Supplementary Material for 'MONET: Multi-Modal Online Continual Learning with Novelty Estimation'

Evelyn Chee, Wynne Hsu, Mong Li Lee
School of Computing, National University of Singapore
{echee,dcshsuw,dcsleeml}@comp.nus.edu.sg

S1. Hyperparameter Setting and Analysis

The hyperparameter values are determined through a systematic local search. Specifically, we search over $\eta_1 \in [0.01, 5.0]$ and $\eta_2 \in [0.01, 5.0]$ for the loss weights, $\alpha \in [0, 1]$ for the uncertainty threshold smoothing factor, and $T \in [1, 1000]$ for the uncertainty threshold update interval. Consistent performance trends are observed across all three datasets. Figure S1 presents the average F1_{all} scores on the CUB200 dataset when varying the individual hyperparameter values.

In Figure S1(a), we observe the impact of the weight on distillation loss, η_1. Initially, the average F1_{all} score improves with increasing η_1 and remains stable within the range of $[0.1, 1.0]$. However, the score drops significantly when η_1 is set too high. Figure S1(b) illustrates the effect of the weight on entropy loss, η_2. Similarly, performance improves as η_2 increases and remains relatively stable within the range of $[0.1, 1.0]$.

In Figure S1(c), we see that performance improves as the smoothing factor α increases up to 0.8 due to the enhanced stability of the uncertainty threshold. However, when α is set too high, the uncertainty threshold does not adapt quickly enough, resulting in large performance drop. Lastly, Figure S1(d) shows that that frequent updates of the uncertainty threshold T are unnecessary. The update interval can be increased up to 100 while still maintaining robust performance.

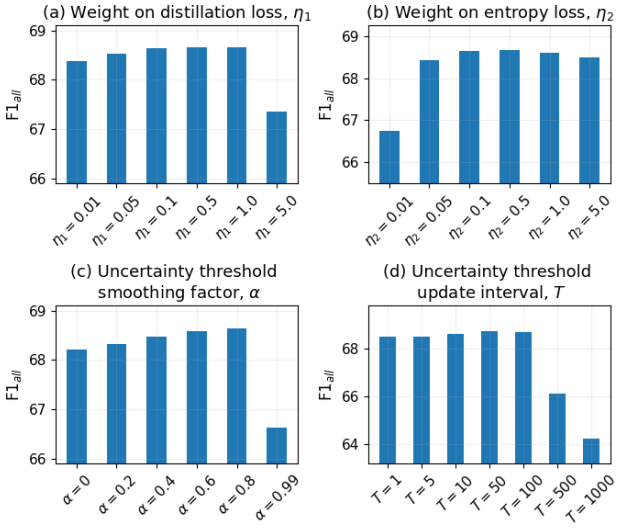

Figure S1. Average F1_{all} score with varying hyperparameters values on the CUB200 dataset.

S2. Analysis on Percentile of Dropped Features

We vary the percentage of features dropped when generating pseudo representations in MONET. Table S1 shows the average F1_{all} scores. The optimal K generally falls between 40% and 60%, with 50% being the most robust. Dropping too few features can result in inaccurate uncertainty thresholds as the pseudo representations do not sufficiently reflect OOD instances, while dropping too many features can reduce diversity of the pseudo representations and diminishes their effectiveness as OOD instances.

Table S1. Average F1_{all} score for different top-K percentile of dropped features.

K	AVE	UESTC-MMEA	CUB200
30	64.45$\pm$2.04	75.28$\pm$0.67	63.71$\pm$0.48
40	65.42$\pm$2.41	**75.49**$\pm$0.93	67.16$\pm$0.51
50	65.52$\pm$2.55	75.40$\pm$0.73	**68.64**$\pm$0.63
60	**65.70**$\pm$2.86	74.69$\pm$0.76	68.52$\pm$0.75
70	65.26$\pm$2.92	73.92$\pm$0.94	67.22$\pm$0.79

S3. Robustness Against Baselines with Varying Uncertainty Threshold Percentiles

In the main paper, baseline results are reported using the 95$^{\text{th}}$ percentile to determine class-specific uncertainty thresholds. Here, we present additional results where different percentile values are applied to the baselines for deriving the uncertainty thresholds. Figure S2 shows the average F1_{all} scores of the baseline methods across percentiles

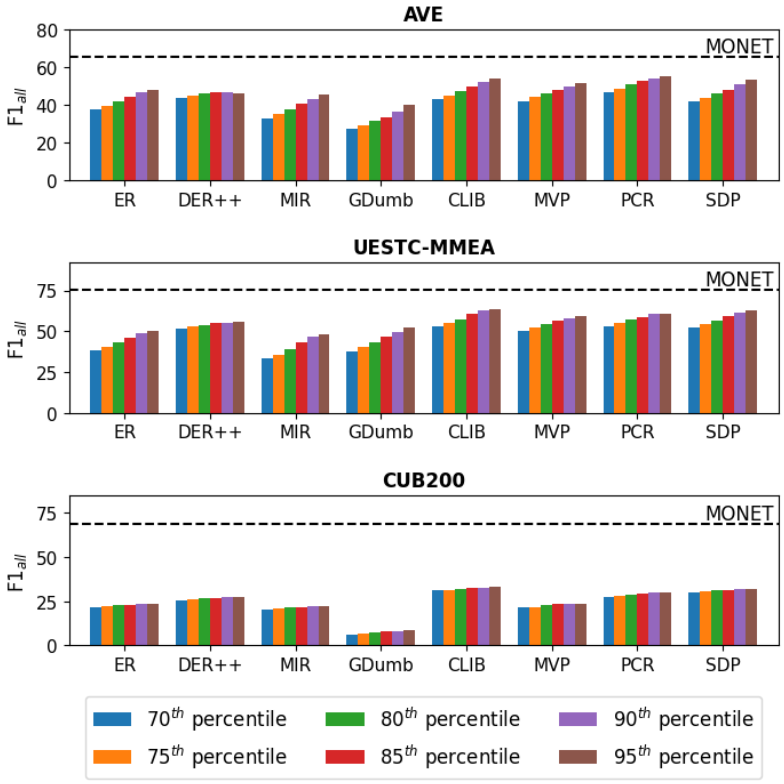

Figure S2. Average F1_{all} score of all baseline methods using different percentiles for computing the uncertainty thresholds.

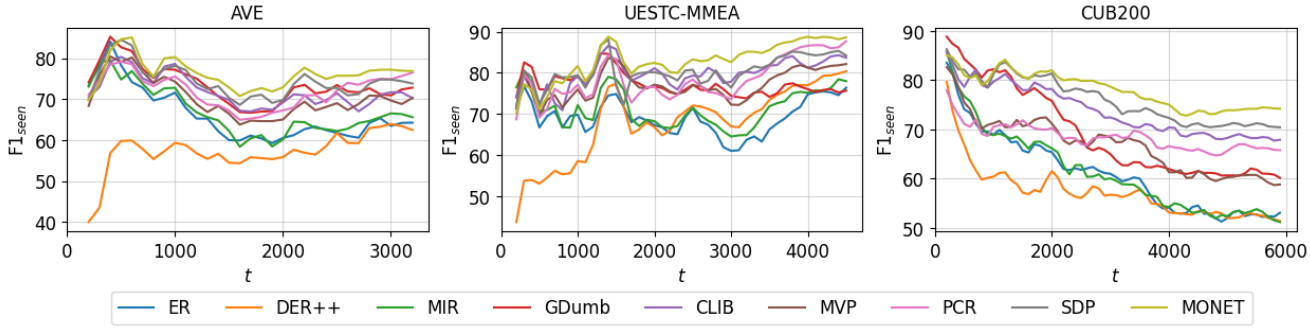

Figure S3. F1_{seen} across time steps t on the three datasets.

ranging from the 70th to the 95th. We see that MONET consistently outperforms the baselines, regardless of the percentile used.

S4. Performance Analysis Across Time Steps

From Figure 3 in the main paper, we see that MONET consistently outperforms the baselines at all time steps in terms of overall classification performance, with the performance gap widening over time. To gain deeper insights, we break down the results and examine the classification accuracy of seen class samples and the novelty detection capabilities of the methods across time steps.

Figure S3 presents the classification accuracy for seen classes (F1_{seen}), showing that MONET increasingly outperforms the baselines as learning progresses. For novelty detection, we observe from the false positive rates in Figure S4 that the baseline methods increasingly misidentify samples from newly learned classes as unknown. In contrast, MONET remains robust in correctly identifying samples from unknown classes as learning progresses.

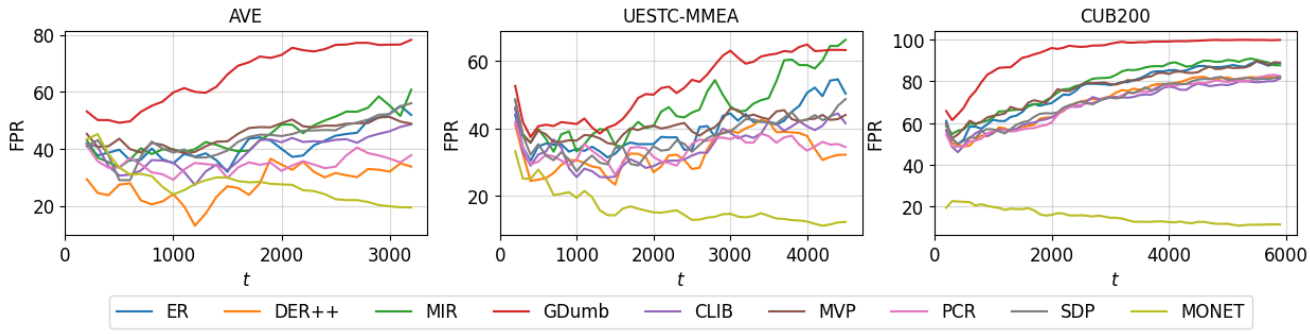

Figure S4. False positive rate (FPR) for identifying samples as unknown across time steps t on the three datasets.

Table S2. Results across different values of standard deviation σ for Gaussian data stream on the CUB200 dataset.

	$\sigma = 0.01$				$\sigma = 0.25$				$\sigma = 0.5$			
Method	$\text{F1}_{all}\uparrow$	$\text{F1}_{seen}\uparrow$	$\text{F1}_{novel}\uparrow$	KLR $\downarrow$	$\text{F1}_{all}\uparrow$	$\text{F1}_{seen}\uparrow$	$\text{F1}_{novel}\uparrow$	KLR $\downarrow$	$\text{F1}_{all}\uparrow$	$\text{F1}_{seen}\uparrow$	$\text{F1}_{novel}\uparrow$	KLR $\downarrow$
ER	32.20$\pm$1.20	66.37$\pm$1.69	60.24$\pm$0.88	50.52$\pm$1.19	17.23$\pm$0.97	55.95$\pm$0.92	33.61$\pm$2.21	25.79$\pm$1.02	14.62$\pm$1.12	55.56$\pm$0.35	23.89$\pm$2.36	24.00$\pm$0.23
DER++	31.69$\pm$0.71	62.52$\pm$0.21	60.17$\pm$0.62	51.00$\pm$1.10	22.17$\pm$0.73	52.85$\pm$0.61	33.58$\pm$2.27	26.91$\pm$0.45	21.39$\pm$1.23	53.21$\pm$0.79	23.91$\pm$2.09	25.46$\pm$0.56
MIR	31.86$\pm$1.27	67.86$\pm$1.46	60.85$\pm$1.47	53.51$\pm$0.33	15.68$\pm$0.59	56.36$\pm$1.04	33.77$\pm$2.40	27.18$\pm$0.67	13.12$\pm$0.51	56.51$\pm$0.51	24.05$\pm$2.23	25.16$\pm$0.84
GDumb	18.93$\pm$0.76	76.83$\pm$1.26	60.74$\pm$1.03	45.89$\pm$1.13	2.75$\pm$0.37	64.64$\pm$0.74	32.85$\pm$2.04	18.14$\pm$0.17	0.93$\pm$0.03	62.24$\pm$0.52	23.24$\pm$2.12	17.52$\pm$0.10
CLIB	32.96$\pm$0.91	78.86$\pm$0.53	61.22$\pm$0.84	35.77$\pm$1.25	29.23$\pm$1.07	71.78$\pm$0.90	34.83$\pm$2.29	16.03$\pm$0.60	29.08$\pm$0.65	72.45$\pm$0.60	24.35$\pm$2.34	13.69$\pm$0.42
MVP	30.86$\pm$0.82	72.27$\pm$0.91	60.89$\pm$0.75	26.34$\pm$1.04	18.00$\pm$1.26	63.00$\pm$0.66	33.75$\pm$2.32	17.14$\pm$0.33	15.76$\pm$1.43	63.33$\pm$0.75	23.80$\pm$2.61	15.13$\pm$0.18
PCR	35.19$\pm$0.71	77.50$\pm$0.37	61.18$\pm$1.19	25.83$\pm$0.59	21.10$\pm$0.82	69.33$\pm$0.36	34.06$\pm$2.49	13.83$\pm$0.15	19.20$\pm$0.83	70.08$\pm$0.67	24.35$\pm$2.38	12.05$\pm$0.46
SDP	30.92$\pm$0.91	80.75$\pm$0.88	61.50$\pm$0.84	29.69$\pm$0.62	28.40$\pm$0.52	74.47$\pm$0.22	34.80$\pm$2.30	12.31$\pm$0.25	28.13$\pm$0.71	74.80$\pm$0.62	24.20$\pm$2.29	11.17$\pm$0.09
MONET	**67.82**$\pm$0.21	**82.61**$\pm$0.45	**63.22**$\pm$2.28	**25.77**$\pm$0.39	**71.35**$\pm$0.58	**76.95**$\pm$0.62	**40.90**$\pm$5.16	**9.29**$\pm$0.27	**72.18**$\pm$0.55	**77.07**$\pm$0.36	**30.86**$\pm$5.33	**8.25**$\pm$0.11

S5. Robustness with Varying Gaussian Spread

We also compare the continual learning methods on Gaussian data streams with different standard deviations σ. A larger standard deviation implies a greater overlap in the data spread between different classes.

Table S2 shows the results of all methods on the CUB200 dataset with σ set to 0.01, 0.25, and 0.5. The baseline methods exhibit a general decline in the average F1_{all} scores as σ increases. In contrast, MONET maintains high and consistent F1_{all} scores across all σ values. MONET also consistently outperforms the baseline methods in classification accuracy on seen classes and demonstrates reduced forgetting, as evidenced by achieving the highest F1_{seen} and lowest KLR scores.