Supplement

This supplement contains additional clarifications of SSIF as well as extra details about our experiments. Specifically, we first motivate why our feature-selection distribution targets features with high-quality splits (Sec 7.1). Then, we describe the datasets used in our experimental analysis and further comment on Q1 and Q2 through a more fine-grained view of the results (Sec 7.2). Finally, we discuss two additional research questions and show that (1) choosing the max value of the split and feature selection distributions results in worse performance, and that (2) SSIF is able to cope with data distributions where IF struggles, commonly known as IF's blind spot (Sec 7.3).

1 Justification of the chosen feature-selection distribution

In this section, we show that our feature-selection distribution assigns high probabilities to features that have more informative split distributions (i.e., features that are more likely to isolate anomalous instances). Let us assume that some features yield truncated normal split distributions $\mathcal{S}|X_{\mathcal{K}} \sim \mathcal{N}(t, \sigma^2_{\mathcal{K}})$ (truncated to $[\min X_{\mathcal{K}}, \max X_{\mathcal{K}}]$) for some constant mean value $t \in (\min X_{\mathcal{K}}, \max X_{\mathcal{K}})$. Intuitively, the lower the variance $\sigma^2_{\mathcal{K}}$, the higher the peak of $p(\mathcal{S}|X_{\mathcal{K}})$, the more informative the feature. We illustrate that this holds in the next proposition.

Proposition 1.1. *Let $\mathcal{S}|X_1, \ldots, \mathcal{S}|X_M$ be M random variables such that, for $k \le M$, $\mathcal{S}|X_k \sim \mathcal{N}(t, \sigma^2_k)$ follows a truncated normal distribution between $[\mathcal{A}, \mathcal{B}]$ for any $t \in [\mathcal{A}, \mathcal{B}]$. Then,*

$$\sigma^2_{k_1} > \sigma^2_{k_2} \implies \mathrm{KL}\left(\mathcal{S}|X_{k_1} \| V\right) < \mathrm{KL}\left(\mathcal{S}|X_{k_2} \| V\right)$$

for any $k_1, k_2 \le M$, which implies $p(\mathcal{K}=k_1) < p(\mathcal{K}=k_2)$, i.e. the feature k_1 has lower probability than k_2.

Proof. For any $k \le M$,

$$\mathrm{KL}\left(\mathcal{S}|X_k \| V\right) = \mathbb{E}_{p(\mathcal{S}|X_k)}\left[\ln p(\mathcal{S}|X_k)\right] - \mathbb{E}_{p(\mathcal{S}|X_k)}\left[\ln p(V)\right] = -\ln(\sqrt{2\pi e}\sigma_k Z_k) - \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{2Z_k} - \mathrm{c}$$

where ϕ, Φ are the density and cumulative of a $\mathcal{N}(0,1)$ variable such that $Z_k = \Phi(\beta_k) - \Phi(\alpha_k)$, with $\alpha_k = \frac{\mathcal{A}-t}{\sigma_k}$ and $\beta_k = \frac{\mathcal{B}-t}{\sigma_k}$. In the first step, we use the alternative definition of KL divergence, while the second step uses that $\mathbb{E}_{p(\mathcal{S}|X_k)}\left[\ln p(V)\right]$ is constant because $p(V)$ is constant, and derives the formula for $\mathbb{E}_{p(\mathcal{S}|X_k)}\left[\ln p(\mathcal{S}|X_k)\right]$ from [**?**]. Finally, for $\sigma^2_{k_1} > \sigma^2_{k_2}$, we get $\mathrm{KL}\left(\mathcal{S}|X_{k_1} \| V\right) < \mathrm{KL}\left(\mathcal{S}|X_{k_2} \| V\right)$ because $\mathrm{KL}\left(\mathcal{S}|X_k \| V\right)$ is decreasing over σ_k. □

2 Further details about experiments (Q1 and Q2).

2.1 Setup

Data. Table 1 illustrates the characteristics (number of instances, number of features, and contamination factor) of the 20 datasets used for the empirical evaluation of our method.

2.2 Results

Q1: Figure 6 shows how the test set AUROC, averaged over all iterations, varies as a function of the percentage of labeled instances c for each method and dataset. SSIF is clearly the best performing in 4 datasets (Annthyroid, Arrhytmia, HeartDisease and SpamBase), while SSDO outperforms all the other methods in another dataset (T06). For all the remaining datasets, there is no clear winner.

Q2: Table 2 shows SSIF and IF test set AUROC, averaged over all iterations, for each dataset in an unsupervised setting. Overall, SSIF performs slightly better than IF at a cost of a higher computational complexity since the unlabeled component has to be estimated for different features.

Dataset	n	M	Contamination
ALOI	1000	27	0.030
Annthyroid	1000	21	0.075
Arrhythmia	256	259	0.047
Cardiotocography	1000	21	0.220
Cover	1000	10	0.010
HeartDisease	157	13	0.045
Letter	1000	32	0.063
PageBlocks	1000	10	0.095
Pima	555	8	0.099
Shuttle	1000	9	0.013
SpamBase	1000	57	0.050
Stamps	340	9	0.091
WBC	454	9	0.022
WDBC	367	30	0.027
T01	1000	79	0.024
T06	1000	79	0.084
T07	1000	79	0.031
T11	1000	79	0.005
T15	4399	10	0.072
T21	1967	10	0.058

Table 1: Real-world and benchmark anomaly detection datasets used for the experiments.

3 Two additional research questions (Q4 and Q5).

Q4: SSIF vs. MaxSSIF. Because supervised tree-based models usually take the max value of the split and feature-selection functions, we test how using the max of $p(\mathcal{S}|X_{\mathcal{K}})$ and $p(\mathcal{K})$ performs in comparison to our sampling approach. We call it MaxSSIF. Figure 7 shows the average AUROC as a function of the percentage of labels c for 5 representative datasets. Overall, the probabilistic version of SSIF performs better or similar than MaxSSIF on 4 datasets out of 5. Moreover, SSIF has a stable performance: when increasing c, SSIF's AUROC varies with limited oscillation, which indicates that SSIF is robust against random effects due to the sampled labels. On the contrary, MaxSSIF often obtains different AUROCs for similar c values (e.g., its AUROC is < 0.5 for $c = 30\%$ and > 0.9 for $c = 35\%$). This is due to the bias introduced by the max-function, which makes MaxSSIF suffer from the randomness of the label selection.

Q5: SSIF performance considering the IF blind spot drawback. One of the most know IF's drawback is the well-known "blind-spot" that is when anomalous instances are surrounded by normal data. In this case, IF can not isolate easily the anomalies and therefore the performance is poor. If some labeled anomalies are available, SSIF overcomes this issue thanks (i) to the labeled component (Sec. 4.1.1), that selects at each step the best split value s to isolate anomalies as soon as possible, and (ii) to the leaf label scaling (Sec. 4.3), that increases the path length of the surrounding normal instances and decreases it for the anomalies.

To show this, we perform experiments to compare SSIF to IF on a 2D toy donut dataset (Figure 8), where the anomalous instances are located in the middle of the donut. We consider three different settings: (a) both labels are available, (b) only normal instances are available, and (c) only anomalous instances are available. For each setting, different percentages of anomalous instances are considered ($c = [1\%, ..., 5\%]$): we consider very low labeled percentages to show that SSIF can significantly outperform IF also when only very few labels are available. For each setting and percentage, 10 iterations are performed to alleviate random effects. Figure 9 shows the results. SSIF outperforms IF in all settings disregarding the percentage of labels provided and achieves significantly higher AUROC. We observe also that setting (c), when only anomalies are present in the training set, outperforms the other two setting: in this case SSIF employs ah higher number of anomalies during the tree construction phase to bias stronger the selected split values.

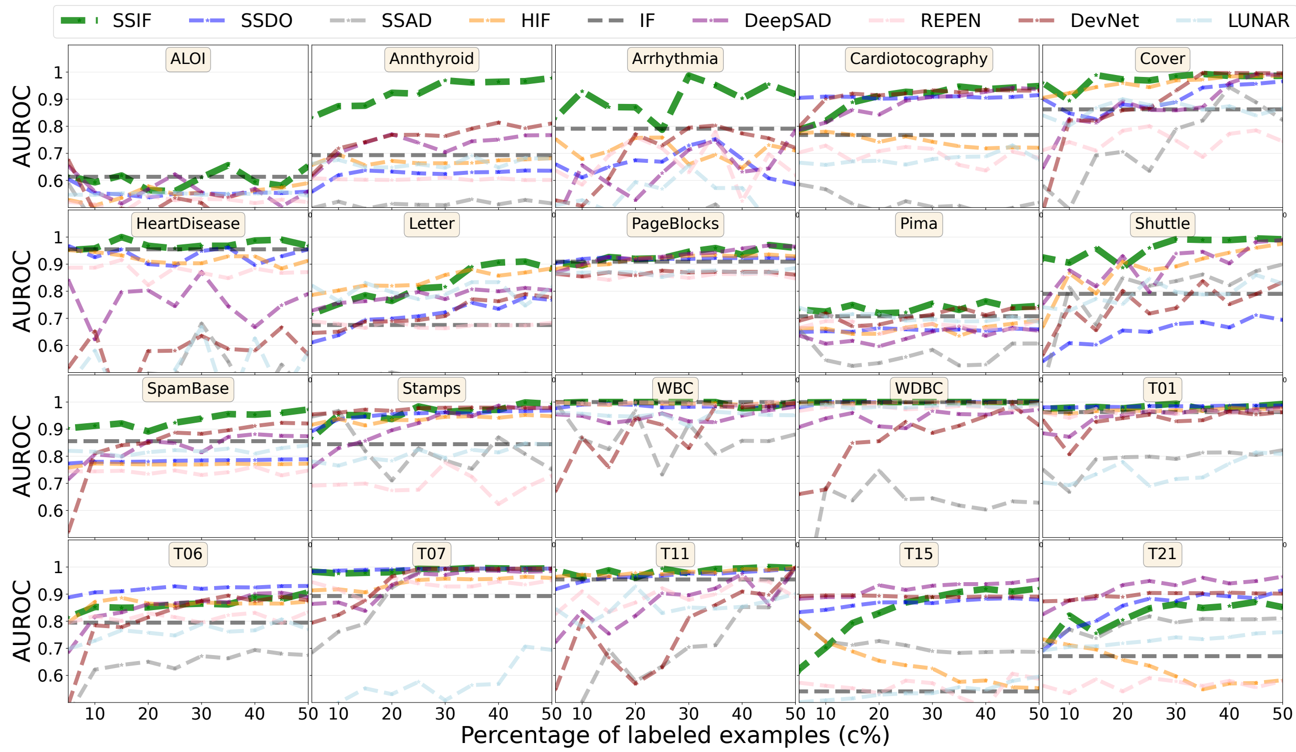

Figure 6: Each dataset's average test set AUROC as a function of the number of days labeled by the user, shown for each method.

Figure 7: Average AUROC as a function of the percentage of labels ($c\%$) for our probabilistic model SSIF and its max-value version MaxSSIF.

Dataset	AUROC(SSIF)	AUROC(IF)
ALOI	0.60 ± 0.02	**0.61 ± 0.03**
Annthyroid	**0.80 ± 0.01**	0.69 ± 0.02
Arrhythmia	**0.83 ± 0.05**	0.79 ± 0.07
Cardiotocography	0.67 ± 0.01	**0.77 ± 0.03**
Cover	**0.89 ± 0.01**	0.86 ± 0.06
HeartDisease	**0.97 ± 0.00**	0.95 ± 0.02
Letter	**0.70 ± 0.01**	0.68 ± 0.02
PageBlocks	0.88 ± 0.01	**0.91 ± 0.01**
Pima	**0.72 ± 0.02**	0.71 ± 0.02
Shuttle	0.78 ± 0.03	**0.79 ± 0.03**
SpamBase	**0.87 ± 0.01**	0.86 ± 0.03
Stamps	**0.88 ± 0.02**	0.84 ± 0.03
WBC	1.00 ± 0.00	1.00 ± 0.00
WDBC	1.00 ± 0.00	1.00 ± 0.00
T01	0.96 ± 0.01	0.96 ± 0.01
T06	**0.82 ± 0.02**	0.79 ± 0.02
T07	0.89 ± 0.03	0.89 ± 0.04
T11	**0.97 ± 0.02**	0.95 ± 0.03
T15	0.48 ± 0.01	**0.54 ± 0.04**
T21	**0.69 ± 0.02**	0.67 ± 0.02

Table 2: Each dataset's average AUROC for SSIF trained in an unsupervised way and IF.

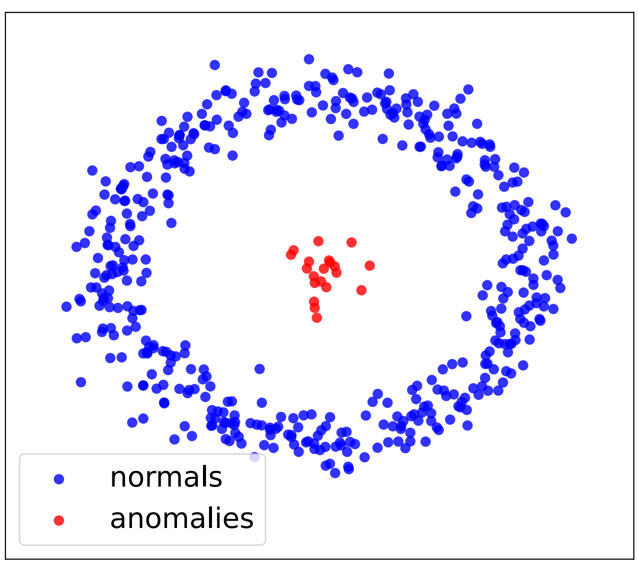

Figure 8: 2D toy donut dataset where anomalies are located in the middle of the normal instances

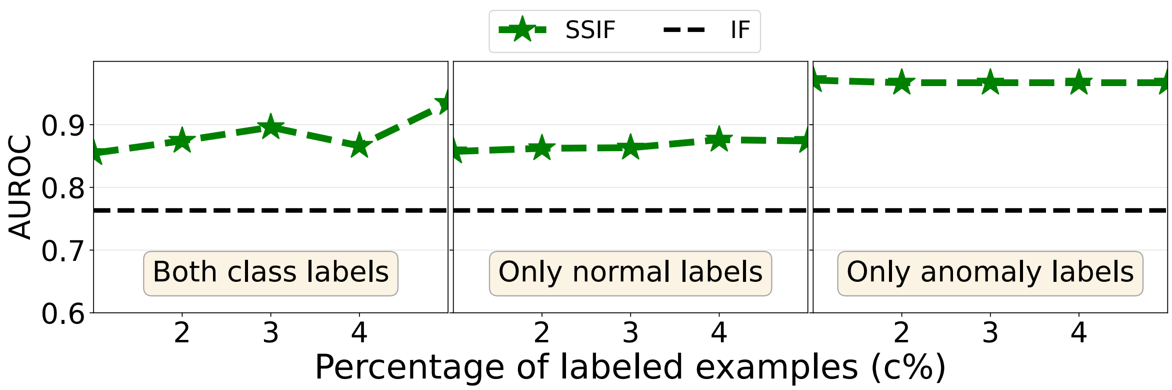

Figure 9: SSIF and IF's average AUROC for different percentage of labels c aggregated over ten iterations on a 2D toy donut dataset. SSIF significantly outperforms IF for every percentage of labels considered.