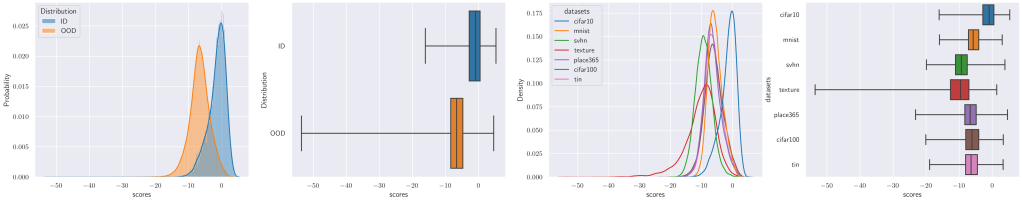


Figure 25: Scores distribution for VIM with CIFAR-10 as In-Distribution.

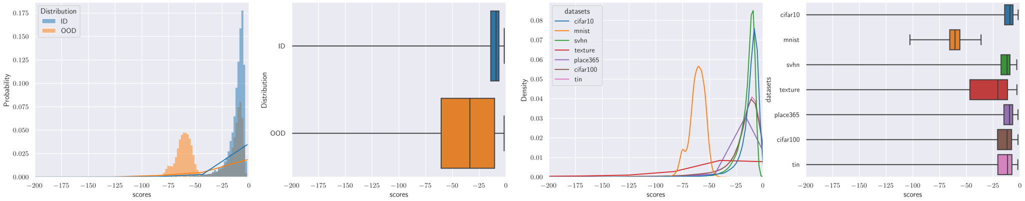


Figure 26: Scores distribution for MDS with CIFAR-10 as In-Distribution.

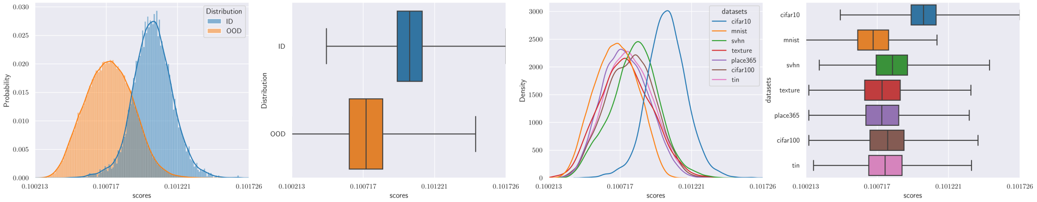


Figure 27: Scores distribution for ODIN with CIFAR-10 as In-Distribution.

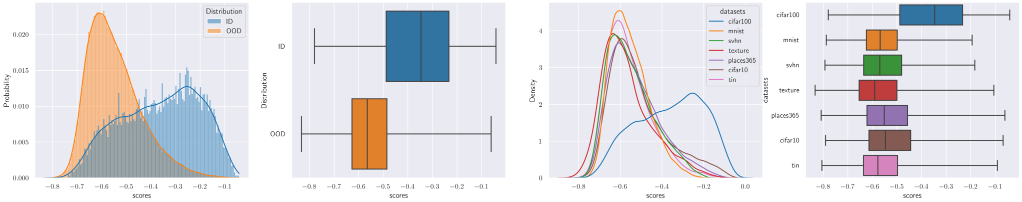


Figure 28: Scores distribution for KNN with cifar-100 as In-Distribution.

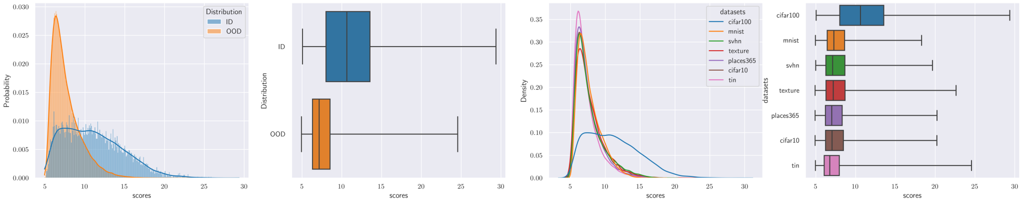


Figure 29: Scores distribution for EBO with cifar-100 as In-Distribution.

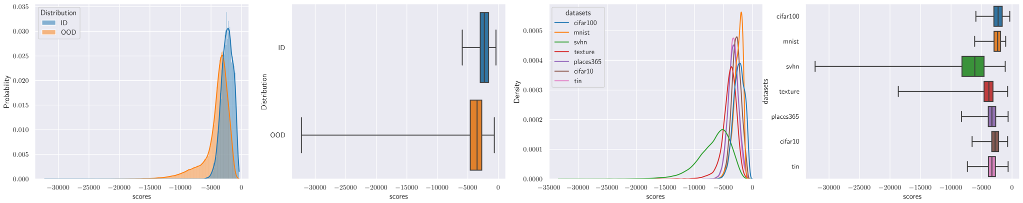


Figure 30: Scores distribution for SSD with cifar-100 as In-Distribution.

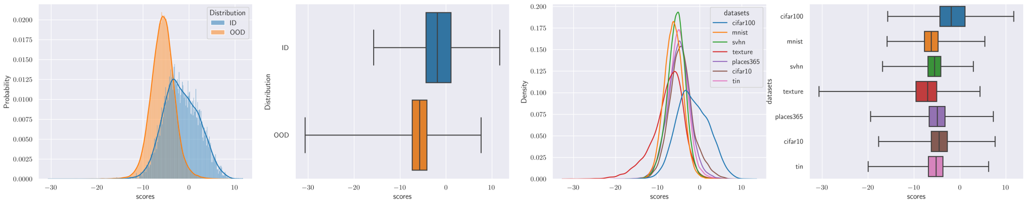


Figure 31: Scores distribution for VIM with cifar-100 as In-Distribution.

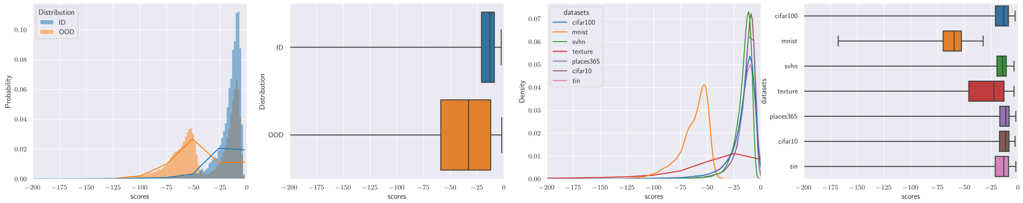


Figure 32: Scores distribution for MDS with cifar-100 as In-Distribution.

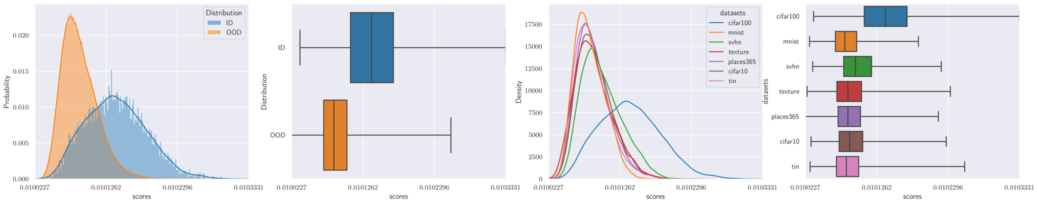


Figure 33: Scores distribution for ODIN with cifar-100 as In-Distribution.