SCOD: From Heuristics to Theory

Vojtech Franc, Jakub Paplham, and Daniel Prusa

Department of Cybernetics, Faculty of Electrical Engineering,
Czech Technical University in Prague, Czech Republic
{xfrancv,paplhjak,prusapa1}@fel.cvut.cz

Abstract. This paper addresses the problem of designing reliable prediction models that abstain from predictions when faced with uncertain or out-of-distribution samples - a recently proposed problem known as Selective Classification in the presence of Out-of-Distribution data (SCOD). We make three key contributions to SCOD. Firstly, we demonstrate that the optimal SCOD strategy involves a Bayes classifier for in-distribution (ID) data and a selector represented as a stochastic linear classifier in a 2D space, using i) the conditional risk of the ID classifier, and ii) the likelihood ratio of ID and out-of-distribution (OOD) data as input. This contrasts with suboptimal strategies from current OOD detection methods and the Softmax Information Retaining Combination (SIRC), specifically developed for SCOD. Secondly, we establish that in a distribution-free setting, the SCOD problem is not Probably Approximately Correct learnable when relying solely on an ID data sample. Third, we introduce POSCOD, a simple method for learning a plugin estimate of the optimal SCOD strategy from both an ID data sample and an unlabeled mixture of ID and OOD data. Our empirical results confirm the theoretical findings and demonstrate that our proposed method, POSCOD, outperforms existing OOD methods in effectively addressing the SCOD problem.

Keywords: out-of-distribution detection · selective classification · optimal strategy · probably approximately correct learning

1 Introduction

Standard methods for learning predictors from data rely on the closed-world assumption, i.e., the training and testing samples are generated from the same distribution, so-called In-Distribution (ID). In real-world applications, ID test samples can be contaminated by samples from another distribution, the so-called Out-Of-Distribution (OOD), which is not represented by the training sample. In recent years, the growing interest in deep learning models capable of handling OOD data has resulted in numerous papers on effective OOD detection (OODD) [4, 7, 8, 13, 15, 24, 26, 34–36, 41]. Notably, despite the practical role of OOD detector as selector of input samples for ID classifier, prior work has not explicitly addressed the consideration of misclassified ID samples in selector design. The concept of classifiers equipped with selectors to reject predictions on likely misclassified input samples, known as selective classification (SC), has been

studied separately in closed-world scenario [5, 11, 12, 31]. Recent research [3, 19, 27, 42] underscores the need for selective classifiers that simultaneously address OODD and SC goals. The newly introduced prediction problem is called Selective Classification in the presence of Out-of-Distribution data (SCOD) [42].

SCOD aims to detect OOD samples, abstaining from predictions on them, while simultaneously minimizing the prediction error on accepted ID samples. To address this problem, [42] introduces the Softmax Information Retaining Combination (SIRC) heuristic strategy. SIRC constructs selectors by combining two scores, one focused on detecting misclassified ID samples and the other on identifying OOD samples. Despite demonstrating superior performance over existing OOD detectors in the SCOD problem, as shown in empirical evidence by [42], it remains unclear whether the SIRC strategy is optimal and whether the SCOD problem can be solved from the available data. In this paper, we address these questions, leveraging the answers to propose a theoretically grounded approach that consistently outperforms existing methods. In particular, we provide the following contributions to the SCOD problem:

1. We demonstrate that the optimal prediction strategy for solving the SCOD problem comprises the Bayes classifier for ID data and a selector represented as a stochastic linear classifier in a 2D space. The input features for this selector are the conditional risk of the ID classifier and the OOD/ID likelihood ratio. Our findings reveal that current OODD methods, as well as the SIRC, yield suboptimal strategies for the SCOD problem.
2. We extend the concept of Probably Approximately Correct (PAC) learnability [10, 33] to address the SCOD problem. Additionaly, we prove that in a distribution-free setting, the SCOD problem is not PAC-learnable when the learning algorithm exclusively depends on an ID data sample.
3. We introduce a method POSCOD for learning the plugin estimate of the optimal SCOD strategy from both an ID data sample and an unlabeled mixture of ID and OOD data. POSCOD simplifies the learning process to i) training an ID classifier through standard cross-entropy loss and ii) training a classifier using the binary cross-entropy (BCE) of a novel corrected sigmoid.
4. We empirically confirm our theoretical findings and demonstrate that our proposed method, POSCOD, outperforms existing OODD methods and SIRC when applied to the SCOD problem.

It is worth noting that while the initial two contributions are theoretical in nature, their practical significance extends to any future SCOD method. The first contribution, characterizing the structure of the optimal strategy, effectively narrows the pool of predictors suitable for the SCOD problem. The second contribution establishes that attempts to devise efficient learning algorithms for SCOD, without assumptions on data distribution and relying solely on ID data, are futile. Notably, many existing OODD and SCOD methods, by making no explicit distribution assumptions and using only the ID sample, are unable to be PAC learners. The proposed algorithm POSCOD serves as a prime example of methods that are in line with these guidelines, and our empirical findings confirm its superior performance compared to existing approaches.

Proofs of all theorems can be found in the Appendix.

2 The SCOD problem and its optimal solution

SCOD is a decision-making problem that aims to design a selective classifier applied to samples from a mixture of ID and OOD. The selective classifier comprises two functions: a classifier of ID data and a selective function (or a *selector* for short). The selector determines which samples are accepted for prediction by the ID classifier and which are rejected as OOD samples or ID samples likely to be misclassified by the ID classifier. In this section, we first define the SCOD problem and the concept of optimal strategy following [42]. Then, we present our first contribution, which shows that *optimal SCOD strategies involve the Bayes classifier of ID data and a selector being a stochastic linear classifier in a 2D space, whose input features are the conditional risk of the ID classifier and the OOD/ID likelihood ratio*. Our result shows that existing OOD detection methods and the state-of-the-art method for SCOD, the SIRC [42], return suboptimal strategies for the SCOD problem.

2.1 Definitions

Data distribution Let $\mathcal{X}$ be a set of observable inputs, and $\mathcal{Y}$ a finite set of labels that can be assigned to ID data. ID samples $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are generated i.i.d. from a joint distribution $p_I \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. OOD samples $x \in \mathcal{X}$ are generated from a distribution $p_O \colon \mathcal{X} \to \mathbb{R}_+$. ID and OOD samples share the same input space $\mathcal{X}$. Let $\emptyset$ be a special label to mark the OOD sample, and $\bar{\mathcal{Y}} = \mathcal{Y} \cup \{\emptyset\}$ an extended label set. At the *deployment stage*, the samples $(x, \bar{y}) \in \mathcal{X} \times \bar{\mathcal{Y}}$ are generated from the joint distribution $p \colon \mathcal{X} \times \bar{\mathcal{Y}} \to \mathbb{R}_+$ defined as a mixture of ID and OOD [10]:

$$p(x, \bar{y}) = \begin{cases} p_O(x)\, \pi_{\mathrm{O}} & \text{if} \quad \bar{y} = \emptyset\,, \\ p_I(x, \bar{y})\, (1 - \pi_{\mathrm{O}}) & \text{if} \quad \bar{y} \in \mathcal{Y}\,, \end{cases} \tag{1}$$

where $\pi_{\mathrm{O}} \in [0, 1]$ is the portion of OOD data in the mixture.

Selective classifier The ultimate goal is to design a reject option strategy $q \colon \mathcal{X} \to \mathcal{D}$, where $\mathcal{D} = \mathcal{Y} \cup \{\text{reject}\}$ is the decision set, which either predicts a label, $q(x) \in \mathcal{Y}$, or rejects the prediction, $q(x) = \text{reject}$. Following [12], we represent the reject option strategy q by a selective classifier (h, c) that comprises the ID classifier $h \colon \mathcal{X} \to \mathcal{Y}$, and a stochastic selector $c \colon \mathcal{X} \to [0, 1]$ which outputs a probability that the input is accepted, i.e.,

$$q(x) = (h, c)(x) = \begin{cases} h(x) & \text{with probability} \quad c(x)\,, \\ \text{reject} & \text{with probability} \quad 1 - c(x)\,. \end{cases} \tag{2}$$

We use the stochastic selector because it turns out to be an optimal solution in the general setting; however, we will show that in most practical settings, the deterministic strategy $c \colon \mathcal{X} \to \{0, 1\}$ suffices.

Evaluation metrics We define three base metrics to evaluate the performance of the SCOD strategy (h, c). One role of the selector $c\colon \mathcal{X} \to [0, 1]$ is to discriminate ID/OOD samples. We consider ID and OOD samples as positive and negative classes, respectively. We evaluate the performance of the selector by the True Positive Rate (TPR) and the False Positive Rate (FPR). The TPR/FPR is the probability that the ID/OOD sample is accepted by the selector c, i.e.,

$$\mathrm{tpr}(c) = \mathbb{E}_{x\sim p_I(x)} c(x) \quad \text{and} \quad \mathrm{fpr}(c) = \mathbb{E}_{x\sim p_O(x)} c(x)\,. \tag{3}$$

The performance of the ID classifier $h\colon \mathcal{X} \to \mathcal{Y}$ on the accepted samples w.r.t. user-defined loss $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is characterized by the selective risk [12]

$$\mathrm{R_S}(h, c) = \frac{\mathbb{E}_{(x,y)\sim p_I(x,y)}[\ell(y, h(x))\, c(x)]}{\mathrm{tpr}(c)}\,,$$

which is defined for non-zero $\mathrm{tpr}(c)$.

Definition 1 (SCOD problem). *Let* $\mathrm{tpr}_{\min} \in (0, 1)$ *be a user-defined minimum acceptable TPR and* $\alpha \in [0, 1]$ *a relative cost associated with not rejecting an OOD sample. The SCOD problem involves solving*

$$\min_{\substack{h\in\mathcal{Y}^{\mathcal{X}}\\ c\in[0,1]^{\mathcal{X}}}} \big[(1-\alpha)\,\mathrm{R_S}(h, c) + \alpha\,\mathrm{fpr}(c)\big] \qquad \textit{s.t.} \qquad \mathrm{tpr}(c) \geq \mathrm{tpr}_{\min}\,, \tag{4}$$

where we assume that both minimizers exist. A selective classifier (h^*, c^*) *that solves (4) is called an optimal SCOD strategy. We refer to* $R(h, c) = (1 - \alpha)\,\mathrm{R_S}(h, c) + \alpha\,\mathrm{fpr}(c)$ *as the SCOD risk.*

Def. 1 is a slight generalization of the formulation proposed in [42], which assumes only the 0/1 loss $\ell(y, y') = [\![y \neq y']\!]$. The analysis and methods in this paper apply to any loss $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ such that $\ell(y, y') = 0$ iff $y = y'$.

The formulation of the SCOD problem (4) is straightforward, intuitive, and offers several advantages over alternative formulations. For example, one could substitute TPR in the constraint with the total coverage $\rho(c) = \mathrm{tpr}(c)(1 - \pi_O) + \mathrm{fpr}(c)\pi_O$. A notable advantage of the SCOD formulation (4) is its independence on the portion of OOD data π_O that is unknown and often non-stationary in practice. Additional benefits over alternatives are discussed in Appendix A.

2.2 The optimal SCOD strategy

In this section, we present our main result, which shows how to construct an optimal SCOD strategy.

Theorem 1. *Let* (h^*, c^*) *be an optimal solution to (4). Then* (h_B, c^*)*, where* h_B *is the Bayes ID classifier*

$$h_B(x) \in \operatorname*{Argmin}_{y'\in\mathcal{Y}} \sum_{y\in\mathcal{Y}} p_I(y \mid x)\ell(y, y')\,, \tag{5}$$

is also optimal to (4).

Theorem 1 ensures that the Bayes ID classifier h_B is an optimal solution to (4). After approximating h^*, e.g., by our best estimate of h_B learned from the data, the search for an optimal selector c leads to the following problem:

Problem 1. (Optimal SCOD selector c for known ID classifier h) *Given ID classifier $h\colon \mathcal{X} \to \mathcal{Y}$, and the user-defined parameters $\mathrm{tpr}_{\min} \in (0,1)$ and $\alpha \in [0,1]$, the optimal selector $c^*\colon \mathcal{X} \to [0,1]$ is a solution to*

$$\min_{c\in[0,1]^{\mathcal{X}}} \big[(1-\alpha)\,\mathrm{R_S}(h,c) + \alpha\,\mathrm{fpr}(c)\big] \qquad s.t. \qquad \mathrm{tpr}(c) \geq \mathrm{tpr}_{\min}\,. \tag{6}$$

Theorem 2. *Let $h\colon \mathcal{X} \to \mathcal{Y}$ be any ID classifier and $r\colon \mathcal{X} \to \mathbb{R}$ its conditional risk $r(x) = \sum_{y\in\mathcal{Y}} p_I(y \mid x)\ell(y, h(x))$. Let $g(x) = p_O(x)/p_I(x)$ be the likelihood ratio of the OOD and ID samples. Then, the set of optimal solutions of Problem 1 contains the selector*

$$c^*(x) = \begin{cases} 0 & \text{if} \quad s(x) > \lambda \\ \tau & \text{if} \quad s(x) = \lambda \\ 1 & \text{if} \quad s(x) < \lambda \end{cases} \qquad \text{with the score} \quad s(x) = r(x) + \frac{\alpha\,\mathrm{tpr}_{\min}}{1-\alpha}\,g(x) \tag{7}$$

for $\alpha \in [0,1)$ and $s(x) = g(x)$ for $\alpha = 1$. The decision threshold $\lambda \in \mathbb{R}$ and the randomization parameter $\tau \in [0,1]$ are implicitly defined by the distribution $(p_O(x), p_I(x,y))$ and the problem parameters $(\ell, \mathrm{tpr}_{\min}, \alpha)$.

Theorems 1 and 2 show that an optimal SCOD strategy (h^*, c^*) can be constructed from the Bayes ID classifier $h^* = h_B$ and a linear stochastic classifier (7) operating in 2D features space, where the first coordinate is the conditional risk $r(x) = \sum_{y\in\mathcal{Y}} p_I(y \mid x)\ell(y, h^*(x))$ and the second coordinate is the OOD/ID likelihood ratio $g(x) = p_O(x)/p_I(x)$. The slope of the linear classifier is determined by the user-defined parameters $(\alpha, \mathrm{tpr}_{\min})$. In general, the selector has to randomize with probability τ for boundary inputs $\mathcal{X}_{s(x)=\lambda} = \{x \in \mathcal{X} \mid s(x) = \lambda\}$. However, if $\mathcal{X}$ is continuous, the set $\mathcal{X}_{s(x)=\lambda}$ has in most cases the probability measure zero, and τ can be arbitrary, i.e., the deterministic $c^*(x) = [\![s(x) \leq \lambda]\!]$ is optimal. Note that Thm. 2 shows how to construct an *optimal selector for an arbitrary ID classifier h*, i.e., not only for the Bayes classifier h_B.

2.3 Relation to existing OODD and SCOD strategies

Single-score strategy Previous work on OODD focuses only on designing a good ID/OOD discriminator while ignoring the performance of the ID classifier on the accepted ID data. OODD methods output a single score $s\colon \mathcal{X} \to \mathbb{R}$ that is used to build a selector $c(x) = [\![s(x) \leq \lambda]\!]$. The goal is defined implicitly or explicitly as the Nayman-Pearson problem [30], and the performance of the selector c is usually evaluated by the ROC curve. Examples of single-score OOD methods involve the MSP score [15], ViM [41], GradNorm [28], etc. Our result shows that all existing single-score methods are not optimal, provided that one wants to solve the SCOD problem, which we verify experimentally in Sec. 6.

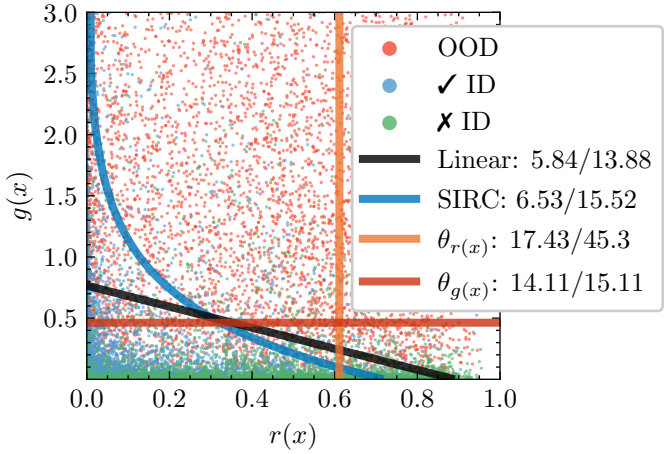

Fig. 1: Selectors in 2D space using scores: $\hat{r}(x)$, a softmax score from an ImageNet classifier, and $\hat{g}(x)$, learned OOD/ID likelihood ratio. OOD data is shown in red, correctly ✓, and incorrectly ✗, classified ID data is shown in blue/green. Selector parameters tuned for 90% TPR and minimal SCOD risk ($\alpha = 0.5$). Metrics: AuSRT ↓ / ScodRisk at TPR=90% ↓, details in Sec. 6.1.

Double-score strategy The SCOD problem we analyze in our paper was formulated in [42], which also proposed the Softmax Information Retaining Combination (SIRC). SIRC is a heuristic strategy to combine two scores $s_1 \colon \mathcal{X} \to \mathbb{R}$ and $s_2 \colon \mathcal{X} \to \mathbb{R}$ into a single one:

$$s_{\text{SIRC}}(x) = -(S_1^{max} - s_1(x))(1 + \exp(-b(s_2(x) - a)) \tag{8}$$

where S_1^{max} is an upper bound on $s_1(x)$, and a and b are hyper-parameters chosen based on a sample of ID data. The score s_{SIRC} is used to build a selector $c(x) = [\![s_{\text{SIRC}}(x) \leq \lambda]\!]$. The authors impose the following informal assumptions on the scores s_1 and s_2. The score s_1 should be i) higher for correctly classified ID samples and ii) lower for misclassified ID samples and OOD samples; in experiments, they set s_1 to the MSP score. The score s_2 should be lower for OOD data compared to ID data; in the experiments, s_2 score is either L_1-norm score [16] or the negative of the Residual score [41]. [42] claim, and we confirm their findings in Sec. 6, that the SIRC strategy performs very well in practice. However, our result shows that the SIRC strategy is not optimal when the scores s_1 and s_2 approach their ideal setting, i.e. when $s_1(x) = r(x)$ and $s_2(x) = g(x)$. Then, the linear combination (7) performs better, which we verify experimentally in Sec. 6. Figure 1 illustrates the proven optimal linear selector, SIRC, and two single-score selectors, functioning as binary classifiers in 2D space. The selectors are separating ImageNet samples as ID data and SSB_Hard as OOD data.

3 SCOD problem is not PAC learnable

In the previous section, we showed how to construct an optimal strategy for the SCOD problem, provided $p_I(y \mid x)$ and $g(x) = p_O(x)/p_I(x)$ are known. The next key question is: Can the SCOD problem be effectively solved using available data? We build on the findings in [10], which establish the non-learnability of the OODD problem in a distribution-free setting [1], particularly when the learning algorithm relies *solely on an ID data sample*. Unlike the SCOD problem, the OODD problem is formulated as a cost-based minimization problem [10], which requires cost assignment for all decision outcomes in the prediction strategy. Extending these insights, we demonstrate that in the distribution-free setting, the SCOD problem is also non-learnable when only ID data is accessible. To achieve this, we broaden the Probably Approximately Correct (PAC) learnability concept [33] for the SCOD problem.

Definition 2. (PAC learnability of SCOD problem) *Let $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function and $(\mathrm{tpr}_{\min}, \alpha) \in (0,1)^2$ user-defined parameters of the SCOD problem (4). A hypothesis class $\mathcal{H} \subset \{(h,c) \in \mathcal{Y}^{\mathcal{X}} \times [0,1]^{\mathcal{X}}\}$ [2] is PAC learnable if there exist a function $m\colon (0,1)^3 \to \mathbb{N}$ and a learning algorithm $A\colon \cup_{m=1}^{\infty}(\mathcal{X}\times\mathcal{Y})^m \to \mathcal{H}$ with the following property: For every $(\varepsilon_1, \varepsilon_2, \delta) \in (0,1)^3$, every OOD distribution $p_O\colon \mathcal{X} \to \mathbb{R}_+$, every ID distribution $p_I\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ and every $\pi_O \in [0,1]$, when running the algorithm A on $m \geq m(\varepsilon_1, \varepsilon_2, \delta)$ examples i.i.d. drawn from $p_I(x,y)$, the algorithm returns a selective classifier (h_m, c_m) such that with a probability of at least $1-\delta$ it holds that*

$$R(h_m, c_m) - R(h^*, c^*) \leq \varepsilon_1 \quad \textit{and} \quad \mathrm{tpr}(c_m) \geq \mathrm{tpr}_{\min} - \varepsilon_2 \,,$$

where $R(h,c) = (1-\alpha)\mathrm{R_S}(h,c) + \alpha\,\mathrm{fpr}(c)$ is the SCOD risk and (h^*, c^*) is an optimal SCOD strategy.

Def. 2 establishes PAC learnability, indicating the existence of an algorithm searching in a hypothesis space $\mathcal{H}$, which can discover an $(\varepsilon_1, \varepsilon_2)$-optimal solution for the SCOD problem (4) with an arbitrarily low probability of failure $\delta \in (0,1)$, given a sufficient number of ID data. Importantly, this guarantee is distribution-free, applying universally to every mixture (1) of ID and OOD data.

Theorem 3. *Let $\mathcal{H} \subset \{(h,c) \in \mathcal{Y}^{\mathcal{X}} \times [0,1]^{\mathcal{X}}\}$ be a non-trivial hypothesis space such that there exist two selective classifiers $(h_1, c_1) \in \mathcal{H}$ and $(h_2, c_2) \in \mathcal{H}$ for which $c_1 \neq c_2$. Then, the hypothesis space $\mathcal{H}$ is not PAC learnable in the sense of Definition 2.*

The implication from Theorem 3 is that, generally, achieving arbitrarily precise approximation of the optimal SCOD strategy using only ID data is unattainable

[1] Distribution-free setting implies learning guarantees for any data distribution.
[2] We use shortcuts $\mathcal{Y}^{\mathcal{X}} = \{h\colon \mathcal{X} \to \mathcal{Y}\}$, $[0,1]^{\mathcal{X}} = \{c\colon \mathcal{X} \to [0,1]\}$.

unless the hypothesis space is trivial [3]. Although our result is negative, it has an important implication: *Attempts to develop an efficient learning algorithm for a SCOD problem that does not make an assumption about the data distribution and uses only ID data are futile.*

4 Plugin estimate of the optimal SCOD strategy

In this section, we leverage the theoretical insights from preceding sections to introduce a method for learning a plugin estimate of the optimal SCOD strategy. We adopt the framework proposed by [18], where in addition to the ID data sample $\mathcal{T}_I = ((x_i^I, y_i^I) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, m)$ generated from $p_I(x, y)$ we also have an unlabeled sample of a mixture of ID and OOD data $\mathcal{T}_U = (x_i^U \in \mathcal{X} \mid i = 1, \ldots, n)$ generated from $p(x) = \pi_O^{\text{tr}}\, p_O(x) + (1 - \pi_O^{\text{tr}}) p_I(x)$. Collecting the data for the unlabeled mixture can, e.g., involve just recording the input samples from a real deployment of the predictor h. We use data $\mathcal{T}_I$ and $\mathcal{T}_U$ to learn a plugin estimate of the optimal SCOD strategy derived in Sec 2.2.

ID classifier We can use any method to train the ID classifier h from $\mathcal{T}_I$ that provides an estimate of the class posterior $p_I(y \mid x)$. In our experiments, we use a CNN with softmax decision layer trained by cross-entropy loss producing $\hat{p}_I(y \mid x)$. Then, we construct the plug-in Bayes ID classifier and conditional risk:

$$\hat{h}_B(x) \in \operatorname*{Argmin}_{y' \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \hat{p}_I(y \mid x) \ell(y, y') \;\text{ and }\; \hat{r}(x) = \sum_{y \in \mathcal{Y}} \hat{p}_I(y \mid x) \ell(y, \hat{h}_B(x)) \,. \quad (9)$$

Note that in case of 0/1-loss, $\ell(y, y') = [\![y \neq y']\!]$, $\hat{h}_B(x) = \operatorname{Argmax}_{y \in \mathcal{Y}} \hat{p}_I(y \mid x)$ is the standard MAP rule, and $\hat{r}(x) = 1 - \max_{y \in \mathcal{Y}} \hat{p}_I(y \mid x)$ is the 1-MSP rule [15].

Selector Once we have $\hat{r}$, it remains to estimate $g(x) = p_O(x)/p_I(x)$, in order to build the plugin estimator of the optimal selector $c^*(x)$ given by (7). We create a sequence $\mathcal{T}_{IU} = ((x_i, z_i) \in \mathcal{X} \times \{I, U\} \mid i = 1, \ldots, n + m)$ by randomly re-shuffling a concatenation of $\mathcal{T}_I$ and $\mathcal{T}_U$, and setting $z_i = I$ when x_i is from $\mathcal{T}_I$ and $z_i = U$ when x_i is from $\mathcal{T}_U$. The sequence $\mathcal{T}_{IU}$ can be seen as a random sample from a mixture

$$p(x, z) = \begin{cases} \pi_U \left(p_O(x)\, \pi_O^{\text{tr}} + p_I(x) \left(1 - \pi_O^{\text{tr}}\right) \right) & \text{if} \quad z = U \\ (1 - \pi_U)\, p_I(x) & \text{if} \quad z = I \end{cases} \quad (10)$$

where $\pi_U = n/(m + n)$ is the *known portion* of $\mathcal{T}_U$ in $\mathcal{T}_{IU}$. It follows directly from (10) that the desired OOD/ID likelihood ratio reads

$$g(x) = \frac{p_O(x)}{p_I(x)} = \frac{p(z = U \mid x)}{p(z = I \mid x)} \frac{1 - \pi_U}{\pi_U \pi_O^{\text{tr}}} - \frac{1 - \pi_O^{\text{tr}}}{\pi_O^{\text{tr}}} \,. \quad (11)$$

[3] The trivial hypothesis space involves a single selector, reducing the SCOD problem to standard prediction under the closed-world assumption - known to be learnable when $\mathcal{H}$ has finite complexity.

We propose to approximate the unknown $p(z \mid x)$ by a *corrected sigmoid model* (CSM) $p(z \mid x; \boldsymbol{\theta}, a)$, the use of which is motivated by Theorem 4.

Theorem 4. *Let $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^d$ be a feature map. Assume that the features $\boldsymbol{\phi}(x)$ computed on the ID and OOD data are normally distributed, i.e, $p_I(x; \boldsymbol{\mu}_I, \boldsymbol{C}) = \mathcal{N}(\boldsymbol{\phi}(x); \boldsymbol{\mu}_I, \boldsymbol{C})$ and $p_O(x; \boldsymbol{\mu}_O, \boldsymbol{C}) = \mathcal{N}(\boldsymbol{\phi}(x); \boldsymbol{\mu}_O, \boldsymbol{C})$. Then, the posterior $p(z \mid x)$ derived from the distribution (10) is an element of $\mathcal{P} = \{p(z \mid x; \boldsymbol{\theta}, a) \mid a = \pi_U(1 - \pi_O^{\text{tr}})/(1 - \pi_U), \boldsymbol{\theta} \in \mathbb{R}^{d+1}\}$ where*

$$p(z = I \mid x; \boldsymbol{\theta}, a) = \frac{1}{1 + |a| + \exp(\boldsymbol{\theta}^T[\boldsymbol{\phi}(x); 1])}\,, \tag{12}$$

is the corrected sigmoid.

We estimate the parameters $(\boldsymbol{\theta}, a)$ of CSM by the Maximum Likelihood (ML) method, which corresponds to minimizing the binary cross-entropy (BCE) of the proposed CSM (12) on the sequence $\mathcal{T}_{IU}$. Let $(\hat{\theta}, \hat{a})$ be the parameters estimated from $\mathcal{T}_{IU}$. By Theorem 4, the unknown π_O^{tr} can be recovered from the parameter $\hat{a}$ by $\hat{\pi}_O^{\text{tr}} = 1 + |\hat{a}| - |\hat{a}|/\pi_U$. Finally, we substitute $p(z \mid x; \hat{\boldsymbol{\theta}}, \hat{a})$ and $\hat{\pi}_O^{\text{tr}}$ into formula (11) to obtain an estimate of the OOD/ID likelihood ratio $\hat{g}$, and use it to obtain a plugin estimate of the optimal score (7), that is,

$$\hat{s}(x) = \hat{r}(x) + \frac{\alpha\,\text{tpr}_{\min}}{1 - \alpha}\hat{g}(x) \quad \text{where} \quad \hat{g}(x) = \frac{p(z = U \mid x; \hat{\boldsymbol{\theta}}, \hat{a})}{p(z = I \mid x; \hat{\boldsymbol{\theta}}, \hat{a})}\frac{1 - \pi_U}{\pi_U \hat{\pi}_O^{\text{tr}}}\,. \tag{13}$$

In the formula for $\hat{g}$ we omit the additive term present in (11), as it is absorbed by the decision threshold λ of the linear selector (7). The value of λ is adjusted on $\mathcal{T}_I$ to achieve the target $\text{tpr}(c) = \text{tpr}_{\min}$. Algo. 1 summarizes the proposed method to learn the Plugin estimate of the Optimal SCOD strategy (POSCOD).

Algorithm 1 POSCOD

Require: ID data $\mathcal{T}_I = ((x_i^I, y_i^I) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, m)$, unlabeled mixture of ID and OOD $\mathcal{T}_U = (x_i^U \in \mathcal{X} \mid i = 1, \ldots, n)$, problem parameters $(\alpha, \text{tpr}_{\min}) \in (0, 1)^2$, target loss $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.

Ensure: Selective classifier $(\hat{h}, \hat{c})$ for the SCOD problem, Def. 1.

1: Train $\hat{p}_I(y \mid x)$ on $\mathcal{T}_I$ using the cross-entropy loss.
2: Construct the plugin Bayes ID classifier $\hat{h}$ and its conditional risk $\hat{r}$ by (9).
3: Create $\mathcal{T}_{IU} = ((x_i, z_i) \in \mathcal{X} \times \{I, U\} \mid i = 1, \ldots, n + m)$ by randomly re-shuffling inputs from $\mathcal{T}_I$ and $\mathcal{T}_U$.
4: Train $\hat{\boldsymbol{\theta}}$ and $\hat{a}$ by minimizing the BCE of the corrected sigmoid (12) on $\mathcal{T}_{IU}$.
5: Compute $\pi_U = n/(n + m)$ and $\hat{\pi}_O^{\text{tr}} = 1 + |\hat{a}| - \frac{|\hat{a}|}{\pi_U}$.
6: Construct the selector score $\hat{s}$ by (13).
7: Tune λ of $\hat{c}(x) = [\![\hat{s}(x) \le \lambda]\!]$ on $\mathcal{T}_I$ to achieve the target $\text{tpr}(\hat{c}) = \text{tpr}_{\min}$.

Computational requirements and implementation POSCOD converts the learning process for the SCOD strategy into two steps. First, it involves training

the ID classifier by minimizing the standard cross-entropy loss. Second, it includes learning a binary classifier by minimizing the BCE of the corrected sigmoid. Implementing this in Pytorch requires modifying two lines of code with the BCE of the standard sigmoid. Notably, POSCOD does not use hyperparameters.

Assumptions POSCOD relies on two assumptions. Firstly, the CSM in Eq. (12) should effectively represent $p(z \mid x)$. Secondly, the OOD prior π_O^{tr} needs to be learnable. Note that ID and *OOD are not required to be transformable to normal distributions*. However, when they are, Thm. 4 guarantees that CSM is exact. Notably, the unknown *OOD prior* π_O^{tr} *in the training sample* $\mathcal{T}_U$ *does not need to match the OOD prior* π_O *in the test sample* because the SCOD problem (4) is independent of π_O; the estimate of π_O^{tr} is used only to adjust the scale of the learned likelihood ratio through (11). To ensure the learnability of π_O^{tr}, the OOD must be a proper novelty distribution with respect to ID, as defined in Def. 4 and Prop. 5 in [2]. OOD is considered proper if it cannot be decomposed into a mixture of ID and any other distribution on $\mathcal{X}$. This assumption is practical and is typically satisfied in real-world scenarios. The suitability of CSM as a proxy for $p(z \mid x)$ should be experimentally validated for specific datasets, a validation we demonstrate in various benchmarks.

Relation to existing work In prior research [18, 27], the standard sigmoid was used to model $p(z \mid x)$. That is, they optimize $\boldsymbol{\theta}$ to fit $p(z \mid x; \boldsymbol{\theta}, a)$ to $\mathcal{T}_{IU}$ while keeping a fixed at 0. Using the standard sigmoid has two drawbacks: i) the model is incorrect because $a = \pi_U(1 - \pi_O^{\text{tr}})/(1 - \pi_U) = 0$ only if $\mathcal{T}_U$ contains clean OOD data, i.e., $\pi_O^{\text{tr}} = 1$; ii) it does not estimate π_O^{tr} (needed to compute $\hat{g}(x)$).

5 Relation to existing literature

Recent insights emphasize the necessity, when designing selectors for ID classifiers in OOD settings, to reject not only OOD samples but also ID samples prone to misclassification. This problem has been termed Unknown Detection [19], Unified Open-Set Recognition [3], and Selective Classification in the presence of Out-of-Distribution (SCOD) data [42]. [42] formally defined the SCOD problem and introduced Softmax Information Retaining Combination (SIRC), a method tailored for SCOD. We analyze the SCOD problem and demonstrate that both existing OODD methods and SIRC deviate from the optimal SCOD strategy that we derived in our paper. We introduce a novel method, POSCOD, learning a plugin estimate of the optimal SCOD strategy and empirically show that it outperforms both SIRC and existing OODD methods.

The PAC learnability of OODD was examined in [10], defining the optimal OODD strategy as an unconstrained cost-based minimization problem. Their results demonstrate that OODD is not PAC-learnable in the distribution-free setting when the learning algorithm relies solely on an ID data sample. We extend their findings and establish that, in the distribution-free setting, SCOD is also not PAC-learnable when only ID data is available.

The work of [18] introduced a method to learn an OODD from an unlabeled mixture of ID and OOD data. Their formulation of the optimal strategy involves constrained optimization, different from the SCOD problem analyzed in our paper. Thanks to the known form of the optimal strategy derived in our paper, our proposed method, POSCOD, simplifies the learning process by i) training an ID classifier using standard cross-entropy loss and ii) training a classifier via the BCE of the novel CSM.

Our work aligns with [27], establishing an optimal strategy for a different SCOD formulation and proposing a plugin estimator. Key distinctions include i) our TPR-constrained SCOD formulation removing depenence on the unknown OOD portion in the test sample (c.f. Appendix A), ii) a novel technique for finding the optimal strategy applicable to all distributions, iii) providing an explicit formula to compute parameters of the optimal strategy and iv) using the CSM instead of the standard sigmoid to model the OOD/ID likelihood ratio.

6 Experiments

In this section, we empirically validate the theoretical results presented in Theorem 2, confirming that the optimal SCOD selector indeed constitutes a linear combination of $g(x)$ and $r(x)$. Additionally, we show that our proposed method, POSCOD, built on the plugin estimate of the optimal SCOD strategy, see Sec. 4, surpasses the current state-of-the-art on real-world datasets. The implementation of the experiments is accessible from: `https://github.com/xfrancv/SCOD`.

6.1 Evaluation metrics

Let $\mathcal{T} = \mathcal{T}_I \cup \mathcal{T}_O$ be a sample of evaluation data where $\mathcal{T}_I = ((x_i^I, y_i^I) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \dots, m)$ is a sample of i.i.d. ID from $p_I(x, y)$ and $\mathcal{T}_O = ((x_i^O, \emptyset) \in \mathcal{X} \times \bar{\mathcal{Y}} \mid i = 1, \dots, n)$ is a sample of i.i.d. OOD from $p_O(x)$. Let $s\colon \mathcal{X} \to \mathbb{R}$ be a score, $c(x; \lambda) = [\![s(x) \leq \lambda]\!]$ a selector and $h\colon \mathcal{X} \to \mathcal{Y}$ an ID classifier. To summarize the performance of a selective classifier (h, c) on the evaluation data $\mathcal{T}$, we employ the *Area Under SCOD Risk - True positive rate* (AuSRT $\downarrow$) curve which evaluates the overall performance of the selective classifier on the SCOD problem, see Def. 1. For any variable x, let $\hat{x}$ be its empirical estimate. The SCOD risk (refer to Eq. (4)) at a given TPR is estimated by $\widehat{\mathrm{R}}(\mathrm{tpr}_{\min}; h, c) = \min_\lambda (1 - \alpha)\, \widehat{\mathrm{R}}_{\mathrm{S}}(h, c) + \alpha\, \widehat{\mathrm{fpr}}(c)$ s.t. $\widehat{\mathrm{tpr}}(c) \geq \mathrm{tpr}_{\min}$. The AuSRT is then computed as the area under the curve given by the points $\big\{ \big(\widehat{\mathrm{R}}(\mathrm{tpr}_{\min}), \mathrm{tpr}_{\min} \big) \mid \mathrm{tpr}_{\min} \in (0, 1] \big\}$. See Appendices D to F for more metrics.

6.2 Experimental setup

Datasets We assess SCOD performance on datasets adopted from the OpenOOD benchmark [43]. Detailed description of the datasets is provided in Appendix B.1.

Models As the ID classifier $h(x)$ on CIFAR-10/100, we use pre-trained ResNet-18 from the OpenOOD benchmark [43]. For $h(x)$ on ImageNet-1K, we use a pre-trained ResNet-50 model from Torchvision [37]. For details, see Appendix B.2.

Learning the likelihood ratio To estimate the OOD/ID likelihood ratio $g(x)$ on real-world data, we train a classifier with BCE of the standard sigmoid (a.k.a logistic regression) and a *corrected sigmoid* model (refer to Sec. 4). We reuse the feature representation of the ID classifier $h(x)$ to learn the likelihood ratio. Up to 50% of OOD samples are reserved for training, and evaluations are done exclusively on unseen samples. Refer to Appendix B.3 for more details.

Double-score parameter setting strategies As mentioned in Sec. 2.3, the current state-of-the-art employs a nonlinear combination $s_{\text{SIRC}}(x)$ of scores $s_1(x)$ and $s_2(x)$ with hyperparameters a and b. In contrast, our proposed approach utilizes a linear combination $s_{\text{Linear}}(x) = s_1(x) + \beta s_2(x)$ with a single hyperparameter β. Two key questions arise when comparing the approaches: Given the optimal parameters a, b, β, which model performs better? More importantly, which model performs better in a scenario when the optimal parameters are unknown? To answer the first question, we search for the best hyperparameters by optimizing the SCOD risk on the test set and refer to scores using the parameters as *tuned*. The *tuned* scores can not be used in practice, however, they provide an upper bound on the model's performance. For practical deployment of s_{SIRC}, [42] offers a heuristic to set the parameters using the empirical mean μ_{s_2} and standard deviation σ_{s_2} of the score $s_2(x)$ on ID data. For the linear strategy, we derive an explicit formula, see Eq. (7). We refer to scores using these parameters as *plugin* and they can be used in practice without test data tuning. We evaluate both *tuned* and *plugin* settings. Details on the tuning can be found in Appendix B.4.

6.3 Experimental validation of theoretical results

To validate Thm. 2, we approximate the likelihood ratio $g(x)$ on real-world data using a logistic regression model trained on features of the ID classifier $h(x)$. This model is impossible to obtain in practice, given the rarity of clean OOD data samples. Nonetheless, it provides the best available approximation of the true likelihood ratio $g(x)$. For the linear strategy, we employ the plug-in conditional risk $\hat{r}(x) = 1 - \text{MSP}$ as the score $s_1(x)$ and use the approximated $\hat{g}(x)$ as the score $s_2(x)$. In the case of SIRC, we use $1 - \hat{r}(x)$ and $-\hat{g}(x)$ as scores to meet the assumptions outlined in Sec. 2.3. The performance of the selectors on ImageNet using these scores is summarized in Tab. 1. The linear strategy consistently outperforms both SIRC and the baseline. The linear strategy with $\hat{r}(x)$ and $\hat{g}(x)$ also surpasses all other combinations of scoring functions employed by contemporary OOD detectors [9, 14–16, 24, 25, 35, 41]. This dominance holds true across all ID datasets, showcasing substantial advantages in both tuned and plugin setups. For detailed results on all datasets, refer to Appendices D to F.

6.4 The POSCOD algorithm

To show the practicality of our results without requiring a clean OOD data sample, we employ a framework using an unlabeled mixture of ID and OOD data, as detailed in Sec. 4. We demonstrate that the linear strategy combining $\hat{r}(x)$ and

Table 1: AuSRT↓ in % for selective classifiers (h, c) from ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$. Results on ID ImageNet and various OOD datasets. Rows represent different scores $s(x)$, showcasing performance for i) tuned, and ii) plugin double-score methods. Red rows use $\hat{g}(x)$ learned on clean OOD data. Relative cost: $\alpha = 0.5$.

Score \ *Dataset*	ssb_hard	ninco	inaturalist	textures	openimage_o
Tuned SIRC	6.59	**6.61**	4.78	7.45	6.21
Tuned Linear	**5.88**	6.67	**4.04**	**4.23**	**5.82**
Plugin SIRC	13.24	10.94	6.39	10.49	8.59
Plugin Linear	6.10	8.05	4.35	4.48	7.37
MSP [15]	17.47	13.59	9.35	12.15	11.05
Likelihood Ratio	14.14	14.04	11.22	11.75	14.00

$\hat{g}(x)$ learned by the proposed POSCOD Algorithm 1 consistently outperforms SIRC and contemporary OOD scoring functions [9, 14–16, 24, 25, 35, 41] across all ID datasets. Additionally, our results reveal that utilizing the standard sigmoid to approximate the likelihood ratio $g(x)$, as done in prior work [18, 27], leads to suboptimal performance. Results for ID ImageNet using the scores $\hat{r}(x)$ and $\hat{g}(x)$ are shown in Tab. 2. In Appendix C, we show that the results hold for a wide range of priors π_O^{tr} and relative costs α. For comprehensive results across all datasets and compared scores, refer to Appendices D to F.

Table 2: Comparison of double-score strategies using logistic regression with standard and corrected sigmoid for approximating the likelihood ratio (LR) $g(x)$ from a mixture of data with $\pi_O^{\text{tr}} = 1/2$ (see Sec. 4). Table displays AuSRT↓ in % for selective classifiers (h, c) from an ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$. Results show practically usable plugin double-score strategies on ID ImageNet. Relative cost: $\alpha = 0.5$.

Sigmoid	*Score* \ *Dataset*	ssb_hard	ninco	inaturalist	textures	openimage_o
Corrected	SIRC	14.27	12.67	8.02	10.50	9.75
	Linear	**6.37**	**7.44**	**4.16**	**4.38**	**6.48**
	LR	14.25	16.67	12.70	13.19	15.65
Standard	SIRC	11.27	9.28	4.34	5.94	7.22
	Linear	7.28	8.83	5.15	5.33	7.22
	LR	13.88	15.21	11.32	11.54	13.57

Comparison with contemporary scoring functions Our results demonstrate that when provided with reasonable estimates $\hat{r}(x)$, $\hat{g}(x)$ of the optimal scores, the linear strategy is a state-of-the-art SCOD selector. However, we refrain from making this claim with currently employed OODD scoring functions. In Tab. 3 we show the performance of SIRC and the linear strategy on ImageNet when combining contemporary scores. Table 3 only shows a subset of the evaluated scores; for SIRC, the best-performing score, and scores [16, 41] mirroring the orig-

inal setup in [42]. For the linear double-score strategy, the three best-performing combinations are shown. In some cases, SIRC outperforms the linear strategy when using contemporary OODD scores. However, with the POSCOD estimate $\hat{g}(x)$, the linear strategy outperforms all other methods by a large margin. For results with all scores on all datasets, refer to Appendices D to F.

Table 3: AuSRT↓ in % points for practically usable strategies. The table contains single-score strategies that when combined with MSP achieve the best results. The likelihood ratio (LR) $\hat{g}(x)$ was approximated by POSCOD. In some cases, SIRC outperforms the linear strategy. When using the plugin LR estimate, the linear strategy significantly outperforms SIRC and all single-score strategies. The best results with contemporary OODD scores are shown in bold. The best results overall are highlighted in green.

	Score \ *Dataset*	ssb_hard	ninco	inaturalist	textures	openimage_o
Single Score	ASH [9]	19.04	13.76	6.95	7.03	8.85
	GradNorm [16]	23.95	22.89	12.95	13.88	17.49
	L_1-norm [16]	32.27	36.06	39.22	28.37	32.86
	ODIN [24]	19.79	16.78	10.08	11.16	11.54
	ReAct [35]	18.88	14.53	7.23	9.00	9.46
	Residual [41]	40.72	35.98	37.03	18.59	31.93
Linear	ASH [9]	18.74	13.47	6.82	**6.95**	**8.66**
	ODIN [24]	17.40	13.51	9.27	12.21	11.02
	ReAct [35]	17.85	13.37	**6.76**	8.62	8.77
	POSCOD LR	6.37	7.44	4.16	4.38	6.48
SIRC	GradNorm [16]	**16.99**	**12.96**	7.70	10.24	9.80
	L_1-norm [16]	17.21	13.54	9.50	11.90	10.96
	Residual [41]	17.49	13.43	9.20	9.54	10.62
	POSCOD LR	14.27	12.67	8.02	10.50	9.75

7 Conclusions

This study addresses the SCOD problem [42], focusing on scenarios in which ID test samples are contaminated by OOD data. Our key contributions include demonstrating that the optimal SCOD strategy involves a Bayes classifier for ID data and a selector corresponding to a stochastic linear classifier in a 2D space. This contrasts with suboptimal strategies used in contemporary OOD detection methods and SIRC [42], the current state-of-the-art on the SCOD problem. We establish the non-learnability of SCOD in a distribution-free setting when relying solely on an ID data sample. This result highlights the inherent challenges of PAC learning for SCOD without access to OOD data. We introduced POSCOD, a method for learning the plugin estimate of the optimal SCOD strategy from both an ID data sample and an unlabeled mixture of ID and OOD data. Empirical validations confirm our theoretical findings and demonstrate that POSCOD outperforms existing OODD methods and SIRC in solving the SCOD problem.

Acknowledgments

The authors acknowledge support for this work from the CTU institutional support (Future Fund).

References

1. Bitterwolf, J., Mueller, M., Hein, M.: In or out? fixing imagenet out-of-distribution detection evaluation. In: ICML (2023), https://proceedings.mlr.press/v202/bitterwolf23a.html
2. Blanchard, G., Lee, G., Scott, C.: Semi-supervised novelty detection. Journal of Machine Learning Research (2010)
3. Cen, J., Luan, D., Zhang, S., Pei, Y., Zhang, Y., Zhao, D., Shen, S., Che, Q.: The devil is in the wrongly-classified samples: towards unified open-set recognition. In: ICLR (2023)
4. Chen, G., Peng, P., Wang, X., Tian, Y.: Adversarial reciprocal points learning for open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8065–8081 (2022)
5. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory **16**(1), 41–46 (1970)
6. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3606–3613 (2014). https://doi.org/10.1109/CVPR.2014.461
7. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint:1802.04865 (2018)
8. Dhamija, A.R., Günther, M., Boult, T.: Reducing network agnostophobia. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
9. Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely simple activation shaping for out-of-distribution detection. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=ndYXTEL6cZz
10. Fang, Z., Li, Y., Lu, J., Dong, J., Han, B., Liu, F.: Is out-of-distribution detection learnable? In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 37199–37213. Curran Associates, Inc. (2022)
11. Franc, V., Prusa, D., Voracek, V.: Optimal strategies for reject option classifiers. Journal of Machine Learning Research **24**(11), 1–49 (2023)
12. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems 30. pp. 4878–4887 (2017)
13. Granese, F., Romanelli, M., Gorla, D., Palamidessi, C., Piantanida, P.: Doctor: A simple method for detecting misclassification errors. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 5669–5681. Curran Associates, Inc. (2021)
14. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 8759–8773. PMLR (Jul 2022)

15. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proceedings of International Conference on Learning Representations (2017)
16. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. In: Advances in Neural Information Processing Systems (2021)
17. Huang, R., Li, Y.: Mos: Towards scaling out-of-distribution detection for large semantic space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8710–8719 (June 2021)
18. Katz-Samuels, J., Nakhleh, J., Nowak, R., Li, Y.: Training ood detectors in their natural habitats. In: ICML (2022)
19. Kim, J., Koo, J., Hwang, S.: A unified benchmark for the unknown detection capability of deep neural networks. Expert Syst. Appl. **229**, 120461 (2021), https://api.semanticscholar.org/CorpusID:244773165
20. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep. (2009)
21. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (Mar 2020). https://doi.org/10.1007/s11263-020-01316-z, http://dx.doi.org/10.1007/s11263-020-01316-z
22. Le, Y., Yang, X.S.: Tiny imagenet visual recognition challenge (2015)
23. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
24. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (2018)
25. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 21464–21475. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf
26. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
27. Narasimhan, H., Krisna Menon, A., Jitkrittum, W., Kumar, S.: Plugin estimators for selective classification with out-of-distribution detection. arXiv preprint:2301.12386v4 (2023)
28. Neal, L., Olson, M., Fern, X., Wong, W.K., Li, F.: Open set learning with counterfactual images. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
29. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011 (2011), http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf
30. Neyman, J., Person, E.: On the use and interpretation of certain test criteria for purpose of statistical inference. Biometrica pp. 175–240 (1928)

31. Pietraszek, T.: Optimizing abstaining classifiers using ROC analysis. In: Proceedings of the 22nd International Conference on Machine Learning. p. 665–672 (2005)
32. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021), https://openreview.net/forum?id=Zkj_VcZ6ol
33. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, USA (2014)
34. Song, Y., Sebe, N., Wang, W.: Rankfeat: Rank-1 feature removal for out-of-distribution detection. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 17885–17898. Curran Associates, Inc. (2022)
35. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 144–157. Curran Associates, Inc. (2021)
36. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 20827–20840. PMLR (Jul 2022)
37. TorchVision maintainers and contributors: Torchvision: Pytorch's computer vision library. GitHub repository (2016), https://github.com/pytorch/vision
38. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(11), 1958–1970 (2008). https://doi.org/10.1109/TPAMI.2008.128
39. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
40. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: a good closed-set classifier is all you need? In: International Conference on Learning Representations (2022)
41. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4911–4920 (2022)
42. Xia, G., Bouganis, C.S.: Augmenting softmax information for selective classification with out-of-distribution data. In: Proceedings of the Asian Conference on Computer Vision (ACCV). pp. 1995–2012 (December 2022)
43. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., Du, X., Zhou, K., Zhang, W., Hendrycks, D., Li, Y., Liu, Z.: Openood: Benchmarking generalized out-of-distribution detection. In: Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmar (2022)
44. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)