D2U: Distance-to-Uniform Learning for Out-of-Scope Detection

Anonymous ACL submission

Abstract

Supervised models trained for single-label classification tasks with cross-entropy loss are implicitly enforced to produce probability distributions that follow a discrete delta distribution in training. Model predictions in test time are expected to be similar to delta distributions given that the classifier determines the class of an input correctly. However, the shape of the predicted probability distribution becomes similar to the uniform distribution when the model cannot infer properly. We exploit this observation for detecting out-of-scope (OOS) utterances in conversational systems. Specifically, we propose a zero-shot post-processing step, called Distance-to-Uniform (D2U), exploiting not only the classification confidence score, but the shape of the entire output distribution. We also introduce a learning procedure that uses D2U for loss calculation in the supervised setup. We conduct experiments using six publicly available datasets. Experimental results show that the performance of out-of-scope detection is improved with our post-processing when there is no OOS training data, as well as with D2U learning procedure when OOS training data is available.

1 Introduction

Automated conversational systems have recently received attention from the research community (Dopierre et al., 2021; Mehri et al., 2020; Qin et al., 2021). In applications such as voice assistants, Spoken Language Understanding (Young et al., 2013) aims to extract meaning from the user inputs, called *utterances*, in order to process and execute desired functionalities. The task of Intent Detection, or Intent Classification, aims to classify user utterance into a set of system-identifiable intents. However, supervised training of such systems can only cover a restricted set of classes, i.e. in-scope (INS) classes. To enhance user experience, the task of Out-of-Scope (OOS) detection (Lin and Xu, 2019a; Xu et al., 2021; Zhan et al., 2021; Shen et al., 2021) distinguishes INS utterances from those that do not belong to the scope of the classifier with dedicated model architectures and loss functions.

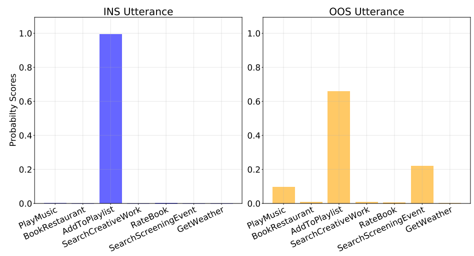

Figure 1: Sample output distributions of an INS classifier predicting the intent of an INS and OOS utterance. Since OOS utterances do not belong to any intent, the classifier is confused, and the prediction gets closer to the uniform distribution.

Modeling OOS detection in a supervised setting can be problematic, since covering every OOS intent in training data is quite challenging. Zero-shot OOS detection is a potential solution to this problem, which examines the prediction confidence of an INS classifier (Hendrycks et al., 2020) to discriminate OOS utterances at inference time. During INS training, cross-entropy loss between model predictions and ground-truth is minimized, resulting in confident predictions for INS utterances that enable a confidence threshold value separating INS and OOS utterances. However, softmax classifiers suffer from overconfident predictions for OOS data (Hendrycks and Gimpel, 2017), which makes it difficult to accurately determine a threshold value.

Figure 1 illustrates output probability distributions of a classifier for predicting the intent class for an INS and OOS utterance. The classifier, trained only on INS utterances, is confused when OOS utterance is given. The model assigns closer probabilities for different classes since there is no *correct* class for this OOS utterance, hence the resulting distribution gets closer to a uniform distribution than a discrete delta distribution.

Based on this observation, we propose to measure the dissimilarity from or distance to the uniform distribution (D2U). Statistical distance or divergence calculations between the prediction distribution and uniform distribution enable a decision boundary to be more accurately determined. Figure 2 illustrates possible benefits of using distance to the uniform distribution with cross-entropy. The subplot at the left shows the distribution of the number of utterances according to their Maximum Likelihood Estimate (MLE) score. The subplot at the right shows the same distribution according to cross-entropy score between prediction probability and uniform distribution. A decision boundary or threshold can be easily determined using D2U's cross-entropy as a post-processing step without any OOS training data.

When OOS training data is available (Larson et al., 2019), D2U can be used as a loss function to minimize the distance between OOS predictions and the uniform distribution. Such a loss function forces OOS predictions to be less confident, and benefit D2U post-processing further. To test our hypothesis that D2U is a useful method for OOS detection, we answer the following research questions:

- **Research Question 1:** Does the application of D2U as a post-processing step on INS classifier predictions increase OOS detection performance when there is no OOS training data?
- **Research Question 2:** Can the performance be boosted by incorporating D2U into the training procedure as a particular loss function when OOS training data is available?

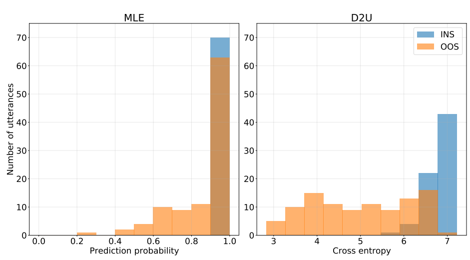

Figure 2: The histogram of prediction probability scores for INS and OOS utterances (MLE) by using a classifier trained on only INS utterances at the left. Instead of MLE, for the same classifier, cross-entropy score between prediction distribution and uniform distribution (D2U) is given at the right. A vertical decision boundary on the x-axis separating INS and OOS utterances can be more accurately determined with D2U.

- **Research Question 3:** Is the performance of OOS detection significantly improved by D2U over existing state-of-the-art methods?

We design dedicated experiments for each research question, and provide results for zero-shot setup with no OOS training data (RQ1), supervised training setup with OOS training data (RQ2), and comparisons against the state-of-the-art OOS detection algorithms from the literature (RQ3). We use six publicly available datasets and report five performance metrics with statistical significance analyses.

This paper is structured as follows. In Section 2, we provide a summary of related studies. Our proposed method is described in Section 3. We report the experimental details in Section 4. We discuss several aspects of our study in Section 5, and conclude the paper in Section 6.

2 Related Work

We divide OOS detection studies into three categories: (i) Confidence-based, (ii) representation-based, and (iii) distance-based methods.

2.1 Confidence-based OOS detection

Threshold-based Methods In earlier studies, detecting OOS utterances is achieved by thresholding the softmax output of INS classifiers (Larson et al., 2019; Feng et al., 2020; Zhang et al., 2020), which reflects the intuition that a classifier network is likely to output a more confident prediction score for a sample that follows its training distribution. The overconfidence problem of softmax classifiers (Hendrycks and Gimpel, 2017), although found to be less apparent in Transformer-based (Vaswani et al., 2017) models (Hendrycks et al., 2020), hinders threshold-based OOS detection performance.

Post-processing Methods The overconfidence problem of softmax classifiers is tackled by post-processing predictions. ODIN (Liang et al., 2018) and SofterMax (Lin and Xu, 2019b) apply temperature scaling for enlarging the confidence gap between INS and OOS instances, since INS logits are ideally further away on the positive axis of the softmax input. Gangal et al. (2020) utilize likelihood ratios with generative classifiers to distinguish OOS predictions. Our proposed method, D2U, is a confidence-based post-processing method when there is no OOS training data available.

2.2 Representation-based OOS detection

Dedicated model architectures or loss functions help represent utterances in a high-dimensional space suitable for OOS detection. Large Margin Cosine Loss (LMCL) ensures that INS intents are tightly clustered (Zeng et al., 2021a), so that OOS utterances are exposed for outlier detection algorithms, such as Local Outlier Factor (Lin and Xu, 2019a). Intent class embeddings (Cavalin et al., 2020) model OOS detection as a reverse dictionary task by mapping intent classes and utterances to the same space. Yilmaz and Toraman (2020) propose a feature representation mechanism that uses KL Divergence to capture the changes in model predictions during sequential processing of utterances.

In order to mitigate data scarcity in OOS detection, Marek et al. (2021) propose a method for generating OOS data with Generative Adversarial Networks. GANs are also utilized for generating high-dimensional vector representations that are hard to distinguish from that of real utterances, providing adversarial signals to the INS classifier (Zeng et al., 2021b; Liang et al., 2021). The adversarial signal supplied during training ensures that the model is more robust to OOS samples, making their detection more achievable.

2.3 Distance-based OOS detection

Distances and divergences are useful tools in OOS detection, since they provide a measure of dissimilarity that can distinguish INS and OOS samples. Xu et al. (2020) utilize Euclidean and Mahalanobis distances with generative classifiers to identify outliers with Gaussian Discriminative Analysis. Mahalanobis distance calculated using representations from the intermediate layers of BERT (Devlin et al., 2019) increases OOS detection performance (Shen et al., 2021). Lee et al. (2018) introduce the confidence loss in Computer Vision for GANs that calculates KL Divergence between the training predictions for OOS samples and uniform distribution, and minimize it to achieve lower confidence values.

The idea of measuring the distance between prediction distribution and uniform distribution is utilized in different learning architectures (Lee et al., 2018; Gangal et al., 2020), but not extensively studied for OOS intent detection. Besides, we explore various distance metrics in zero-shot OOS detection and different distance-to-uniform training procedures in supervised setup.

3 Distance-to-Uniform Calculation for OOS Detection

3.1 D2U post-processing for zero-shot OOS detection

Supervised classifiers trained on INS data model the ground truth labels with a discrete delta function that corresponds to the label, given as follows.

$$\delta_{c_i}(x) = \begin{cases} 1, & \text{if } x = c_i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

For the data instance i, c_i is the ground-truth label indicating the correct class. The cross-entropy loss between softmax model output and discrete delta function is given as follows.

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} \delta_{c_i}(x) \log \hat{P}(u_i) \quad (2)$$

Here, $\hat{P}(u_i)$ is the output probability distribution of the model for utterance u_i in a batch of N utterances, and c_i is the correct class label for the given utterance. This criterion implicitly forces the model to generate confident predictions for a given data point with maximal confidence score assigned to the ground-truth class label, and low prediction scores for the other classes. When an OOS utterance is given to an intent classifier that is trained using only INS data, the classifier gets confused, i.e., the output probability distribution is more dissimilar to a delta distribution than what an INS utterance would result in. In other words, output distributions of OOS samples get closer to the uniform distribution than that of INS samples, an observation that we exploit for OOS detection.

The conventional methods for OOS detection make use of a pre-determined threshold value on the Maximum Likelihood Estimate (MLE) score assigned to the predicted label, given as follows.

$$OOS(u_i) = \begin{cases} 1, & \text{if } max(\hat{P}(u_i)) < \theta \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Here, θ is a pre-defined threshold value between 0 and 1, and $max(\hat{P}(u_i))$ is the MLE score, which considers only the prediction confidence and ignores the shape of the distribution. We exploit the information conveyed by the shape of the entire prediction distribution by first calculating a distance between the output distribution $\hat{P}$ and the uniform distribution U before applying the threshold, given as follows.

$$OOS(u_i) = \begin{cases} 1, & \text{if } dst(\hat{P}(u_i), U) < \theta \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The distance determined by the $dst(.)$ function between $\hat{P}(u_i)$ and U can be calculated with various distance metrics. We experiment with geometric distance calculations, such as Euclidean distance and Cosine distance; as well as statistical distance calculations, such as Jensen-Shannon distance and symmetrized Kullback-Leibler divergence. The distance value calculated by the $dst(.)$ function can be intuitively interpreted as the level of confidence of the model. When the distance value is low, the model is less confident and more confused, since the output distribution assigns closer scores for each class.

This is an architecture-agnostic zero-shot post-processing step which can be generalized to any classification model trained with cross-entropy loss with no need for OOS training data. OOS detection in test time is achieved by a function of the prediction distribution given by D2U.

3.2 Distance metrics for post-processing

We examine a number of geometric and statistical distance measures listed as follows.

- Bray Curtis Distance (BC): For two probability distributions, u and v, the Bray Curtis distance is given as $\sum_i |u_i - v_i| / \sum_i |u_i + v_i|$.
- Canberra Distance (Cbr): Canberra distance between u and v is $\sum_i (|u_i - v_i|/(u_i + v_i))$.
- Cosine Distance (Cos): Derived from the Cosine similarity, the Cosine distance is formulated as $1 - (u \cdot v / ||u||_2 ||v||_2)$ where $||.||_2$ is the L_2 norm.
- Euclidean Distance (Euc): The Euclidean distance between u and v is given as $||u - v||_2$.
- Hellinger Distance (Helng): The Hellinger distance between u and v is $||\sqrt{u} - \sqrt{v}||_2 / \sqrt{2}$.
- Cross-Entropy (CE): Cross-Entropy is a measure of dissimilarity between distributions u and v given as $-\sum_i u_i \log v_i$.
- Symmetrized KL Divergence (KL): The symmetrized Kullback-Leibler divergence is given as $[KL(u, v) + KL(v, u)]/2$ where $KL(u, v) = \sum_i u_i \log(u_i/v_i)$.
- Jenshen Shannon Distance (JS): Jensen Shannon Distance generalizes KL divergence between u and v as $(KL(u, m)/2) + (KL(v, m)/2)$ where m is the mean of two distributions.

3.3 D2U training procedure for supervised OOS detection

When OOS training data is available, we propose D2U loss function, as shown in Figure 3, to increase the similarity between OOS output probability distributions and the uniform distribution. We use pretrained BERT (Devlin et al., 2019) as the classifier network. The loss function for INS utterances, L_{ins}, is still cross-entropy between true label and prediction, given as follows.

$$L_{ins} = -\frac{1}{N_{ins}} \sum_{i=1}^{N_{ins}} \delta(c_i) \log \hat{P}(u_i) \tag{5}$$

For OOS utterances, the loss L_{oos}, is calculated against the uniform distribution, given as follows.

$$L_{oos} = \frac{1}{N_{oos}} \sum_{i=1}^{N_{oos}} dst(\hat{P}(u_i), U) \tag{6}$$

The total loss is the weighted average over a batch of utterances containing N_{ins} number of INS utterances and N_{oos} number of OOS utterances, given as follows.

$$L_{total} = \frac{N_{ins} L_{ins} + N_{oos} L_{oos}}{N_{ins} + N_{oos}} \tag{7}$$

As the $dst(.)$ function in Equation 6, we experiment with differentiable functions; such as cross-entropy, KL divergence, and Sinkhorn distance (Cuturi, 2013), named as D2U-CE, D2U-KL, and D2U-S, respectively. These functions treat the model

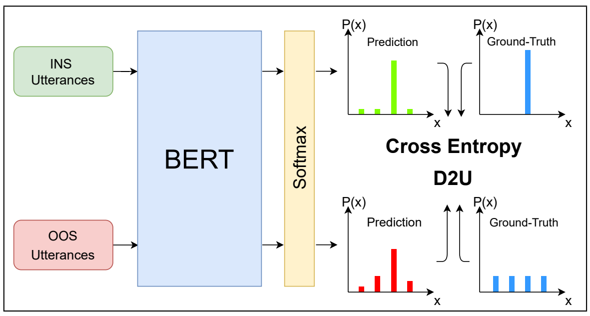

Figure 3: The architecture of the proposed supervised learning method, D2U. INS utterances are learned with conventional cross-entropy loss against true label distribution, while OOS loss is calculated against the uniform distribution.

output and ground truth as probability distributions and provide a differentiable measure. We do not modify the loss calculation for INS utterances so as not to affect the INS classification performance.

Note that this architecture does not model the OOS intent as a separate class. Therefore, post-processing is applied as described in Section 3.1 in test time. Since the loss function incorporates D2U into training, the performance gain by the post-processing is expected to be increased.

4 Experiments

4.1 Datasets

We use six publicly available intent classification datasets, some of which include labeled OOS data. CLINC (Larson et al., 2019) is a dataset with 150 INS intent classes targeting various domains with curated OOS data. We use the OOS split of CLINC to augment other existing intent detection datasets that do not include labeled OOS data; which are ACID (Acharya and Fung, 2020), Banking (Casanueva et al., 2020), HWU64 (Liu et al., 2019), and SNIPS (Coucke et al., 2018). We observe that HWU64 has many short and noisy utterances, we therefore remove any utterances with length less than or equal to three words.

TOP (Gupta et al., 2018) is an intent detection dataset that generalizes conventional intent labeling with semantic parsing. The intent labels follow a hierarchical structure with potentially many labels for an utterance. However, we take only root intent class label into account to be consistent with other datasets. The utterances with the intent labels "UNSUPPORTED" and "UNSUPPORTED_NAVIGATION" are treated as OOS. We give the main statistics of the datasets in Table 1.

The variety of the number of classes, average length (number of words), and vocabulary size provides a wide spectrum for understanding different OOS detection scenarios. For instance, TOP dataset can be considered a low resource setup since the number of OOS utterances is significantly lower than INS utterances.

Table 1: The statistics of the datasets used in this study.

	ACID	Banking	CLINC	HWU64	SNIPS	TOP
INS	22,172	13,081	22,500	23,431	13,784	36,668
OOS	16,000	16,000	16,000	16,000	16,000	3,653
Total	38,172	29,081	38,500	39,431	29,784	40,321
Vocabulary	25,083	26,702	25,810	26,069	30,100	12,610
Avg. Len.	8.95	9.85	8.39	7.25	8.65	8.93
Classes	175	77	150	46	7	16

4.2 Evaluation metrics

To assess the performance of OOS detection, we report the scores of Receiver Operating Curve Area Under Curve (ROC AUC), False Positive Rate at 90% OOS True Positive Rate (FPR90), and False Negative Rate at 90% OOS True Negative Rate (FNR90). Since these performance metrics are independent of a threshold value used for decision boundary, they provide a means of fair comparison. We also report weighted OOS Recall and weighted OOS F1 metrics based on the threshold value that maximizes the Youden's J statistic (Youden, 1950) on a validation set.

Compared to Precision, Recall is arguably a more critical performance metric for OOS detection; since Recall considers Type II error, meaning that OOS utterances are mislabeled as INS. In this case, the voice assistant would execute a task that the user does not intent to do. We argue that ROC is a more generic measure that considers the performances of varying thresholds, than Recall and F1 considering only a fixed threshold.

4.3 Baseline approaches

We provide results for important baseline methods. In the experiments, BERT (Devlin et al., 2019) with softmax layer is used as the classifier network. For RQ1, the baseline zero-shot post-processing approaches are listed below.

- **MLE** (Hendrycks and Gimpel, 2017; Hendrycks et al., 2020): The confidence score of a classifier trained only on INS utterances is used with thresholding.
- **Softmax temperature scaling (Temp)** (Liang et al., 2018; Lin and Xu, 2019b): As a modification to the MLE setup, the softmax input is applied a temperature value of 10^3.
- **Standard deviation (Stdev)**: We use the standard deviation of the distribution before thresholding since OOS predictions would have lower standard deviation.
- **Entropy (Ent)** (Shen et al., 2021): The entropy of the prediction distribution, $H(\hat{P}(u_i))$, is calculated before applying the threshold, as follows.

$$OOS(u_i) = \begin{cases} 1, & \text{if } H(\hat{P}(u_i)) > \theta \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Table 2: **RQ1:** D2U-zero with various distance metrics vs. post-processing baselines in zero-shot setup. Row-wise best scores are given in bold. (↑) and (↓) indicate that higher and lower scores are better, respectively. "•" indicates statistically significant differences with two-tailed paired t-test at a 95% interval (with Bonferroni correction $p < 0.0125$) in pairwise comparison between D2U-zero and all baselines except the ones marked with "∘".

Metric	Dataset	Baselines				D2U-zero							
		MLE	Temp	Stdev	Ent	CE	BC	Cbr	Cos	JS	Euc	KL	Helng.
ROC AUC (↑)	ACID	90.93	91.83∘	91.54	91.98	92.01	91.40	90.18	91.54	92.08	91.54	**92.14•**	92.13
	Banking	94.92	96.15	95.88	96.66	**97.03•**	96.89	96.09	95.88	96.99	95.88	96.91	96.97
	CLINC	95.34	96.16	95.52	95.90	**96.32•**	96.09	96.26	95.52	96.24	95.52	96.20	96.25
	HWU64	79.29	80.46	79.95	80.90	80.32	**81.49•**	80.49	79.95	81.16	79.95	80.92	81.05
	SNIPS	95.45	96.20	95.50	95.70	**96.33•**	95.61	95.83	95.50	95.86	95.50	96.15	96.04
	TOP	74.23	73.23	74.26	74.19	73.24	74.32	**74.36•**	74.26	74.18	74.26	73.25	73.76
FPR90 (↓)	ACID	25.00	22.10∘	20.70∘	19.85∘	21.40	24.80	28.60	20.70	20.90	20.70	**19.70•**	20.70
	Banking	13.30	10.00	11.30	7.85	7.30	6.60	12.00	11.30	6.20	11.30	6.30	**6.10•**
	CLINC	9.30	7.80∘	9.30	8.00∘	**7.50•**	8.10	7.70	9.30	7.70	9.30	7.80	7.60
	HWU64	58.90	54.10	55.70	**48.12**	52.50	51.60	55.00	55.70	52.20	55.70	52.60	51.90
	SNIPS	10.40	**8.40**	10.30	10.71	**8.40**	10.40	10.30	10.30	9.90	10.30	8.60	8.70
	TOP	**51.38**	53.00	**51.38**	66.65	53.00	**51.38**	**51.38**	**51.38**	51.50	**51.38**	53.00	51.75
FNR90 (↓)	ACID	27.30	25.67	26.22	**19.70•**	21.46∘	20.60∘	26.70	26.22	20.45∘	26.22	21.38∘	20.75∘
	Banking	14.36	10.49	11.37	**7.20•**	8.01∘	8.79∘	13.68	11.37	7.88∘	11.37	7.69∘	7.79∘
	CLINC	11.80	9.16∘	11.60	8.90	8.36	8.11	**7.56•**	11.60	7.98	11.60	8.67	8.31
	HWU64	51.97	52.26	51.75	52.80∘	52.01	47.14	**47.01•**	51.75	48.63	51.75	51.03	49.57
	SNIPS	11.14	**9.29**	11.14	11.80	**9.29**	11.14	11.14	11.14	10.71	11.14	9.71	10.29
	TOP	67.88	69.05	67.89	**51.50**	69.07	67.89	67.90	67.89	68.17	67.89	69.04	68.46

For RQ2, we use D2U zero-shot cross-entropy post-processing (D2U-zero) as the baseline method, since we examine any improvement in supervised setup over zero-shot. For RQ3, we compare supervised D2U with the following baselines.

- **Large Margin Cosine Loss (LMCL)** (Zeng et al., 2021b): During INS training, the Cosine distance among class centroids is increased up to a margin. We set the margin as 0.35, and scaling factor as 30.
- **Domain Regularization Module (DRM)** (Shen et al., 2021): DRM introduces domain logits for regularization during INS training. We slightly modify the design and apply sigmoid to domain logits before dividing the classification logits for training stability.
- **BERT-Binary (Binary)** (Devlin et al., 2019): The "bert-base-uncased" model fine-tuned as a binary classifier for OOS detection.
- **Entropy Regularization (Reg.)** (Zheng et al., 2020): Entropy of OOS training predictions are maximized with a loss function given as follows.

$$L_{oos} = -\frac{1}{N_{oos}} \sum_{i=1}^{N_{oos}} H(\hat{P}(u_i)) \quad (9)$$

4.4 Experimental design

The experiments are designed with respect to our research questions (RQ 1-3). First, we fine-tune a BERT classifier (Devlin et al., 2019) for INS intent detection with cross-entropy loss, and apply different D2U post-processing methods for RQ1. Then, we fix the post-processing method, and examine the effect of supervised D2U training for RQ2. Lastly, we compare D2U with state-of-the art baselines for RQ3 to assess the performance gain of our method.

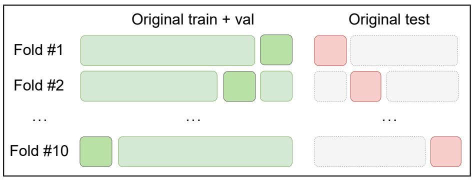

Figure 4: Modified leave-one-out 10-fold split strategy that complies with original splits. At each fold, only 10% of test data is included, while 90% of training data is retained and the remaining 10% is used as validation.

To avoid potential annotator-dependent effects as noted by Larson et al. (2019) and comply with the original splits, we modify 10-fold leave-one-out cross-validation as illustrated in Figure 4. The validation splits are used to find confidence threshold values for Recall and F1 calculations. We validate statistically significant differences in the average performances of 10-folds with the two-tailed paired t-test at a 95% interval with Bonferroni correction. Note that the test splits do not overlap in order to satisfy the independence criterion of t-test.

4.5 Experimental results

RQ1: D2U in zero-shot setup. In Table 2, we report ROC AUC, FPR90, and FNR90 scores for different post-processing methods applied to a BERT-

Table 3: **RQ2:** D2U training compared to zero-shot. "•" indicates statistically significant differences with the two-tailed paired t-test at a 95% interval in pairwise comparison between D2U-zero and best supervised.

Data	Method	ROC↑	FPR90↓	FNR90↓	REC↑	F1↑
ACID	D2U-zero	92.01	21.40	21.46	86.43	88.69
	D2U-CE	96.75	**7.30•**	7.96	95.98	95.55
	D2U-KL	**96.78•**	7.90	**7.76•**	**96.31•**	**96.01•**
	D2U-S	95.88	8.80	9.54	93.18	93.78
Banking	D2U-zero	97.03	7.30	8.01	91.47	91.67
	D2U-CE	**99.36•**	**1.00•**	**0.23•**	96.66	96.55
	D2U-KL	99.25	1.70	0.39	**97.47•**	**97.42•**
	D2U-S	98.79	2.00	2.12	95.90	95.88
CLINC	D2U-zero	96.32	7.50	8.36	91.31	91.75
	D2U-CE	97.48	5.10	6.18	93.27	92.84
	D2U-KL	97.29	5.20	**4.93•**	93.33	92.91
	D2U-S	**97.69•**	**3.90•**	5.36	**94.53•**	**94.53•**
HWU64	D2U-zero	80.32	52.50	52.01	**76.83**	**76.03•**
	D2U-CE	**87.37•**	31.70	**37.05•**	74.58	68.18
	D2U-KL	87.19	**30.30•**	41.50	75.27	69.41
	D2U-S	82.23	47.80	49.10	74.28	68.30
SNIPS	D2U-zero	96.33	8.40	9.29	88.35	88.43
	D2U-CE	98.61	2.70	2.86	89.47	89.52
	D2U-KL	**99.16•**	**1.60•**	**1.57•**	88.59	88.64
	D2U-S	98.39	2.80	2.29	**90.29**	**90.36**
TOP	D2U-zero	73.24	53.00	69.07	84.54	86.14
	D2U-CE	97.42	6.25	**4.03•**	94.55	95.01
	D2U-KL	**97.50•**	**5.88•**	4.10	**95.17•**	**95.51•**
	D2U-S	94.94	12.00	15.61	92.13	92.95

based INS classifier with no OOS training data. Our proposed method, D2U-zero, statistically significantly outperforms all baselines in all datasets with respect to ROC AUC score. Using cross-entropy for D2U-zero has better performance in majority of cases, compared to other distance metrics. The reason for its success might be that cross-entropy is the loss function used in the training procedure of the model. In terms of FPR90 and FNR90, D2U-zero does not always outperform all baselines. Though, the cases when baselines outperform are not statistically significant. This shows that the baseline methods can optimize FPR90 and FNR90 individually but cannot outperform D2U in terms of ROC which considers Type I and Type II errors simultaneously. Entropy (Shen et al., 2021) is a strong baseline that performs better than other baselines with respect to all performance metrics.

RQ2: D2U in supervised setup. Next, we report the effect of D2U training on OOS detection in Table 3. Since our concern here is to observe any improvement over zero-shot setup, we fix post-processing method as cross-entropy for all methods due to its performance in the previous experiment. The results show that using D2U as a loss function statistically significantly improves the performance of D2U-zero in almost all cases. KL divergence loss (D2U-KL) and Cross-Entropy loss (D2U-CE) are effective D2U methods in ACID, Banking, HWU64, SNIPS, and TOP datasets, whereas Sinkhorn distance (D2U-S) is effective in CLINC dataset.

RQ3: D2U versus state-of-the-art. The performances of state-of-the-art baseline OOS detection models, regardless of zero-shot or supervised, and D2U methods are compared in Table 4, with extensive results reported in the Appendix. MLE, softmax temperature (Temp.), Entropy, LMCL, and DRM are zero-shot OOS detection setups, whereas entropy regularization (Reg.) and BERT-Binary (Binary) are supervised setups. D2U statistically significantly outperforms other baselines in most datasets, although Binary is a strong baseline method that outperforms D2U in ACID and Banking datasets and challenges it in HWU64 and TOP, which is not statistically significant. ACID, Banking and TOP datasets contain domain-specific utterances; from insurance, banking, and navigation applications, respectively. This might cause a trivial detection for the BERT-based binary classifier. HWU64 contains generic utterances like queries and questions which may coincide with the OOS split and disturb the training process of D2U.

5 Discussion

5.1 Limitations

We acknowledge some limitations to our study. All methods in our study, including baselines, use BERT (Devlin et al., 2019) as the classifier network but one can experiment with other multiclass prediction models. In addition, the generalization ability of D2U to other neural networks such as LSTM and CNN are not investigated.

Except for CLINC and TOP, the datasets are augmented with the OOS data from CLINC. However, we argue that this approach is nontrivial since the majority of the OOS training data is sampled from Wikipedia (Larson et al., 2019) and remains OOS for other datasets. Moreover, D2U has effective performance on CLINC and TOP datasets which are specifically designed with OOS utterances.

5.2 Qualitative analysis

We provide a qualitative analysis on the effect of D2U training. We illustrate the model output distributions for INS utterance "*get me to*

Table 4: **RQ3:** D2U vs. OOS detection baselines. The bold score is the best. The underlined score is the best that baseline achieves when D2U outperforms, or vice versa. "•" indicates statistically significant differences with the two-tailed paired t-test at a 95% interval (with Bonferroni correction $p < 0.0071$) in pairwise comparisons between D2U and all baselines except the ones marked with "○". If baseline outperforms, "•" indicates the difference (with Bonferroni correction $p < 0.0167$) in pairwise comparisons between the baseline and our best version.

Train	ACID					Banking					CLINC				
	ROC	FPR	FNR	REC	F1	ROC	FPR	FNR	REC	F1	ROC	FPR	FNR	REC	F1
MLE (Hendrycks et al., 2020)	90.9	25.0	67.9	84.9	87.5	94.9	13.3	14.4	89.4	89.6	95.3	9.3	11.8	90.2	90.8
Temp. (Liang et al., 2018)	91.8	22.1	69.1	85.8	88.2	96.2	10.0	10.5	90.3	90.5	96.2	7.8	9.2	90.1	90.7
Entropy (Shen et al., 2021)	92.0	19.9	51.5	86.9	89.0	96.7	7.9	7.2	91.6	91.8	95.9	8.0	8.9	90.3	90.8
Binary (Devlin et al., 2019)	**97.2**	**6.2**	**6.7**	**96.5**	**96.1**	**99.9**	**0.2**	**0.2**	**97.9**	**97.8**	85.6	48.6	31.4	88.3	86.1
LMCL (Zeng et al., 2021a)	94.1	15.6	66.5	88.4	90.1	97.2	6.3	8.1	92.6	92.6	96.3	7.4	9.9	90.8	91.3
DRM (Shen et al., 2021)	93.2	19.9	62.5	86.8	89.0	96.1	13.1	11.4	90.6	90.5	95.9	8.5	9.7	91.0	91.4
Reg. (Zheng et al., 2020)	96.0	10.3	7.1	95.6	95.1	99.0	2.4	0.9	96.8	96.8	97.3○	6.5	6.8	93.3○	92.9○
D2U-CE-CE (ours)	96.8	7.3	8.0	96.0	95.6	99.4	1.0	0.2	96.7	96.6	97.5	5.1	6.2	92.3	92.8
D2U-KL-CE (ours)	96.8	7.9	7.8	96.3	96.0	99.3	1.7	0.4	97.5	97.4	97.3	5.2	**4.9**	93.3	92.9
D2U-S-CE (ours)	95.9	8.8	9.5	93.2	93.8	98.8	2.3	2.1	95.9	95.9	**97.7•**	**3.9•**	5.4	**94.5•**	**94.5•**

Train	HWU64					SNIPS					TOP				
	ROC	FPR	FNR	REC	F1	ROC	FPR	FNR	REC	F1	ROC	FPR	FNR	REC	F1
MLE (Hendrycks et al., 2020)	79.3	58.9	52.0	73.0	73.7	95.5	10.4	11.1	88.3○	88.4○	74.2	51.4	67.9	84.9	86.6
Temp. (Liang et al., 2018)	80.5	54.1	52.3	76.7	76.5	96.2	8.4	9.3	88.4○	88.5○	73.2	53.0	69.1	84.5	86.1
Entropy (Shen et al., 2021)	80.9	48.1	52.8	77.7	77.3	95.7	10.7	11.8	88.2○	88.3○	74.2	66.7	51.5	84.8	86.5
Binary (Devlin et al., 2019)	**88.0**	35.8○	**31.0**	74.7	67.8	98.9○	1.7○	2.0○	86.2○	86.2○	97.3○	**4.4•**	5.8○	**97.0•**	**97.0•**
LMCL (Zeng et al., 2021a)	84.3	43.0	49.0	**80.3•**	**80.2•**	85.2	49.5	31.9	67.8	66.3	70.6	69.8	66.5	57.8	66.3
DRM (Shen et al., 2021)	79.3	56.9	50.3	73.4	74.0	93.6	13.4	12.0	87.9○	87.9○	77.0	50.1	62.5	81.7	84.4
Reg. (Zheng et al., 2020)	83.4	46.5	45.2	74.0	67.0	98.6○	2.5○	2.7○	88.4○	88.5○	96.5	7.5	7.1○	94.5	94.9
D2U-CE-CE (ours)	87.4	31.7	37.1	74.6	68.2	98.6	2.7	2.9	89.5	89.5	97.4	6.3	**4.0•**	94.6	95.0
D2U-KL-CE (ours)	87.2	**30.3•**	41.5	75.3	69.4	**99.2•**	**1.6•**	**1.6•**	88.6	88.6	**97.5•**	5.9	4.1	95.2	95.5
D2U-S-CE (ours)	82.2	47.8	49.1	74.3	68.3	98.4	2.8	2.3	**90.3•**	**90.4•**	94.9	12.0	15.6	92.1	93.0

ritzville by 4 via the freeway." belonging to the "GET_DIRECTIONS" intent, and the OOS utterance "*how many skating rinks are available in the south pacific tomorrow at 10*" taken from the TOP dataset in Figure 5. We observe that the OOS utterance results in an overconfident prediction in the BERT MLE model whereas the prediction distribution of D2U-CE is quite similar to uniform distribution.

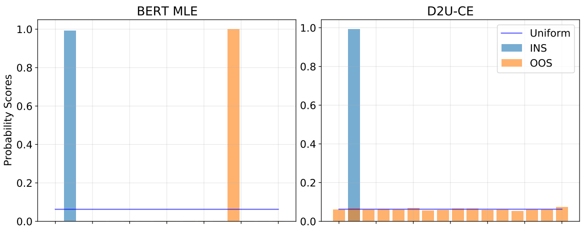

Figure 5: The effect of D2U on prediction distributions.

5.3 INS performance

In Table 5, we analyze if OOS detection models deteriorate the performance of INS detection. We set MLE as a baseline, which does not modify training procedure. The results show that the performance of INS classification is not dramatically deteriorated by the supervised models including D2U in SNIPS and TOP, whereas it is even improved in the remaining datasets. Although D2U's INS performance is similar to other supervised models, D2U has better OOS performance than others, as observed in Table 4.

We do not include BERT-Binary, which has no capability of INS classification. BERT-Binary has a challenging OOS performance in Table 4, but D2U has advantage of showing state-of-the-art performances for both INS and OOS detection.

Table 5: Weighted F1 score for INS classification.

Method	Datasets					
	ACID	Banking	CLINC	HWU64	SNIPS	TOP
MLE	80.74	84.91	95.67	81.97	98.14	98.60
LMCL	85.69	89.64	95.83	82.08	97.86	98.56
DRM	88.70	89.89	96.25	82.49	98.14	98.68
Reg.	86.60	91.22	96.38	82.55	97.43	98.32
D2U-CE	86.40	90.95	96.42	82.31	97.84	98.26
D2U-KL	86.26	90.69	96.22	82.72	98.01	98.29
D2U-S	86.50	91.52	96.34	82.16	97.43	98.37

6 Conclusion

In this study, we improve confidence-based OOS detection performance with a distance calculation between classifier prediction and uniform distribution. In zero-shot setup, our proposed method serves as an architecture-agnostic post-processing step to emphasize the distinction between INS and OOS utterances. With use of OOS training data in the supervised setup, we bring closer OOS predictions to uniform distribution with a specialized loss calculation. Experimental results demonstrate that D2U improves OOS detection performance over existing baselines. We plan to extend our study to different neural network architectures and deep learning tasks, such as other out-of-domain tasks and other research areas.

References

Shailesh Acharya and Glenn Fung. 2020. Using optimal embeddings to learn new intents with few examples: An application in the insurance domain.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

Paulo Cavalin, Victor Henrique Alves Ribeiro, Ana Appel, and Claudio Pinhanez. 2020. Improving out-of-scope detection in intent classification by using embeddings of the word graph space of the classes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3952–3961, Online. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. ProtAugment: Intent detection meta-learning through unsupervised diverse paraphrasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2454–2466, Online. Association for Computational Linguistics.

Yulan Feng, Shikib Mehri, Maxine Eskenazi, and Tiancheng Zhao. 2020. "none of the above": Measure uncertainty in dialog response retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2013–2020, Online. Association for Computational Linguistics.

Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. 2020. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7764–7771.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792, Brussels, Belgium. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. 2018. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*.

Chaojie Liang, Peijie Huang, Wenbin Lai, and Ziheng Ruan. 2021. Gan-based out-of-domain detection using both in-domain and out-of-domain samples. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7663–7667. IEEE.

Shiyu Liang, Yixuan Li, and R Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*.

Ting-En Lin and Hua Xu. 2019a. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5496, Florence, Italy. Association for Computational Linguistics.

Ting-En Lin and Hua Xu. 2019b. A post-processing method for detecting unknown intent of dialogue system via pre-trained deep neural network classifier. *Knowledge-Based Systems*, 186:104979.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. In *10th International Workshop on Spoken Dialogue Systems Technology 2019*.

Petr Marek, Vishal Ishwar Naik, Anuj Goyal, and Vincent Auvray. 2021. Oodgan: Generative adversarial network for out-of-domain data generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 238–245.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*.

Libo Qin, Fuxuan Wei, Tianbao Xie, Xiao Xu, Wanxiang Che, and Ting Liu. 2021. GL-GIN: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 178–188, Online. Association for Computational Linguistics.

Yilin Shen, Yen-Chang Hsu, Avik Ray, and Hongxia Jin. 2021. Enhancing the generalization for intent classification and out-of-domain detection in SLU. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2443–2453, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Hong Xu, Keqing He, Yuanmeng Yan, Sihong Liu, Zijun Liu, and Weiran Xu. 2020. A deep generative distance-based classifier for out-of-domain detection with mahalanobis space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1452–1460.

Keyang Xu, Tongzheng Ren, Shikun Zhang, Yihao Feng, and Caiming Xiong. 2021. Unsupervised out-of-domain detection via pre-trained transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1052–1061, Online. Association for Computational Linguistics.

Eyup Halit Yilmaz and Cagri Toraman. 2020. *KLOOS: KL Divergence-Based Out-of-Scope Intent Detection in Human-to-Machine Conversations*, page 2105–2108. Association for Computing Machinery, New York, NY, USA.

WJ Youden. 1950. Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Zijun Liu, Yanan Wu, Hong Xu, Huixing Jiang, and Weiran Xu. 2021a. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 870–878, Online. Association for Computational Linguistics.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Hong Xu, and Weiran Xu. 2021b. Adversarial self-supervised learning for out-of-domain detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5631–5639.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Y.S. Lam. 2021. Out-of-scope intent detection with self-supervision and discriminative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3521–3532, Online. Association for Computational Linguistics.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, S Yu Philip, Richard Socher, and Caiming Xiong. 2020. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1198–1209.

A Appendix

We report Receiver Operating Curve Area Under Curve, False Positive Rate at 90% OOS True Positive Rate, False Negative Rate at 90% OOS True Negative Rate, weighted OOS Recall, and weighted OOS F1 scores in Tables 6, 7, 8, 9, 10 respectively. Different training procedures, baseline and proposed, are reported in rows and different post-processing methods, baseline and proposed, are reported in columns.

Baseline training methods are BERT-based in-scope classifier (MLE) (Larson et al., 2019; Devlin et al., 2019), Large Margin Cosine Loss (LMCL) (Zeng et al., 2021a), Domain Regularization Module (DRM) (Shen et al., 2021), entropy regularization (Reg.) (Zheng et al., 2020), and BERT-binary classifier (Binary) (Devlin et al., 2019). Post-processing methods are not applicable for Binary training since it models OOS detection as a binary classification problem. Baseline post-processing methods are Maximum Likelihood Estimate (MLE) (Gangal et al., 2020; Zhang et al., 2020), softmax temperature (Temp) (Liang et al., 2018; Lin and Xu, 2019b), standard deviation (Stdev), and entropy (Ent) (Shen et al., 2021).

Table 6: Average ROC AUC score of 10-Fold binary OOS Detection. Row-wise highest scores are given in bold.

Data	Training	MLE	Temp	Stdev	Ent	CE	BC	Cbr	Cos	JS	Euc	KL	Helng.
ACID	MLE	90.93	91.83	91.54	91.98	92.01	91.40	90.18	91.54	92.08	91.54	**92.14**	92.13
	LMCL	94.05	94.07	**94.23**	94.04	93.72	89.31	85.54	**94.23**	93.88	**94.23**	93.91	93.89
	DRM	93.23	92.43	93.54	93.95	91.70	**94.07**	93.67	93.54	94.00	93.54	92.92	93.85
	Reg.	95.98	96.56	96.15	96.50	97.06	96.82	**97.10**	96.15	96.84	96.15	96.75	96.81
	Binary	**97.19**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	95.63	96.22	95.85	96.17	**96.75**	96.43	96.72	95.85	96.47	95.85	96.42	96.46
	D2U-KL	95.47	96.25	95.65	96.06	**96.78**	96.47	96.77	95.65	96.49	95.65	96.37	96.45
	D2U-S	94.46	95.14	94.80	95.24	**95.88**	95.55	95.61	94.80	95.63	94.80	95.54	95.61
Banking	MLE	94.92	96.15	95.88	96.66	**97.03**	96.89	96.09	95.88	96.99	95.88	96.91	96.97
	LMCL	97.19	97.20	**97.32**	97.15	96.88	94.07	91.61	97.32	97.01	**97.32**	97.04	97.03
	DRM	96.12	96.97	96.70	97.47	96.56	**98.06**	97.93	96.70	97.97	96.70	97.28	97.88
	Reg.	98.95	99.12	99.03	99.13	**99.19**	99.16	99.18	99.03	99.18	99.03	99.16	99.17
	Binary	**99.88**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	99.07	99.28	99.14	99.24	**99.36**	99.26	99.29	99.14	99.28	99.14	99.29	99.29
	D2U-KL	98.99	99.19	99.07	99.15	**99.25**	99.19	99.22	99.07	99.22	99.07	99.22	99.22
	D2U-S	97.83	98.53	98.12	98.43	**98.79**	98.61	98.65	98.12	98.65	98.12	98.63	98.64
CLINC	MLE	95.34	96.16	95.52	95.90	**96.32**	96.09	96.26	95.52	96.24	95.52	96.20	96.25
	LMCL	**96.31**	96.30	96.30	96.14	95.81	92.51	86.08	96.30	95.95	96.30	96.02	95.99
	DRM	95.85	94.47	96.00	96.19	93.78	**96.29**	95.73	96.00	95.88	96.00	95.07	95.63
	Reg.	97.29	97.63	97.31	97.47	97.58	97.52	97.55	97.31	97.62	97.31	**97.65**	97.64
	Binary	**85.57**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	97.08	97.47	97.14	97.30	**97.48**	97.27	97.31	97.14	97.42	97.14	**97.48**	97.45
	D2U-KL	96.86	**97.29**	96.90	97.04	**97.29**	97.07	97.12	96.90	97.20	96.90	97.26	97.23
	D2U-S	96.71	97.54	96.85	97.15	**97.69**	97.22	97.33	96.85	97.42	96.85	97.53	97.48
HWU64	MLE	79.29	80.46	79.95	80.90	80.32	**81.49**	80.49	79.95	81.16	79.95	80.92	81.05
	LMCL	84.28	84.37	84.98	85.17	**85.33**	84.04	82.50	84.98	85.28	84.98	85.27	85.27
	DRM	79.32	79.05	80.00	80.67	78.68	**81.55**	80.50	80.00	80.75	80.00	79.66	80.31
	Reg.	83.38	86.34	84.19	85.68	87.05	86.78	**87.22**	84.19	86.70	84.19	86.51	86.62
	Binary	**88.02**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	83.64	86.58	84.42	85.80	87.37	87.09	**87.60**	84.42	86.98	84.42	86.77	86.89
	D2U-KL	83.46	86.44	84.14	85.50	87.19	86.71	**87.27**	84.14	86.64	84.14	86.54	86.62
	D2U-S	79.67	81.74	80.42	81.69	82.23	**82.86**	82.57	80.42	82.41	80.42	82.13	82.27
SNIPS	MLE	95.45	96.20	95.50	95.70	**96.33**	95.61	95.83	95.50	95.86	95.50	96.15	96.04
	LMCL	85.18	87.54	88.20	90.45	**93.15**	89.91	93.08	88.20	91.69	88.20	91.87	91.76
	DRM	93.58	94.47	93.63	93.82	**94.58**	93.75	93.95	93.63	93.98	93.63	94.43	94.13
	Reg.	98.61	98.74	98.61	98.64	**98.76**	98.62	98.63	98.61	98.65	98.61	98.73	98.67
	Binary	**98.91**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	98.51	98.60	98.52	98.53	**98.61**	98.52	98.54	98.52	98.54	98.52	98.60	98.56
	D2U-KL	98.97	99.15	98.99	99.01	**99.16**	98.99	99.02	98.99	99.04	98.99	99.14	99.09
	D2U-S	98.24	98.37	98.25	98.30	**98.39**	98.29	98.32	98.25	98.32	98.25	98.37	98.34
TOP	MLE	74.23	73.23	74.26	74.19	73.24	74.32	**74.36**	74.26	74.18	74.26	73.25	73.76
	LMCL	70.62	70.11	70.72	70.88	70.11	70.58	**71.29**	70.72	70.95	70.72	70.43	70.70
	DRM	76.97	76.59	76.99	76.96	76.61	77.06	**77.13**	76.99	76.98	76.99	76.60	76.83
	Reg.	96.45	**96.57**	96.45	96.47	96.57	96.45	96.47	96.45	96.49	96.45	96.56	96.52
	Binary	**97.29**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	97.30	**97.42**	97.30	97.33	**97.42**	97.31	97.34	97.30	97.35	97.30	97.41	97.38
	D2U-KL	97.39	**97.50**	97.41	97.43	**97.50**	97.42	97.43	97.41	97.44	97.41	97.49	97.46
	D2U-S	94.68	**94.94**	94.71	94.76	**94.94**	94.75	94.78	94.71	94.80	94.71	94.92	94.87

Table 7: Average FPR90 score of 10-Fold binary OOS Detection. Row-wise lowest scores are given in bold.

Data	Training	MLE	Temp	Stdev	Ent	CE	BC	Cbr	Cos	JS	Euc	KL	Helng.
ACID	MLE	25.00	22.10	20.70	19.85	21.40	24.80	28.60	20.70	20.90	20.70	**19.70**	20.70
	LMCL	15.60	15.50	14.80	**13.83**	17.40	37.10	46.20	14.80	16.50	14.80	16.50	16.50
	DRM	19.90	20.80	18.40	**13.51**	24.80	14.40	16.30	18.40	15.00	18.40	18.80	15.30
	Reg.	10.30	8.60	9.70	8.41	7.20	7.40	**7.10**	9.70	7.50	9.70	8.20	7.60
	Binary	**6.17**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	10.30	8.20	9.40	8.54	7.30	7.50	7.40	9.40	**7.20**	9.40	7.70	7.30
	D2U-KL	10.70	8.90	9.90	**7.56**	7.90	8.00	7.80	9.90	8.10	9.90	8.10	8.10
	D2U-S	11.90	10.30	11.00	9.57	**8.80**	9.00	9.30	11.00	9.10	11.00	9.40	9.30
Banking	MLE	13.30	10.00	11.30	7.85	7.30	6.60	12.00	11.30	6.20	11.30	6.30	**6.10**
	LMCL	6.30	6.20	**5.50**	7.00	7.10	19.40	25.80	**5.50**	6.80	**5.50**	6.80	6.80
	DRM	13.10	8.40	10.20	6.06	9.20	**4.60**	4.90	10.20	5.20	10.20	7.30	5.30
	Reg.	2.40	1.80	2.20	**0.36**	1.60	1.60	1.40	2.20	1.70	2.20	1.60	1.70
	Binary	**0.16**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	2.40	1.60	2.10	**0.23**	1.00	1.00	1.00	2.10	1.00	2.10	1.00	1.00
	D2U-KL	2.70	1.60	2.50	**0.39**	1.70	1.70	1.60	2.50	1.70	2.50	1.80	1.70
	D2U-S	5.80	3.70	5.20	2.25	2.00	2.60	**1.90**	5.20	2.60	5.20	3.40	2.80
CLINC	MLE	9.30	7.80	9.30	8.00	**7.50**	8.10	7.70	9.30	7.70	9.30	7.80	7.60
	LMCL	7.40	**7.30**	7.60	8.51	8.60	18.30	35.40	7.60	8.40	7.60	8.50	8.40
	DRM	8.50	11.80	8.50	**7.64**	13.70	7.80	9.50	8.50	8.30	8.50	9.40	8.20
	Reg.	6.50	5.10	6.60	5.13	5.10	**4.80**	4.90	6.60	5.00	6.60	5.10	5.20
	Binary	**48.58**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	6.70	5.40	6.70	5.62	**5.10**	5.90	5.40	6.70	5.50	6.70	5.50	5.60
	D2U-KL	6.50	5.50	6.50	**4.89**	5.20	5.90	5.80	6.50	5.70	6.50	5.60	5.70
	D2U-S	7.10	4.50	7.10	6.38	**3.90**	4.80	4.60	7.10	4.60	7.10	4.80	4.50
HWU64	MLE	58.90	54.10	55.70	**48.12**	52.50	51.60	55.00	55.70	52.20	55.70	52.60	51.90
	LMCL	43.00	42.90	38.10	41.84	37.30	41.70	44.90	38.10	37.50	38.10	**37.10**	37.30
	DRM	56.90	51.10	53.00	**49.62**	54.30	52.40	53.00	53.00	50.80	53.00	51.50	51.40
	Reg.	46.50	32.70	41.20	41.84	**29.60**	31.30	**29.60**	41.20	31.80	41.20	32.60	32.10
	Binary	**35.77**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	44.90	35.10	40.80	40.43	31.70	31.70	**30.60**	40.80	32.50	40.80	33.50	33.10
	D2U-KL	43.80	33.70	39.90	42.48	**30.30**	31.50	31.10	39.90	33.00	39.90	33.30	33.10
	D2U-S	57.60	48.10	53.00	**46.84**	47.80	47.40	48.50	53.00	47.30	53.00	48.10	47.60
SNIPS	MLE	10.40	8.40	10.30	10.71	**8.40**	10.40	10.30	10.30	9.90	10.30	8.60	8.70
	LMCL	49.50	41.90	36.70	29.86	21.10	30.10	**16.40**	36.70	24.10	36.70	24.10	24.10
	DRM	13.40	10.20	13.40	10.86	**10.20**	13.40	13.30	13.40	12.40	13.40	10.40	11.50
	Reg.	2.50	**2.20**	2.50	2.29	**2.20**	2.50	2.50	2.50	2.40	2.50	**2.20**	2.30
	Binary	**1.71**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	**2.70**	**2.70**	**2.70**	3.14	**2.70**	**2.70**	**2.70**	**2.70**	**2.70**	**2.70**	**2.70**	**2.70**
	D2U-KL	2.10	**1.60**	2.10	2.14	**1.60**	2.10	2.10	2.10	1.90	2.10	**1.60**	**1.60**
	D2U-S	2.90	2.80	2.90	3.00	2.80	2.90	2.90	2.90	2.80	2.90	2.80	**2.70**
TOP	MLE	**51.38**	53.00	**51.38**	66.65	53.00	**51.38**	**51.38**	**51.38**	51.50	**51.38**	53.00	51.75
	LMCL	69.75	70.13	69.50	**65.06**	70.75	70.63	71.50	69.50	70.50	69.50	70.25	70.25
	DRM	**50.13**	51.88	**50.13**	59.79	51.88	**50.13**	50.13	**50.13**	50.88	**50.13**	51.63	50.63
	Reg.	7.50	7.25	7.50	**4.90**	7.25	7.50	7.50	7.50	7.50	7.50	7.25	7.25
	Binary	**4.43**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	6.13	6.25	6.13	**4.04**	6.25	6.13	6.13	6.13	6.13	6.13	6.25	6.13
	D2U-KL	6.25	5.88	6.25	**3.57**	5.88	6.25	6.38	6.25	6.25	6.25	5.88	6.00
	D2U-S	11.63	11.88	11.63	12.94	12.00	11.63	**11.50**	11.63	11.63	11.63	11.88	11.75

Table 8: Average FNR90 score of 10-Fold binary OOS Detection. Row-wise lowest scores are given in bold.

Data	Training	MLE	Temp	Stdev	Ent	CE	BC	Cbr	Cos	JS	Euc	KL	Helng.
ACID	MLE	27.30	25.67	26.22	**19.70**	21.46	20.60	26.70	26.22	20.45	26.22	21.38	20.75
	LMCL	16.36	16.30	**14.80**	15.80	16.57	28.78	38.57	**14.80**	15.58	**14.80**	15.30	15.39
	DRM	16.42	20.92	16.23	15.00	23.88	**13.31**	14.48	16.23	13.95	16.23	18.94	14.33
	Reg.	10.80	8.79	10.39	8.70	7.61	8.32	**7.48**	10.39	8.08	10.39	8.30	8.20
	Binary	**6.70**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	11.11	9.11	10.12	8.00	**7.96**	8.40	8.03	10.12	8.41	10.12	8.61	8.47
	D2U-KL	12.84	9.99	12.04	9.00	7.76	8.05	**7.53**	12.04	8.17	12.04	8.85	8.35
	D2U-S	12.69	10.41	11.88	10.20	**9.54**	9.68	10.00	11.88	9.87	11.88	9.76	9.71
Banking	MLE	14.36	10.49	11.37	**7.20**	8.01	8.79	13.68	11.37	7.88	11.37	7.69	7.79
	LMCL	8.05	7.95	6.78	**6.30**	9.09	18.60	25.96	6.78	8.44	6.78	8.34	8.37
	DRM	11.43	9.71	9.90	7.20	12.12	**4.76**	5.90	9.90	5.47	9.90	8.99	6.19
	Reg.	0.94	**0.42**	0.52	1.80	**0.42**	0.55	0.46	0.52	0.55	0.52	0.55	0.55
	Binary	**0.20**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	0.42	0.26	0.26	1.30	**0.23**	0.26	0.26	0.26	0.26	0.26	0.26	0.26
	D2U-KL	1.50	0.59	0.72	1.80	**0.39**	0.59	0.72	0.72	0.52	0.72	0.42	0.49
	D2U-S	4.85	3.58	3.68	3.70	**2.12**	2.38	2.57	3.68	2.41	3.68	2.74	2.57
CLINC	MLE	11.80	9.16	11.60	8.90	8.36	8.11	**7.56**	11.60	7.98	11.60	8.67	8.31
	LMCL	9.91	9.96	9.76	**8.30**	10.76	21.51	45.62	9.76	10.44	9.76	10.13	10.20
	DRM	9.69	16.31	9.67	7.90	21.27	**7.80**	10.36	9.67	8.84	9.67	12.62	9.69
	Reg.	6.82	5.42	6.73	6.00	**5.11**	5.18	5.36	6.73	5.20	6.73	5.36	5.18
	Binary	**31.40**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	7.40	6.13	7.33	6.60	6.18	6.18	6.11	7.33	6.04	7.33	6.09	**6.02**
	D2U-KL	6.60	5.07	6.42	6.20	4.93	**4.82**	**4.82**	6.42	4.87	6.42	4.96	**4.82**
	D2U-S	8.33	5.98	8.16	6.50	**5.36**	5.84	5.53	8.16	5.84	8.16	6.00	5.89
HWU64	MLE	51.97	52.26	51.75	52.80	52.01	47.14	**47.01**	51.75	48.63	51.75	51.03	49.57
	LMCL	48.97	48.55	47.91	**37.40**	47.01	47.86	52.18	47.91	47.52	47.91	47.78	47.78
	DRM	**50.34**	62.91	50.47	51.10	62.91	51.88	57.31	50.47	57.39	50.47	60.81	59.74
	Reg.	45.17	41.62	45.13	**34.70**	40.60	41.88	40.09	45.13	40.77	45.13	41.50	40.81
	Binary	**31.00**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	45.43	38.59	45.17	**36.10**	37.05	41.58	37.09	45.17	40.34	45.17	39.10	39.87
	D2U-KL	45.56	42.74	45.64	**35.90**	41.50	40.47	39.19	45.64	41.84	45.64	42.56	41.88
	D2U-S	50.00	49.23	49.96	48.70	49.10	45.77	**45.21**	49.96	47.99	49.96	49.06	48.63
SNIPS	MLE	11.14	9.29	11.14	11.80	**9.29**	11.14	11.14	11.14	10.71	11.14	9.71	10.29
	LMCL	31.86	30.86	31.43	30.30	**23.29**	31.86	26.00	31.43	29.43	31.43	28.29	28.43
	DRM	12.00	10.57	12.00	14.30	10.57	12.00	12.00	12.00	11.71	12.00	**10.43**	11.71
	Reg.	2.71	2.14	2.57	3.10	**2.00**	2.43	2.43	2.57	2.43	2.57	2.29	2.29
	Binary	**2.00**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	3.29	3.00	3.29	2.90	**2.86**	3.29	3.29	3.29	3.29	3.29	3.14	3.14
	D2U-KL	2.86	1.86	2.86	2.30	**1.57**	2.71	2.57	2.86	2.29	2.86	1.86	2.14
	D2U-S	3.43	2.43	3.43	3.10	**2.29**	3.29	3.14	3.43	2.86	3.43	2.43	2.86
TOP	MLE	67.88	69.05	67.89	**51.50**	69.07	67.89	67.90	67.89	68.17	67.89	69.04	68.46
	LMCL	**66.54**	72.06	66.61	69.88	72.97	**66.54**	68.03	66.61	69.61	66.61	71.76	71.35
	DRM	62.51	62.65	62.51	**50.38**	62.66	62.51	62.51	62.51	62.39	62.51	62.68	62.55
	Reg.	7.06	6.07	7.05	7.50	**6.02**	7.06	6.76	7.05	6.63	7.05	6.17	6.44
	Binary	**5.75**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	5.10	4.22	5.08	6.13	**4.03**	4.93	4.42	5.08	4.62	5.08	4.32	4.48
	D2U-KL	5.25	4.25	5.21	6.38	**4.10**	5.22	4.55	5.21	4.67	5.21	4.32	4.54
	D2U-S	14.98	15.59	14.99	**11.63**	15.61	14.98	15.00	14.99	15.10	14.99	15.64	15.19

Table 9: Average Recall score of 10-Fold binary OOS Detection. Row-wise highest scores are given in bold.

Data	Training	MLE	Temp	Stdev	Ent	CE	BC	Cbr	Cos	JS	Euc	KL	Helng.
ACID	MLE	84.86	85.76	87.23	86.88	86.43	85.80	84.10	87.23	87.48	87.23	**87.81**	87.18
	LMCL	88.35	**88.79**	**88.79**	87.47	86.84	80.77	77.23	**88.79**	87.04	**88.79**	87.26	87.07
	DRM	86.81	89.53	88.83	90.02	89.63	**90.73**	90.33	88.83	90.70	88.83	90.23	90.56
	Reg.	95.58	95.76	95.92	96.03	96.00	96.08	**96.25**	95.92	96.05	95.92	96.06	96.05
	Binary	**96.45**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	95.11	95.28	95.44	95.65	95.98	95.92	**96.03**	95.44	95.90	95.44	95.86	95.90
	D2U-KL	95.64	95.91	96.04	96.23	96.31	96.30	**96.38**	96.04	96.26	96.04	96.25	96.25
	D2U-S	92.43	93.05	**93.39**	93.34	93.18	92.85	91.56	**93.39**	93.12	**93.39**	93.36	93.26
Banking	MLE	89.36	90.29	90.88	91.62	91.47	91.57	89.78	90.88	92.04	90.88	**92.09**	92.04
	LMCL	92.56	92.78	93.19	92.65	92.43	86.81	85.04	**93.19**	92.65	93.19	92.53	92.68
	DRM	90.59	93.32	92.01	93.69	92.83	94.45	94.10	92.01	**94.55**	92.01	93.49	94.30
	Reg.	96.83	96.98	96.93	97.17	97.40	97.42	97.40	96.93	97.40	96.93	97.35	**97.47**
	Binary	**97.84**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	**96.76**	96.63	96.68	96.49	96.66	96.63	96.61	96.68	96.54	96.68	96.54	96.54
	D2U-KL	97.10	97.20	97.25	97.10	97.47	**97.57**	97.54	97.25	97.44	97.25	97.47	97.44
	D2U-S	94.52	95.41	95.41	96.02	95.90	95.87	95.80	95.41	96.04	95.41	**96.12**	96.07
CLINC	MLE	90.22	90.05	90.29	90.25	**91.31**	91.07	91.04	90.29	90.82	90.29	90.27	90.45
	LMCL	90.78	90.80	91.20	**91.40**	91.20	88.44	81.44	91.20	91.27	91.20	91.11	91.15
	DRM	90.95	92.53	90.96	91.75	91.85	92.73	92.75	90.96	**92.93**	90.96	92.73	92.76
	Reg.	93.29	93.31	93.49	93.38	93.31	93.33	93.33	93.49	**93.56**	93.49	93.49	**93.56**
	Binary	**88.31**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	92.35	92.84	92.60	92.87	93.27	**93.33**	93.11	92.60	93.13	92.60	93.31	93.00
	D2U-KL	93.16	**93.64**	93.35	93.05	93.33	93.13	93.02	93.35	93.62	93.35	93.09	93.55
	D2U-S	93.42	94.55	93.93	94.58	94.53	94.67	94.65	93.93	94.67	93.93	94.62	**94.87**
HWU64	MLE	73.02	76.65	74.79	**77.66**	76.83	77.34	75.57	74.76	77.40	74.79	77.57	77.31
	LMCL	80.33	80.24	81.32	81.38	81.92	80.75	79.10	81.32	81.95	81.32	81.83	**81.98**
	DRM	73.44	77.01	76.29	77.22	76.77	**77.99**	77.51	76.29	77.69	76.29	77.22	77.43
	Reg.	73.95	74.58	74.43	74.43	**74.94**	74.82	74.73	74.43	74.70	74.43	74.79	**74.79**
	Binary	**74.70**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	73.23	**75.09**	73.89	74.22	74.58	74.40	74.76	73.89	74.16	73.89	74.16	74.10
	D2U-KL	73.74	74.76	74.07	74.58	75.27	74.97	**75.54**	74.07	74.43	74.07	75.27	74.79
	D2U-S	74.25	**75.00**	74.16	74.37	74.28	74.37	73.89	74.16	74.40	74.16	74.55	74.73
SNIPS	MLE	88.29	**88.41**	88.29	88.24	88.35	88.29	88.06	88.29	88.24	88.29	**88.41**	88.06
	LMCL	67.76	71.76	74.65	79.35	83.12	79.94	**85.82**	74.65	80.71	74.65	80.65	80.71
	DRM	87.88	**89.24**	87.88	88.41	89.18	87.88	88.06	87.88	88.88	87.88	89.12	88.88
	Reg.	88.41	88.94	88.35	87.94	**89.53**	88.47	89.06	88.35	89.35	88.35	88.88	89.35
	Binary	**86.18**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	88.24	89.41	88.65	87.76	**89.47**	89.18	89.12	88.65	89.18	88.65	**89.47**	89.24
	D2U-KL	88.53	88.59	88.53	88.00	88.59	**88.94**	88.76	88.53	88.71	88.53	88.65	88.76
	D2U-S	87.00	90.29	86.41	87.82	90.29	89.35	89.35	86.41	89.65	86.41	**90.35**	90.29
TOP	MLE	**84.85**	84.52	**84.85**	84.75	84.54	**84.85**	84.71	**84.85**	84.71	**84.85**	84.49	83.14
	LMCL	57.78	68.92	57.53	61.81	**75.63**	58.43	67.12	57.53	65.35	57.53	68.10	64.83
	DRM	81.68	72.93	81.68	79.17	72.87	81.68	**81.78**	81.68	79.49	81.68	72.80	75.56
	Reg.	94.51	95.23	94.51	94.76	95.07	94.77	**95.60**	94.51	95.07	94.51	95.04	94.97
	Binary	**96.95**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	93.85	94.34	93.84	94.00	94.55	94.00	**94.90**	93.84	94.56	93.84	94.33	94.49
	D2U-KL	94.21	95.07	94.21	94.29	95.17	94.52	**95.55**	94.21	95.25	94.21	95.17	95.35
	D2U-S	91.29	91.92	91.29	91.70	**92.13**	91.29	91.36	91.29	91.90	91.29	91.97	91.59

Table 10: Average F1 score of 10-Fold binary OOS Detection. Row-wise highest scores are given in bold.

Data	Training	MLE	Temp	Stdev	Ent	CE	BC	Cbr	Cos	JS	Euc	KL	Helng.
ACID	MLE	87.51	88.21	89.24	89.03	88.69	88.24	86.95	89.24	89.45	89.24	**89.71**	89.24
	LMCL	90.14	90.46	**90.47**	89.52	89.05	84.60	81.99	**90.47**	89.20	**90.47**	89.35	89.22
	DRM	89.00	90.83	90.42	91.34	90.83	**91.85**	91.51	90.42	91.81	90.42	91.40	91.70
	Reg.	95.10	95.28	95.46	95.58	95.58	95.67	**95.89**	95.46	95.62	95.46	95.65	95.63
	Binary	**96.07**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	94.43	94.60	94.80	95.09	95.55	95.47	**95.61**	94.80	95.46	94.80	95.42	95.46
	D2U-KL	95.19	95.55	95.64	95.88	96.01	95.99	**96.08**	95.64	95.94	95.64	95.93	95.93
	D2U-S	93.16	93.66	**93.90**	93.87	93.78	93.51	92.59	**93.90**	93.71	**93.90**	93.89	93.81
Banking	MLE	89.60	90.53	91.03	91.76	91.67	91.74	90.04	91.03	92.17	91.03	**92.22**	92.18
	LMCL	92.63	92.87	**93.27**	92.75	92.52	87.21	85.42	**93.27**	92.72	**93.27**	92.61	92.74
	DRM	90.50	93.16	91.89	93.54	92.60	94.36	93.99	91.89	**94.42**	91.89	93.31	94.19
	Reg.	96.75	96.91	96.85	97.12	97.35	97.38	97.35	96.85	97.35	96.85	97.30	**97.43**
	Binary	**97.79**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	**96.66**	96.54	96.59	96.37	96.55	96.53	96.50	96.59	96.43	96.59	96.43	96.43
	D2U-KL	97.03	97.13	97.18	97.02	97.42	**97.52**	97.50	97.18	97.39	97.18	97.41	97.39
	D2U-S	94.48	95.36	95.36	95.98	95.88	95.86	95.80	95.36	96.02	95.36	**96.07**	96.03
CLINC	MLE	90.75	90.65	90.82	90.80	**91.75**	91.54	91.50	90.82	91.32	90.82	90.84	91.00
	LMCL	91.27	91.29	91.61	**91.79**	91.59	88.98	82.71	91.61	91.68	91.61	91.54	91.57
	DRM	91.39	92.66	91.39	92.08	91.94	92.96	92.93	91.39	**93.12**	91.39	92.90	92.97
	Reg.	92.89	92.89	93.12	92.96	92.91	92.95	92.94	93.12	**93.22**	93.12	93.13	**93.22**
	Binary	**86.07**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	91.64	92.26	91.98	92.30	92.84	**92.87**	92.63	91.98	92.61	91.98	92.86	92.46
	D2U-KL	92.70	**93.28**	92.90	92.54	92.91	92.65	92.55	92.90	93.23	92.90	92.58	93.15
	D2U-S	93.46	94.55	93.92	94.58	94.53	94.64	94.61	93.92	94.66	93.92	94.63	**94.86**
HWU64	MLE	73.72	76.50	75.08	**77.33**	76.03	76.96	75.26	75.06	76.80	75.08	77.07	76.79
	LMCL	80.18	80.08	81.09	81.16	**81.77**	80.52	78.90	81.09	81.69	81.09	81.53	81.74
	DRM	74.01	76.66	76.06	76.82	76.11	**77.68**	76.98	76.06	77.30	76.06	76.90	77.00
	Reg.	66.95	68.19	67.81	67.85	**68.80**	68.52	68.49	67.81	68.21	67.81	68.37	68.37
	Binary	**67.83**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	65.55	**69.02**	66.65	67.39	68.18	67.69	68.60	66.65	67.29	66.65	67.45	67.31
	D2U-KL	66.42	68.42	67.16	68.23	69.41	68.91	**69.88**	67.16	67.81	67.16	69.37	68.46
	D2U-S	69.28	**69.93**	68.67	68.85	68.30	68.39	67.82	68.67	68.64	68.67	68.95	69.31
SNIPS	MLE	88.37	**88.49**	88.37	88.31	88.43	88.37	88.13	88.37	88.31	88.37	**88.49**	88.13
	LMCL	66.27	71.04	74.22	79.27	83.13	79.87	**85.91**	74.22	80.68	74.22	80.61	80.68
	DRM	87.94	**89.29**	87.94	88.46	89.23	87.94	88.11	87.94	88.93	87.94	89.17	88.94
	Reg.	88.46	88.99	88.40	87.98	**89.57**	88.52	89.10	88.40	89.40	88.40	88.93	89.40
	Binary	**86.15**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	88.29	89.45	88.69	87.81	**89.52**	89.23	89.17	88.69	89.22	88.69	**89.52**	89.28
	D2U-KL	88.58	88.64	88.58	88.05	88.64	**89.00**	88.82	88.58	88.76	88.58	88.70	88.82
	D2U-S	87.02	90.36	86.43	87.87	90.36	89.42	89.42	86.43	89.71	86.43	**90.42**	90.36
TOP	MLE	**86.57**	86.13	**86.57**	86.50	86.14	**86.57**	86.47	**86.57**	86.48	**86.57**	86.11	85.19
	DRM	84.41	77.62	84.41	82.42	77.56	84.41	**84.48**	84.41	82.64	84.41	77.53	79.50
	Reg.	94.93	95.55	94.93	95.14	95.42	95.15	**95.85**	94.93	95.41	94.93	95.39	95.32
	LMCL	66.25	75.13	66.04	69.48	**80.25**	66.71	73.74	66.04	72.34	66.04	74.48	71.90
	Binary	**96.95**	-	-	-	-	-	-	-	-	-	-	-
	D2U-CE	94.42	94.83	94.41	94.55	95.01	94.55	**95.29**	94.41	95.00	94.41	94.82	94.95
	D2U-KL	94.71	95.43	94.71	94.78	95.51	94.97	**95.84**	94.71	95.58	94.71	95.51	95.67
	D2U-S	92.29	92.78	92.29	92.61	**92.95**	92.29	92.35	92.29	92.77	92.29	92.82	92.53