In- or Out-of-Distribution Detection via Dual Divergence Estimation (Supplementary Material)

Sahil Garg[1,*] **Sanghamitra Dutta**[2] **Mina Dalirrooyfard**[1] **Anderson Schneider**[1] **Yuriy Nevmyvaka**[1]

[1]Dept. of Machine Learning Research, Morgan Stanley, New York, New York, USA
[2]Dept. of Electrical and Computer Engineering, University of Maryland, College Park, Maryland, USA
[*]Corresponding Author: sahil.garg@morganstanley.com, sahil.garg.cs@gmail.com

1 PROOFS

We first recall the definition of $\hat{D}(\mathbf{X}\|\mathbf{X}^{in})$ here.

$$\hat{D}(\mathbf{X}\|\mathbf{X}^{in}) = \max_{\hat{f}(.)\in\mathcal{H}} \frac{1}{m}\sum_{\mathbf{x}_j\in\mathbf{X}} \hat{f}(\mathbf{x}_j) - \log \sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}(\mathbf{x}_i^{in})} + \log N \tag{1}$$

We define the value inside the max expression for any particular f as $\hat{D}_f(\mathbf{X}\|\mathbf{X}^{in})$.

$$\hat{D}_{\hat{f}}(\mathbf{X}\|\mathbf{X}^{in}) = \frac{1}{m}\sum_{\mathbf{x}_j\in\mathbf{X}} \hat{f}(\mathbf{x}_j) - \log \sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}(\mathbf{x}_i^{in})} + \log N \tag{2}$$

1.1

Proof of Theorem 1.

Let $m = \min_{\mathbf{x}_i^{in}} \hat{f}^*(\mathbf{x}_i^{in})$. By way of contradiction assume that there is $\mathbf{x}_j$ such that $\hat{f}^*(\mathbf{x}_j) < m$. Define function $\bar{f}$ as follows. For any $\mathbf{x}_i \in \mathbf{X}$,

$$\bar{f}(\mathbf{x}_i) = \begin{cases} \hat{f}^*(\mathbf{x}_i) & \text{if } f(\mathbf{x}_i) \geq m \\ (\hat{f}^*(\mathbf{x}_i) + m)/2 & \text{otherwise} \end{cases}$$

We show that $\hat{D}_{\bar{f}}(\mathbf{X}\|\mathbf{X}^{in}) > \hat{D}_{\hat{f}^*}(\mathbf{X}\|\mathbf{X}^{in})$. To see this, first note that for any $\mathbf{x}_i^{in} \in \mathbf{X}^{in}$, $\bar{f}(\mathbf{x}_i^{in}) = \hat{f}^*(\mathbf{x}_i^{in})$. So we have

$$\log \sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} \frac{e^{\bar{f}(\mathbf{x}_i^{in})}}{|\mathbf{X}^{in}|} = \log \sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} \frac{e^{\hat{f}^*(\mathbf{x}_i^{in})}}{|\mathbf{X}^{in}|}. \tag{3}$$

Moreover, we have that for any $\mathbf{x}_i \in \mathbf{X}$, $\bar{f}(\mathbf{x}_i) \geq \hat{f}^*(\mathbf{x}_i)$. This is because if $\hat{f}^*(\mathbf{x}_i) \geq m$, then $\bar{f}(\mathbf{x}_i) = \hat{f}^*(\mathbf{x}_i)$ and if $\hat{f}^*(\mathbf{x}_i) < m$, then $\bar{f}(\mathbf{x}_i) = (m + \hat{f}^*(\mathbf{x}_i))/2 > \hat{f}^*(\mathbf{x}_i)$. Since there is at least one $\mathbf{x}_j$ such that $\hat{f}^*(\mathbf{x}_j) < m$, we have that

$$\sum_{\mathbf{x}_j\in\mathbf{X}} \frac{\bar{f}(\mathbf{x}_j)}{|\mathbf{X}|} > \sum_{\mathbf{x}_j\in\mathbf{X}} \frac{\hat{f}^*(\mathbf{x}_j)}{|\mathbf{X}|}. \tag{4}$$

By Eq 3 and 4 and the definition of $\hat{D}_f(\mathbf{X}\|\mathbf{X}^{in})$, we have that $\hat{D}_{\bar{f}}(\mathbf{X}\|\mathbf{X}^{in}) > \hat{D}_{\hat{f}^*}(\mathbf{X}\|\mathbf{X}^{in})$. Since $\hat{D}_{\hat{f}^*}(\mathbf{X}\|\mathbf{X}^{in}) = \hat{D}(\mathbf{X}\|\mathbf{X}^{in}) = \max \hat{D}_f(\mathbf{X}\|\mathbf{X}^{in})$, this is a contradiction. So for all $\mathbf{x}_j \in \mathbf{X}$, we have $\hat{f}^*(\mathbf{x}_j) \geq m$. □

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

1.2

Proof of Theorem 2. By way of contradiction assume that $\max_{\mathbf{x}_j \in \mathbf{X}} \hat{f}^*(\mathbf{x}_j) < \max_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} \hat{f}^*(\mathbf{x}_i^{in})$. Let $m = \max_{\mathbf{x}_j \in \mathbf{X}} \hat{f}^*(\mathbf{x}_j)$. Define $f_0(x) = \min(\hat{f}^*(x), m)$. Thus, $f_0(x) \leq \hat{f}^*(x)$ for all x with strict inequality $f_0(x) < \hat{f}^*(x)$ for at least some $\mathbf{x}_i^{in} \in \mathbf{X}^{in}$.

We will now that show that $\hat{D}_{f_0}(\mathbf{X} \| \mathbf{X}^{in}) > \hat{D}_{\hat{f}^*}(\mathbf{X} \| \mathbf{X}^{in})$. To see this, note that:

$$\log \sum_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} \frac{e^{f_0(\mathbf{x}_i^{in})}}{|\mathbf{X}^{in}|} < \log \sum_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} \frac{e^{\hat{f}^*(\mathbf{x}_i^{in})}}{|\mathbf{X}^{in}|}. \tag{5}$$

This leads to:

$$\begin{aligned}
\hat{D}_{f_0}(\mathbf{X} \| \mathbf{X}^{in}) &= \sum_{\mathbf{x}_j \in \mathbf{X}} \frac{f_0(\mathbf{x}_j)}{|\mathbf{X}|} - \log \sum_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} \frac{e^{f_0(\mathbf{x}_i^{in})}}{|\mathbf{X}^{in}|} \\
&> \sum_{\mathbf{x}_j \in \mathbf{X}} \frac{f_0(\mathbf{x}_j)}{|\mathbf{X}|} - \log \sum_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} \frac{e^{\hat{f}^*(\mathbf{x}_i^{in})}}{|\mathbf{X}^{in}|} \\
&= \sum_{\mathbf{x}_j \in \mathbf{X}} \frac{\hat{f}^*(\mathbf{x}_j)}{|\mathbf{X}|} - \log \sum_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} \frac{e^{\hat{f}^*(\mathbf{x}_i^{in})}}{|\mathbf{X}^{in}|} \text{ since } \hat{f}^*(\mathbf{x}_j) \leq m \text{ for all } \mathbf{x}_j \in \mathbf{X} \text{ making } \hat{f}^*(\mathbf{x}_j) = f_0(\mathbf{x}_j) \\
&= \hat{D}_{\hat{f}^*}(\mathbf{X} \| \mathbf{X}^{in}).
\end{aligned} \tag{6}$$

This is a contradiction since $\hat{D}_{\hat{f}^*}(\mathbf{X} \| \mathbf{X}^{in}) = \hat{D}(\mathbf{X} \| \mathbf{X}^{in}) = \max_{f \in \mathcal{H}} \hat{D}_f(\mathbf{X} \| \mathbf{X}^{in})$. So, we have $\max_{\mathbf{x}_j \in \mathbf{X}} \hat{f}^*(\mathbf{x}_j) \geq \max_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} \hat{f}^*(\mathbf{x}_i^{in})$. □

1.3

Proof of Theorem 3. By the definition of $\mathbf{X}^{ood}$, since for any $\mathbf{x}_j \in \mathbf{X}^{ood}$ we have $\hat{f}^*(\mathbf{x}_j) > \log \sum_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} e^{\hat{f}^*(\mathbf{x}_i^{in})}$, we obtain $\frac{1}{m} \sum_{\mathbf{x}_j \in \mathbf{X}^{ood}} \hat{f}^*(\mathbf{x}_j) > \log \sum_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} e^{\hat{f}^*(\mathbf{x}_i^{in})}$ and thus by Eq 2 we get that $\hat{D}_{\hat{f}^*}(\mathbf{X}^{ood} \| \mathbf{X}^{in}) > \log N$. Now since $\hat{D}(\mathbf{X}^{ood} \| \mathbf{X}^{in}) = \max_{\hat{f}} \hat{D}_{\hat{f}}(\mathbf{X}^{ood} \| \mathbf{X}^{in})$, we have that $\hat{D}(\mathbf{X}^{ood} \| \mathbf{X}^{in}) \geq \hat{D}_{\hat{f}^*}(\mathbf{X}^{ood} \| \mathbf{X}^{in}) > \log N$. Note that the function attaining a maximum in $\hat{D}(\mathbf{X}^{ood} \| \mathbf{X}^{in})$ is not necessarily $\hat{f}^*$ and we don't make a such assumption.

□

1.4

Proof of Theorem 4. First we consider present → past direction in computing KL-D. We show present data points as a set P and historical past data points as H. For any subset T of the historical past, we show the KL divergence between present and this subset of the past by $\hat{D}_{kl}(P||T)$. Recall that we can assume there is a neural net function $\boldsymbol{f} : f_{P \to T}$ optimizing $\hat{D}_{kl}(P||T)$ for episodes $T \subseteq H$ of the past data. Formally we assume that

$$\hat{D}_{kl}(P||T) = \overline{\boldsymbol{f}^P} - \log \overline{e^{\boldsymbol{f}^T}}$$

For $T = P$, we have that $\hat{D}_{kl}(P||P)$ is near zero since the KL-D between P and P is zero. So we have

$$\hat{D}_{kl}(P||P) = \overline{\boldsymbol{f}^P} - \log \overline{e^{\boldsymbol{f}^P}} = O(1)$$

So $\overline{\boldsymbol{f}^P} \leq \log \overline{e^{\boldsymbol{f}^P}} + O(1)$.

Now using the fact that the replay samples R are taken with respect to the present bins distribution, we show that $|\log \overline{e^{\boldsymbol{f}^R}} - \log \overline{e^{\boldsymbol{f}^P}}| \leq d$. There is $\alpha > 0$ such that for each present bin B, if B has n_B present data points we sample αn_B

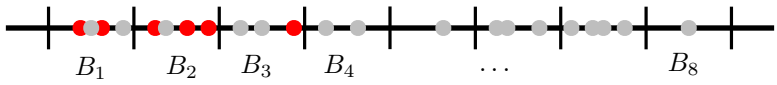

Figure 1: DV representation of data points in one dimension. The grey points represent past data and the red points represent present data. Observe that if we remove all the points in $B_4, \dots, B_8$, the mean value of past data points will decrease.

past points in B. Moreover for any present data x_P and replay data x_R in B we have $x_P - d \leq x_R \leq x_P + d$. Let B^R be the set of replay samples in B and B^P be the set of present samples in B. So

$$\alpha e^{-d} \sum_{x \in B^P} e^{\boldsymbol{f}(x)} \leq \alpha \sum_{x \in B^P} e^{\boldsymbol{f}(x-d)} \leq \sum_{x \in B^R} e^{\boldsymbol{f}(x)} \leq \alpha \sum_{x \in B^P} e^{\boldsymbol{f}(x+d)} \leq \alpha e^{d} \sum_{x \in B^P} e^{\boldsymbol{f}(x)}$$

So we have $\overline{e^{\boldsymbol{f}^P}} \cdot e^{-d} \leq \overline{e^{\boldsymbol{f}^R}} \leq \overline{e^{\boldsymbol{f}^P}} \cdot e^{d}$. And so $|\log \overline{e^{\boldsymbol{f}^R}} - \log \overline{e^{\boldsymbol{f}^P}}| \leq d$. This means that

$$0 \leq \hat{D}_{kl}(P||R) = \overline{\boldsymbol{f}^P} - \log \overline{e^{\boldsymbol{f}^R}} \leq \overline{\boldsymbol{f}^P} - \log \overline{e^{\boldsymbol{f}^P}} + d + O(1) \leq O(d)$$

The replay $\rightarrow$ present direction is very similar, we note it here for completeness.

Again there is a neural net function $\boldsymbol{g} : f_{R \rightarrow P}$ optimizing $\hat{D}(T||P)$ for episodes $T \subseteq H$ of the past data. This means:

$$\hat{D}_{kl}(T||P) = \overline{\boldsymbol{g}^T} - \log \overline{e^{\boldsymbol{g}^P}}$$

For $T = P$, we have

$$\hat{D}_{kl}(P||P) = \overline{\boldsymbol{g}^P} - \log \overline{e^{\boldsymbol{g}^P}} = O(1)$$

So $\overline{\boldsymbol{g}^P} \leq \log \overline{e^{\boldsymbol{g}^P}} + O(1)$. For every bin B, we have that the difference in value between any present sample and replay sample is at most d, and hence $|\overline{\boldsymbol{g}^P} - \overline{\boldsymbol{g}^R}| \leq d$ So we have

$$0 \leq \hat{D}_{kl}(R||P) = \overline{\boldsymbol{g}^R} - \log \overline{e^{\boldsymbol{g}^P}} \leq \overline{\boldsymbol{g}^P} - \log \overline{e^{\boldsymbol{g}^P}} + d + O(1) \leq O(d)$$

□

1.5

Lemma 1 *The value of the function* $\hat{D}_f(\mathbf{X_a}\|\mathbf{X_b}) := \frac{1}{|\mathbf{X_a}|} \sum_{\mathbf{x}_j \in \mathbf{X_a}} f(\mathbf{x}_j) - \log \frac{1}{|\mathbf{X_b}|} \sum_{\mathbf{x}_i \in \mathbf{X_b}} e^{f(\mathbf{x}_i)}$ *is unchanged if we replace* $f(x)$ *with* $\tilde{f}(x) = f(x) + c$ *for some constant* c *for all* $x \in R^k$*, given any two sets* $\mathbf{X_a}$ *and* $\mathbf{X_b} \subseteq R^k$*.*

Proof of Lemma 1.

$$\begin{aligned}
\hat{D}_{\tilde{f}}(\mathbf{X_a}\|\mathbf{X_b}) &= \frac{1}{|\mathbf{X_a}|} \sum_{\mathbf{x}_j \in \mathbf{X_a}} \tilde{f}(\mathbf{x}_j) - \log \frac{1}{|\mathbf{X_b}|} \sum_{\mathbf{x}_i \in \mathbf{X_b}} e^{\tilde{f}(\mathbf{x}_i)} \\
&= \frac{1}{|\mathbf{X_a}|} \sum_{\mathbf{x}_j \in \mathbf{X_a}} (f(\mathbf{x}_j) + c) - \log \frac{1}{|\mathbf{X_b}|} \sum_{\mathbf{x}_i \in \mathbf{X_b}} e^{f(\mathbf{x}_j)+c} \\
&= c + \frac{1}{|\mathbf{X_a}|} \sum_{\mathbf{x}_j \in \mathbf{X_a}} f(\mathbf{x}_j) - \log \frac{e^c}{|\mathbf{X_b}|} \sum_{\mathbf{x}_i \in \mathbf{X_b}} e^{f(\mathbf{x}_j)} \\
&= c + \frac{1}{|\mathbf{X_a}|} \sum_{\mathbf{x}_j \in \mathbf{X_a}} f(\mathbf{x}_j) - \log e^c - \log \frac{1}{|\mathbf{X_b}|} \sum_{\mathbf{x}_i \in \mathbf{X_b}} e^{f(\mathbf{x}_j)} \\
&= \frac{1}{|\mathbf{X_a}|} \sum_{\mathbf{x}_j \in \mathbf{X_a}} f(\mathbf{x}_j) - \log \frac{1}{|\mathbf{X_b}|} \sum_{\mathbf{x}_i \in \mathbf{X_b}} e^{f(\mathbf{x}_j)} = \hat{D}_f(\mathbf{X_a}\|\mathbf{X_b}).
\end{aligned} \tag{7}$$

□

Next, we prove Theorem 5. We define: $\hat{D}_f(\mathbf{X_a}\|\mathbf{X_b}) := \frac{1}{|\mathbf{X_a}|}\sum_{\mathbf{x}_j\in\mathbf{X_a}} f(\mathbf{x}_j) - \log\frac{1}{|\mathbf{X_b}|}\sum_{\mathbf{x}_i\in\mathbf{X_b}} e^{f(\mathbf{x}_i)}$ for any $\mathbf{X_a}, \mathbf{X_b} \subseteq R^k$. Then, $\hat{D}(\mathbf{X_a}\|\mathbf{X_b}) = \max_{f\in\mathcal{H}} \hat{D}_f(\mathbf{X_a}\|\mathbf{X_b})$ where $\mathcal{H}$ is the set of all learnable functions defined as follows: $\mathcal{H} \subseteq \{f : R^k \to R\}$ such that (i) $-\infty < f(x) < \infty$ for all $x \in R^k$ and (ii) If $f_1, f_2, g \in \mathcal{H}$, then functions of the form $f_1(x)I(g(x) \geq \tau) + f_2(x)I(g(x) < \tau)$ (which are essentially entirely derived from functions in $\mathcal{H}$) also lie in $\mathcal{H}$. Here $I(\cdot)$ is the indicator function. This stems from the intuition that if we are able to learn some functions on $R^k \to R$, then a function that is entirely derived from those functions should also be learnable.

Proof of Theorem 5. We have $\hat{D}(\mathbf{X}\|\mathbf{X}^{in}) = \max_{f\in\mathcal{H}} \hat{D}_f(\mathbf{X}\|\mathbf{X}^{in})$ and $\hat{D}(\bar{\mathbf{X}}\|\mathbf{X}^{in}) = \max_{f\in\mathcal{H}} \hat{D}_f(\bar{\mathbf{X}}\|\mathbf{X}^{in})$. Now, $\hat{f}_1 \in \mathcal{H}$ is such that $\hat{D}(\mathbf{X}\|\mathbf{X}^{in}) = \hat{D}_{\hat{f}_1}(\mathbf{X}\|\mathbf{X}^{in})$. We also let $\hat{f}_2 \in \mathcal{H}$ be a function such that $\hat{D}(\bar{\mathbf{X}}\|\mathbf{X}^{in}) = \hat{D}_{\hat{f}_2}(\bar{\mathbf{X}}\|\mathbf{X}^{in})$.

Observe that, $\hat{D}(\bar{\mathbf{X}}\|\mathbf{X}^{in}) = \max_{f\in\mathcal{H}} \hat{D}_f(\mathbf{X}\|\mathbf{X}^{in}) \geq \frac{1}{|\bar{\mathbf{X}}|}\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \hat{f}_1(\mathbf{x}_j) - \log\frac{1}{|\mathbf{X}^{in}|}\sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}_1(\mathbf{x}_i^{in})} = \hat{D}_{\hat{f}_1}(\bar{\mathbf{X}}\|\mathbf{X}^{in})$.

By way of contradiction, let us assume strict inequality: $\hat{D}_{\hat{f}_1}(\bar{\mathbf{X}}\|\mathbf{X}^{in}) < \hat{D}(\bar{\mathbf{X}}\|\mathbf{X}^{in})$.

Then, plugging in $\hat{f}_2$, we get,

$$\frac{1}{|\bar{\mathbf{X}}|}\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \hat{f}_1(\mathbf{x}_j) - \log\frac{1}{|\mathbf{X}^{in}|}\sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}_1(\mathbf{x}_i^{in})} < \frac{1}{|\bar{\mathbf{X}}|}\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \hat{f}_2(\mathbf{x}_j) - \log\frac{1}{|\mathbf{X}^{in}|}\sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}_2(\mathbf{x}_i^{in})}. \tag{8}$$

Or,

$$\begin{aligned}
\hat{D}(\mathbf{X}\|\mathbf{X}^{in}) &= \frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}_j\in\mathbf{X}} \hat{f}_1(\mathbf{x}_j) - \log\frac{1}{|\mathbf{X}^{in}|}\sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}_1(\mathbf{x}_i^{in})} \\
&< \frac{1}{|\bar{\mathbf{X}}|}\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \hat{f}_2(\mathbf{x}_j) + \frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}_j\in\mathbf{X}} \hat{f}_1(\mathbf{x}_j) - \frac{1}{|\bar{\mathbf{X}}|}\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \hat{f}_1(\mathbf{x}_j) - \log\frac{1}{|\mathbf{X}^{in}|}\sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}_2(\mathbf{x}_i^{in})} \\
&= \frac{1}{|\bar{\mathbf{X}}|}\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \hat{f}_2(\mathbf{x}_j) + \frac{1}{|\mathbf{X}|}\left(\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \hat{f}_1(\mathbf{x}_j) + \sum_{\mathbf{x}_j\in\mathbf{X}\setminus\mathbf{X}} \hat{f}_1(\mathbf{x}_j)\right) - \frac{1}{|\bar{\mathbf{X}}|}\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \hat{f}_1(\mathbf{x}_j) - \log\frac{1}{|\mathbf{X}^{in}|}\sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}_2(\mathbf{x}_i^{in})} \\
&= \frac{1}{|\mathbf{X}|}\left(\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \frac{|\mathbf{X}|}{|\bar{\mathbf{X}}|}\hat{f}_2(x) + (1 - \frac{|\mathbf{X}|}{|\bar{\mathbf{X}}|})\hat{f}_1(x) + \sum_{\mathbf{x}_j\in\mathbf{X}\setminus\mathbf{X}} \hat{f}_1(\mathbf{x}_j)\right) - \log\frac{1}{|\mathbf{X}^{in}|}\sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}_2(\mathbf{x}_i^{in})} \\
&\leq \frac{1}{|\mathbf{X}|}\left(\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \hat{f}_2(\mathbf{x}_j) + \sum_{\mathbf{x}_j\in\mathbf{X}\setminus\bar{\mathbf{X}}} \hat{f}_1(\mathbf{x}_j)\right) - \log\frac{1}{|\mathbf{X}^{in}|}\sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}_2(\mathbf{x}_i^{in})},
\end{aligned} \tag{9}$$

where the last line holds because, without loss of generality, we can assume that $\hat{f}_2(x) < \hat{f}_1(x)$. This is because, if the $\hat{f}_2(x)$ happens to be greater than $\hat{f}_1(x)$ at some values of x, using Lemma 1, we can always redefine another $\hat{f}_2(x) \in \mathcal{H}$ as the old $\hat{f}_2(x) - c$ where the constant c is an offset that is chosen appropriately, e.g., $c = \max_{x\in R} |\hat{f}_1(x) - \text{old}\hat{f}_2(x)| < \infty$ since $\hat{f}_1$ and old $\hat{f}_2$ also belong to $\mathcal{H}$.

Let us now define a function $\tilde{f}(x)$ as follows: $\tilde{f}(x) = \hat{f}_1(x)I(\hat{f}_1(x) > \tau) + \hat{f}_2(x)I(\hat{f}_1(x) \leq \tau)$. This function attains the following values over the subsets $\bar{\mathbf{X}}$, $\mathbf{X}\setminus\bar{\mathbf{X}}$ and $\mathbf{X}^{in}$: $\tilde{f}(x) = \begin{cases} \hat{f}_1(x), & x \in \mathbf{X}\setminus\bar{\mathbf{X}} \\ \hat{f}_2(x), & x \in \mathbf{X}^{in} \cup \bar{\mathbf{X}}, \end{cases}$ since $\hat{f}_1(x) > \tau$ for $x \in \mathbf{X}\setminus\bar{\mathbf{X}}$ and $\hat{f}_1(x) \leq \tau$ for $x \in \mathbf{X_{in}} \cup \bar{\mathbf{X}}$. The function $\tilde{f}$ also belongs to $\mathcal{H}$ because of its form that is entirely derived from other functions in $\mathcal{H}$.

But this means that we now have a function $\tilde{f}(x) \in \mathcal{H}$, such that $\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}_j\in\mathbf{X}} \tilde{f}(\mathbf{x}_j) - \log\frac{1}{|\mathbf{X}^{in}|}\sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\tilde{f}(\mathbf{x}_i^{in})} = \hat{D}_{\tilde{f}}(\mathbf{X}\|\mathbf{X}^{in}) > \hat{D}(\mathbf{X}\|\mathbf{X}^{in})$ which is a contradiction since $\hat{D}(\mathbf{X}\|\mathbf{X}^{in}) = \max_{f\in\mathcal{H}} \hat{D}_f(\mathbf{X}\|\mathbf{X}^{in})$.

Thus, the strict inequality ($\hat{D}_{\hat{f}_1}(\bar{\mathbf{X}}\|\mathbf{X}^{in}) < \hat{D}(\bar{\mathbf{X}}\|\mathbf{X}^{in})$) does not hold, and we have:

$$\hat{D}(\bar{\mathbf{X}}\|\mathbf{X}^{in}) = \hat{D}_{\hat{f}_1}(\bar{\mathbf{X}}\|\mathbf{X}^{in}) = \frac{1}{|\bar{\mathbf{X}}|}\sum_{\mathbf{x}_j\in\bar{\mathbf{X}}} \hat{f}_1(\mathbf{x}_j) - \log\frac{1}{|\mathbf{X}^{in}|}\sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} e^{\hat{f}_1(\mathbf{x}_i^{in})}$$

□

2 MORE ON EXPERIMENTAL ANALYSIS

2.1 VISUALIZATIONS

(a) SUN (b) Places (c) iNaturalist

(d) Textures (e) Animation (f) Arabic

(g) Tumors (h) Chest Xray (i) M. Pox

(j) M. Posters (k) YouTube T. (l) Shells P.

Figure 2: With our method *DDE** for OOD detection in WideResnet101, Imagenet dataset (ID set in blue) vs OOD test sets (in red) are shown to be separated in the respective dual functional spaces.

2.2 ANALYSIS FOR VITS

In Table 1, we present results for OOD detection in ViT-L-16. Note that OOD detection in pretrained Vision Transformers is under explored. The results suggest that all the methods are fundamentally limited in their capability for OOD detection in ViTs. It is only for a few OOD datasets such as USPS, Alzeihmers, Arabic Characters, Sign Language, Shells Pebbles, that we observe good performance across a majority of the methods. While our methods, *DDE** and *DDE-SM**, are signfiicantly superior w.r.t. all the methods for ViT-L-16, the ViT does limit even our proposed OOD detectors in comparison to the WideResnet.

2.3 VARYING SAMPLE SIZE IN IMAGENET (ID) DATASET

We perform a new ablation study for our method (DDE*) by varying the sample size (N) on the Imagenet (ID) dataset. Following the same experimental setup as in Table 1 in the paper, we present results in the table below. "All" refers to using all the samples in the Imagenet dataset for OOD detection which is the same as column "DDE*" in Table 1 of the paper. For a given sample size, we randomly select samples from the Imagenet dataset and use only those samples for the entire experiment including tuning the hyperparameters. We perform 10 random trials, and correspondingly report mean and standard deviation of the FPR95 scores for each of the 51 test OOD sets. We observe that, for many of the OOD test sets, even a small sample size of a few thousands ($N = 3000$) suffices to achieve high OOD detection rate. However, on the extreme end, using a sample size of 100 is clearly not enough. Note that the numbers from this study should not be compared to other methods in the paper, since the latter use the entire Imagenet dataset.

2.4 BATCH INFERENCE ON TEST SET

Following the same experimental setup as in Table 1 in the paper, we perform an ablation study of OOD detection in a test set splitting it into (100) small ordered batches of equal sizes; batch size across the OOD test set varies from 2 to 430 with median value of 28. We believe batch inference of a test set to strongly resemble real world scenarios of continual lifelong learning. For attaining a reasonable sample size in a test set (though not necessary), we augment each batch of test samples with (300) randomly selected samples from ID training set (i.e. Imagenet dataset) and (300) samples from the OOD validation set (same as discussed in the paper, generated from ID samples in Imagenet by simple perturbations proposed by Hendrycks et al. [2019]). We perform 10 trials to account for randomness in selecting the samples for augmentation. In the table below, referring to this batch-inference based online version of our method as "DDE-Online", we present mean and standard deviation of FPR95 scores (from 10 trials) for each of the OOD test sets. In addition, for a comparison, we present the original results for our method ("DDE*") as well as the best of all the baselines (which is different for each OOD test set) from Table 1 in the paper. It is interesting to note that the standard deviation of FPR95 scores is low and that it performs even better than "DDE*" for many test sets. Even for most of the other cases, "DDE-Online" has lower FPR95 than the best of the baselines.

2.5 FIXED ESTIMATOR TUNED FOR TEST SETS

As per the reviewer's suggestion, in the table below, we present results from an ablation study on generalization of the estimator. We optimize the dual function for estimating KL-divergence between the ID training set and the OOD validation set. Using this dual function, we perform OOD detection across all the OOD test sets. This highly compute efficient variant of our method is referred as "DDEv". Optionally, we fine tune for a given test set using 10% or 20% of the original compute cost of our method (DDEvt10 and DDEvt20). For a comparison, we also present results for the default version of our method DDE* and the best of the baselines from Table 1 in the paper. FPR95 scores in the table below suggest that the estimator does generalize to many OOD test sets, and it further benefits from fine tuning.

2.6 MIXTURE OF ID & OOD SAMPLES IN TEST SET

We evaluate our approach for OOD detection on test sets containing both ID and OOD samples. We augment each OOD test set with (3000) ID test samples. Besides this change, evaluation setup is same as for Table 1 in the paper. Results for this setting are denoted as "DDE-mixed". In the table below, for each test set, we report FPR95 scores for OOD samples, as well as for ID samples in parenthesis. In addition, for a comparison, we present the results of "DDE*" as well as the best of all the baselines from Table 1 in the paper. Our method detects ID samples in each test set with a very high accuracy (FPR95 >

Dataset	msp	mls	odin	ebo	gn	react	gm	knn	dice	ash	wm	klm	cider	ige	dde*	dde-sm*
ID Test↑	95	94	93	94	95	93	92	95	95	93	94	93	94	94	**96**	95
OOD Val.	77	74	75	68	68	67	67	79	68	66	83	73	79	68	**54**	59
SUN	96	96	94	98	97	98	98	99	99	97	–	96	96	98	**91**	**91**
Places	96	95	**94**	97	96	96	95	98	98	96	–	96	95	97	96	**94**
iNaturalist	93	93	90	95	94	94	98	99	96	92	–	93	93	95	76	**66**
Textures	95	93	94	90	91	88	89	94	90	85	–	94	95	90	71	**58**
Agr. Crop	85	89	78	98	96	97	98	98	98	98	–	96	85	97	40	**33**
Animation	–	99	99	–	–	–	–	–	–	–	–	99	–	–	**18**	19
B. Tumors	91	88	86	91	93	89	**13**	30	92	66	–	96	67	92	31	31
C. Xray	90	91	84	–	–	–	80	–	–	–	–	90	93	–	**4**	7
Faces in W.	91	88	87	93	95	90	59	83	91	90	–	92	88	93	**30**	35
Fastfood	98	98	97	98	98	97	96	98	96	97	–	97	98	97	62	**50**
Gemstone	–	–	99	–	–	–	97	99	–	99	–	–	99	–	78	**63**
Lego	–	–	99	–	–	99	98	99	–	99	–	98	97	–	**66**	**66**
Plant D.	99	–	99	–	–	–	–	–	–	–	–	–	–	–	27	**20**
USPS	**0**	**0**	**0**	**0**	–	–	–	**0**	–	–	–	**0**	**0**	–	**0**	**0**
Alzeihmers	17	3	7	4	14	1	2	98	53	**0**	–	68	76	8	**0**	**0**
B. Cells	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**11**	12
B. Logos	98	98	98	98	98	98	95	98	99	98	–	98	98	98	**23**	25
Captcha	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**0**	**0**
Cards	98	98	98	96	97	96	94	96	97	94	–	97	97	96	47	**36**
Arabic	48	36	43	19	15	18	**0**	1	2	1	29	66	12	24	**0**	**0**
Chess	94	93	93	90	91	90	82	86	90	87	95	96	90	90	**74**	80
C. Fine Art	99	99	99	98	98	98	97	98	97	97	–	99	99	98	68	**48**
Coffee B.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**3**	5
Colon S.	–	–	–	–	–	–	84	94	–	–	–	–	–	–	**0**	**0**
Covid CT S.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**6**	7
Diamonds	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**12**	15
E. Faces	–	–	99	–	–	–	90	98	–	–	–	–	99	–	**10**	14
H. Eyes	–	–	–	–	–	–	99	–	–	–	–	–	–	–	**13**	15
Fire & S.	–	–	–	98	–	98	68	69	90	98	95	–	82	99	**57**	63
H.W. Eng.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**0**	**0**
Excavation	–	–	99	99	99	97	–	–	99	99	–	93	–	99	23	**22**
Eyes	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**29**	31
H.W. Math	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**8**	**8**
H. & B.	99	99	98	99	99	99	99	–	–	99	–	96	–	99	**8**	10
I. Food	97	96	96	95	96	94	92	96	93	94	–	97	96	95	75	**56**
Lego M. F.	–	–	99	–	–	99	98	99	–	99	–	98	97	–	66	**61**
Licence P.	81	81	79	91	96	91	66	81	94	89	–	94	64	92	**26**	34
Meat Q.	–	–	–	–	–	99	–	–	–	–	–	–	–	–	**0**	**0**
M. Pox	–	–	99	–	99	99	83	99	98	99	–	99	–	99	55	**52**
M. Posters	87	82	83	75	78	72	75	81	74	71	–	88	84	75	**43**	52
Orna. P.	–	–	99	–	–	–	–	–	–	99	–	98	–	–	**11**	15
Paintings	96	96	95	98	98	97	**41**	43	98	97	67	–	88	98	51	51
Pollen G.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**6**	8
QR C.	86	83	77	–	–	97	75	–	–	47	–	95	99	–	**0**	**0**
Railway T.	–	–	–	–	–	–	–	–	–	–	–	99	–	–	21	**19**
Weed C.	76	72	72	71	71	70	63	85	76	68	–	87	78	71	**33**	40
YouTube T.	84	78	79	76	77	72	75	83	79	72	–	83	83	75	59	**51**
Weather	99	98	98	98	98	98	95	97	99	97	–	98	95	98	**87**	89
Sign L.	34	25	30	6	11	4	**0**	2	1	1	98	65	38	8	**0**	1
Stairs	98	98	98	99	99	99	98	98	99	98	99	99	97	99	**51**	71
Shells P.	**0**	**0**	**0**	**0**	93	90	–	**0**	93	85	–	**0**	**0**	–	**0**	**0**

Table 1: Evaluation results for OOD detection in ViT-L-16 pretrained on Imagenet-1k using the metric FPR95 (↓). Due to space constraints, we display method names in lower case and use "–" wherever FPR95 is 100. Best scores are shown in bold and the second best scores are underlined.

Dataset	All	N=30000	N=10000	N=3000	N=1000	N=100
ID Test ↑	95	95±1	94±1	94±1	95±2	96±4
OOD Validation	31	46±6	47±5	42±5	35±5	33±22
SUN	18	29±8	33±6	32±9	33±23	44±35
Places	10	32±12	37±8	34±9	34±23	45±35
iNaturalist	11	22±9	28±9	24±6	28±17	39±31
Textures	15	27±10	38±13	32±8	34±12	61±37
Agriculture Crop	0	3±3	9±7	8±4	17±28	19±21
Animation	6	14±7	20±7	18±6	19±11	33±29
Brain Tumors	3	8±5	11±5	12±4	11±2	35±33
Chest Xray	4	9±5	14±6	14±6	12±4	44±37
Faces in the Wild	9	16±8	23±8	19±8	24±26	37±33
Fastfood	10	27±8	35±9	33±9	27±7	44±35
Gemstones	4	10±6	17±9	18±6	16±6	40±33
LEGO	0	3±3	6±5	6±3	15±28	40±37
Plant Diseases	2	8±5	13±5	13±5	15±13	47±35
USPS	1	5±3	9±5	8±4	7±2	29±31
Alzeihmers	1	4±4	6±4	6±3	5±2	28±36
Blood Cells	1	5±4	9±6	9±4	17±24	21±25
Brand Logos	0	0±0	1±2	1±1	1±1	7±16
Captcha	0	0±0	0±0	0±0	0±0	0±0
Cards	11	17±7	23±12	19±7	21±9	45±32
Arabic Handwritten Characters	4	7±4	10±5	9±3	8±2	18±20
Chess	1	5±4	10±5	11±5	11±4	32±34
Chinese Fine Art	1	4±4	9±5	9±5	17±28	44±36
Coffee Beans	1	4±3	6±5	6±3	5±3	18±24
Colonoscopy	1	3±2	6±4	5±2	5±2	22±28
Covid CT Scan	3	7±5	11±5	11±5	12±6	24±27
Diamonds	3	5±3	9±5	7±3	7±2	35±33
Emotional Faces	5	13±7	20±10	17±7	22±26	39±34
Human Eyes	5	11±6	19±7	17±9	15±5	28±30
Fire & Smoke	0	0±0	1±1	0±1	10±30	11±20
English Handwritten Characters	2	4±3	7±4	6±3	6±2	15±18
Excavation	0	2±2	4±3	4±2	13±29	41±40
Eyes	3	5±3	7±4	6±3	9±10	35±34
Handwritten Math Symbols	1	3±3	±4	6±3	6±2	15±19
Bart and Homer	0	1±1	3±3	3±2	13±29	16±20
Indian Food	13	23±11	29±9	30±8	32±19	41±33
Lego Minifigures	0	3±3	7±4	6±4	15±29	32±35
Licence Plates	0	0±0	0±1	0±1	0±1	20±38
Meat Quality	0	0±1	2±3	1±1	1±2	11±26
Monkeypox	8	14±7	21±7	21±6	29±25	44±35
Movie Posters	14	26±10	35±9	31±9	26±10	37±30
Ornamental Plants	0	3±4	6±5	6±3	14±29	21±26
Paintings	1	5±4	9±5	8±4	16±28	21±26
Pollen Grain	1	4±4	9±6	11±5	9±3	25±27
QR Codes	1	2±2	4±3	3±2	3±2	12±16
Railway Tracks	1	2±2	6±5	6±4	14±29	31±34
Weed Crop	4	6±4	9±6	9±4	17±28	27±25
YouTube Thumbnail	5	22±11	31±8	27±7	31±24	49±33
Weather	14	27±8	33±12	29±7	31±20	51±40
Sign Language	1	3±4	6±5	6±3	4±1	22±24
Stairs	0	0±0	1±2	1±1	0±1	9±22
Shells or Pebbles	22	26±6	31±9	31±8	26±9	53±35

Table 2: Evaluation results for OOD detection in WideResnet101 pretrained on Imagenet-1k using the metric FPR95 (↓).

Dataset	Best of the Baselines	DDE*	**DDE-Online**
SUN	12	18	14±1
Places	34	10	14±0
iNaturalist	12	11	8±1
Textures	12	15	10±1
Agriculture Crop	0	0	3±1
Animation	21	6	4±1
Brain Tumors	14	3	6±1
Chest Xray	7	4	4±0
Faces in the Wild	19	9	5±1
Fastfood	47	10	14±1
Gemstone	39	4	11±1
LEGO	2	0	4±1
Plant Diseases	14	2	6±1
USPS	12	1	1±0
Alzeihmers	4	1	1±0
Blood Cells	6	1	6±1
Brand Logos	0	0	0±0
Captcha	0	0	0±0
Cards	59	11	9±1
Arabic Handwritten Characters	4	4	0±0
Chess Pieces	9	1	7±1
Chinese Fine Art	2	1	7±1
Coffee Beans	10	1	4±1
Colonoscopy	1	1	2±1
Covid CT Scans	11	3	6±1
Diamonds	31	3	5±1
Emotional Faces	15	5	4±0
Human Eyes	20	5	5±1
Fire & Smoke	0	0	0±0
English Handwritten Characters	8	2	2±1
Excavation	1	0	2±1
Eyes	11	3	5±1
Handwritten Math Symbols	10	1	1±1
Bart and Homer	0	0	1±0
Indian Food	49	13	14±1
LEGO Minifigures	1	0	4±1
Licence Plates	0	0	0±0
Meat Quality	0	0	0±0
Monkeypox	50	8	9±1
Movie Posters	37	14	13±1
Ornamental Plants	10	0	3±2
Paintings	2	1	5±1
Pollen Grain	12	1	6±2
QR Codes	5	1	0±0
Railway Tracks	1	1	2±1
Weed Crops	26	4	7±1
YouTube Thumbnails	40	5	17±2
Weather	58	14	16±1
Sign Language	10	1	2±1
Stairs	0	0	0±0
Shells or Pebbles	59	22	14±1

Table 3: Evaluation results for OOD detection in WideResnet101 pretrained on Imagenet-1k using the metric FPR95 (↓).

Dataset	Best of the Baselines	DDE*	DDEv	DDEvt10	DDEvt20
SUN	12	18	33	22	21
Places	34	10	32	23	16
iNaturalist	12	11	29	15	15
Textures	12	15	82	65	43
Agriculture Crop	0	0	0	0	0
Animation	21	6	6	6	6
Brain Tumors	14	3	5	6	6
Chest Xray	7	4	3	7	7
Faces in the Wild	19	9	7	6	6
Fastfood	47	10	35	24	17
Gemstone	39	4	26	10	9
LEGO	2	0	1	3	3
Plant Diseases	14	2	4	10	10
USPS	12	1	2	4	4
Alzeihmers	4	1	1	2	2
Blood Cells	6	1	2	7	7
Brand Logos	0	0	0	0	0
Captcha	0	0	0	0	0
Cards	59	11	50	21	15
Arabic Handwritten Characters	4	4	3	5	5
Chess Pieces	9	1	5	6	6
Chinese Fine Art	2	1	2	11	11
Coffee Beans	10	1	2	3	3
Colonoscopy	1	1	0	0	0
Covid CT Scans	11	3	3	4	4
Diamonds	31	3	16	6	6
Emotional Faces	15	5	4	6	6
Human Eyes	20	5	6	6	6
Fire & Smoke	0	0	0	0	0
English Handwritten Characters	8	2	0	1	1
Excavation	1	0	0	1	1
Eyes	11	3	8	5	5
Handwritten Math Symbols	10	1	1	2	2
Bart and Homer	0	0	0	0	0
Indian Food	49	13	31	18	13
LEGO Minifigures	1	0	0	3	3
Licence Plates	0	0	0	0	0
Meat Quality	0	0	0	0	0
Monkeypox	50	8	31	14	11
Movie Posters	37	14	23	15	12
Ornamental Plants	10	0	2	4	4
Paintings	2	1	2	4	4
Pollen Grain	12	1	6	7	7
QR Codes	5	1	0	3	3
Railway Tracks	1	1	0	1	1
Weed Crops	26	4	10	5	5
YouTube Thumbnails	40	5	27	18	14
Weather	58	14	51	29	21
Sign Language	10	1	2	4	4
Stairs	0	0	0	0	0
Shells or Pebbles	59	22	44	28	20

Table 4: Evaluation results for OOD detection in WideResnet101 pretrained on Imagenet-1k using the metric FPR95 (↓).

94). As for detecting OOD samples, for many of the test sets, our method achieves lower FPR95 scores (as desired) w.r.t. the best of the baselines.

2.7 DETAILS ON DUAL DIVERGENCE ESTIMATION VIA DEEP NEURAL NETWORKS

As it has been explored in the previous works for dual divergence estimation [Belghazi et al., 2018], we employ a lightweight deep neural net, independent of the pretrained DNN, as a dual function approximator. The neural dual function is optimized via maximization of the divergence measure w.r.t. the weight parameters. Large batch size (10k in our experiments) is recommended to avoid otherwise high variance in estimating the measure [Song and Ermon, 2019].

Besides, DNNs present the challenge of overfitting. In the context of divergence estimation, it means that if we perform a very large number of batch updates, the estimate can eventually diverge. In practice, a few hundred batch updates with low learning rate (5e-4 in our experiments) suffice to converge before the phenomenon of divergence may start to take place after a few thousand batch updates.

The neural architecture, along with the hyperparameters such as learning rate, and number of batch updates, can be automatically tuned such that 5% of the samples in ID set are identified as OOD as it is the standard practice in all the previous works on OOD detection in pretrained networks (corresponding to metric FPR95). Furthermore, as suggested in previous works, one can also minimize false positive rates on a validation set of OOD samples which is generated via various kinds of perturbations performed on ID samples [Hendrycks et al., 2019].

It is also worth noting that, while the architecture and all the hyperparameters are fixed after the tuning on the ID set, the weight parameters of a DNN are optimized independently for each test set. This is because the dual function is unique to the problem of dual divergence estimation between a test set and the ID set. Despite this, we find in our experimental analysis that the average compute time for OOD detection in a test set (of size in few thousand) is in seconds.

2.8 DATASETS FOR ID DETECTION

US stocks prices. We started with the 1000 stocks from the constituents of the Russell 3000 index that have the highest liquidity. This dataset is publicly available, though very large in size to be released as a single file. After performing necessary preprocessing and checks on data quality issues, we use 982 of those stocks. The returns are evaluated every 5 minutes, for the period of from May 2021 to May 2022, i.e. 7800 timesteps.

ECG dataset is available on Kaggle.[1]

2.9 TEST DATASETS FOR OOD DETECTION

All the new datasets are available at Kaggle. For the previously benchmarked test OOD datasets, we obtained the preprocessed versions from the respective sources.

References

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of International Conference on Machine Learning*, 2018.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *Proceedings of International Conference on Learning Representations*, 2019.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

[1] `https://www.kaggle.com/datasets/shayanfazeli/heartbeat?select=ptbdb_abnormal.csv`

Dataset	Best of the Baselines	DDE*	DDE-mixed
SUN	12	18	37 (99)
Places	34	10	36 (98)
iNaturalist	12	11	17 (98)
Textures	12	15	20 (98)
Agriculture Crop	0	0	25 (95)
Animation	21	6	10 (98)
Brain Tumors	14	3	1 (96)
Chest Xray	7	4	1 (96)
Faces in the Wild	19	9	7 (96)
Fastfood	47	10	23 (97)
Gemstone	39	4	5 (98)
LEGO	2	0	14 (95)
Plant Diseases	14	2	5 (95)
USPS	12	1	0 (98)
Alzeihmers	4	1	0 (96)
Blood Cells	6	1	11 (98)
Brand Logos	0	0	1 (94)
Captcha	0	0	0 (94)
Cards	59	11	9 (98)
Arabic Handwritten Characters	4	4	2 (98)
Chess Pieces	9	1	16 (94)
Chinese Fine Art	2	1	28 (95)
Coffee Beans	10	1	0 (98)
Colonoscopy	1	1	4 (94)
Covid CT Scans	11	3	0 (96)
Diamonds	31	3	0 (98)
Emotional Faces	15	5	10 (99)
Human Eyes	20	5	3 (97)
Fire & Smoke	0	0	2 (94)
English Handwritten Characters	8	2	0 (98)
Excavation	1	0	9 (95)
Eyes	11	3	1 (98)
Handwritten Math Symbols	10	1	0 (97)
Bart and Homer	0	0	6 (94)
Indian Food	49	13	20 (97)
LEGO Minifigures	1	0	11 (94)
Licence Plates	0	0	1 (94)
Meat Quality	0	0	0 (97)
Monkeypox	50	8	5 (97)
Movie Posters	37	14	35 (98)
Ornamental Plants	10	0	0 (94)
Paintings	2	1	21 (95)
Pollen Grain	12	1	7 (96)
QR Codes	5	1	0 (98)
Railway Tracks	1	1	9 (94)
Weed Crops	26	4	2 (98)
YouTube Thumbnails	40	5	38 (97)
Weather	58	14	35 (99)
Sign Language	10	1	0 (96)
Stairs	0	0	12 (94)
Shells or Pebbles	59	22	31 (98)

Table 5: Evaluation results for OOD detection in WideResnet101 pretrained on Imagenet-1k using the metric FPR95 (↓).