Supplementary Materials

Anonymous Authors

A COMPARISON METHODS

(1) **Random**: Construct a coreset consisting of examples chosen from the full training set by uniform random sampling.
(2) **Forgetting** [7]: Construct a coreset composed of examples with the highest forgetting scores. The forgetting score counts how many times the forgetting happens during model training, i.e. an example was misclassified in the current epoch after being correctly classified in the previous epoch.
(3) **Entropy** [1]: Construct a coreset of examples with the highest entropy score. It is an uncertainty-based method. Entropy indicates the uncertainty of a sample given a certain classifier and training epoch, and examples with higher entropy are more important for model training.
(4) **EL2N** [4]: Construct a coreset of examples with the highest EL2N score. As an approximation of the GraNd score [4], which measures the average contribution of each sample to the decline of the training loss at early epochs across several independent runs, the EL2N score measures the data difficulty or importance by the L2 norm of error vectors.
(5) **Area under the margin (AUM)** [5]: Construct a coreset consisting of examples with the lowest AUM score. AUM is a data difficulty and importance metric that identifies noisy and mislabeled data by observing a network's training dynamics. (It measures the probability gap between the target class and the next largest class across all training epochs.)
(6) **Coverage-Centric Coreset Selection (CCS)** [8]: CCS jointly considers overall data coverage across a distribution and the importance of individual examples by employing a modified stratified sampling technique. In our experiments, AUM is used as the metric for determining the importance within the CCS framework.

B PERFORMANCE AT HIGH SELECION RATES

We provide more comparison results between EVA and other SOTA baselines at high selection rates. As reported in Tab. 1, EVA consistently exhibits superior performance in the majority of cases. Notably, EVA outperforms the full dataset at high selection rates, for instance, it achieves 98.80% accuracy using only half of the OrganAMNIST data, compared to the 98.39% accuracy with the full dataset, underlining its capability to maintain or even enhance model performance despite utilizing a pruned dataset.

Table 1: High selection rate performance on OrganAMNIST and OrganSMNIST with ResNet-18. The models trained with the full datasets achieves 98.39% and 91.76%, respectively. The first and second best results in each column are marked in red and blue, respectively.

	OrganAMNIST			OrganSMNIST		
α	50%	70%	90%	50%	70%	90%
Random	98.14 ±0.04	98.29 ±0.15	98.44 ±0.43	89.63 ±0.68	90.23 ±0.72	91.16 ±0.23
Forgetting [7]	98.46 ±0.16	98.31 ±0.82	98.70 ±0.44	91.18 ±0.69	91.46 ±0.38	91.58 ±0.26
Entropy [1]	97.93 ±0.44	98.32 ±0.03	98.50 ±0.55	90.41 ±0.27	91.31 ±0.50	91.36 ±0.78
EL2N [4]	98.39 ±0.89	98.24 ±0.04	98.48 ±1.27	89.94 ±0.57	90.43 ±0.90	91.53 ±0.61
AUM [5]	98.13 ±0.02	98.54 ±0.76	98.55 ±0.51	89.81 ±0.52	90.97 ±0.87	91.58 ±0.03
CCS [8]	97.55 ±0.19	97.56 ±0.60	97.07 ±0.81	85.53 ±0.11	85.69 ±0.15	86.38 ±0.91
EVA (Ours)	98.80 ±0.56	98.96 ±0.73	98.68 ±1.45	91.21 ±0.98	91.94 ±0.74	91.89 ±0.96

C EFFECTIVENESS ON NATURAL IMAGE DATASET

The applicability of EVA extends beyond medical imagery, as evidenced by our exploration of its effectiveness on natural image datasets such as CIFAR-10 and CIFAR-100. As detailed in Tab. 2 and Tab. 3, our method demonstrates robust performance across varying selection rates. On CIFAR-10, EVA attains 90.50% accuracy at 30% selection rate, closely approaching the full dataset's accuracy benchmark of 93.06%. In the more complex CIFAR-100 dataset, EVA achieves commendable results at low selection rates, i.e., reaching 62.93% accuracy with 30% of the full dataset. In addition, EVA 's performance consistently outpaces various SOTA baselines across different low selection rates tested, showcasing its generalizability and the robustness of its coreset selection efficacy in diverse image contexts.

D PARAMETER SETTINGS

We train the ResNet-18 over 200 epochs with a batch size of 256. For networks update, SGD optimizer with momentum of 0.9 and weight decay of 0.0005 is used. The learning rate is initialized as 0.1 and decays with the cosine annealing scheduler.
Besides, this section details the optimal window combinations identified for each dataset and selection rate assessed in our study. We represent each window combination of early and late stages, (t_e, t_E)+(t_l, t_L), more concisely as (t_e, t_l), since we set the window size to $K = 10$ throughout our experiments. For each dataset, We list the optimal (t_e, t_l) of every selection rate α as follows in the format of (t_e, t_l, α).

- For OrganAMNIST, the optimal settings are (1, 190, 2%), (1, 190, 5%), (100, 190, 10%), (1, 150, 20%), (90, 150, 30%), (90, 150, 50%), (1, 190, 70%), (90, 190, 90%).
- For OrganSMNIST, the optimal settings are (90, 100, 2%), (150, 190, 5%), (100, 190, 10%), (100, 190, 20%), (170, 190, 30%), (150, 190, 50%), (90, 150, 70%), (1, 100, 90%).
- For CIFAR-10, the optimal settings are (1, 100, 2%), (1, 150, 5%), (1, 190, 10%), (1, 100, 20%), (100, 190, 30%).
- For CIFAR-100, the optimal settings are (170, 190, 2%), (100, 190, 5%), (170, 190, 10%), (100, 190, 20%), (90, 190, 30%).

Table 2: Performances on natural image dataset CIFAR-10 with ResNet-18. The model trained with the full dataset achieves 93.06% accuracy.

α	2%	5%	10%	20%	30%
Random	41.64 ± 0.92%	58.62 ± 0.29%	71.64 ± 0.47%	84.57 ± 0.33%	89.79 ± 0.32%
Forgetting [7]	36.20 ± 0.24%	41.67 ± 0.58%	52.29 ± 0.33%	76.00 ± 1.45%	90.27 ± 0.88%
Entropy [1]	32.08 ± 0.42%	47.70 ± 0.39%	60.52 ± 0.15%	75.69 ± 0.90%	86.46 ± 0.57%
EL2N [4]	10.54 ± 0.49%	15.94 ± 0.61%	23.45 ± 0.86%	42.62 ± 0.68%	81.69 ± 1.27%
AUM [5]	14.66 ± 0.52%	18.54 ± 0.59%	25.35 ± 0.22%	49.51 ± 1.10%	73.92 ± 0.72%
CCS [8]	43.95 ± 1.66%	51.45 ± 2.18%	71.78 ± 1.98%	85.53 ± 0.97%	89.70 ± 0.65%
EVA (Ours)	**46.27 ± 0.37%**	**61.75 ± 0.57%**	**73.73 ± 0.42%**	**85.12 ± 0.68%**	**90.50 ± 0.49%**

Table 3: Performances on natural image dataset CIFAR-100 with ResNet-18. The model trained with the full dataset achieves 78.46% accuracy.

α	2%	5%	10%	20%	30%
Random	13.35 ± 0.39%	20.53 ± 0.93%	37.10 ± 1.01%	53.68 ± 1.33%	62.74 ± 0.15%
Forgetting [7]	6.86 ± 0.08%	10.14 ± 0.32%	16.87 ± 0.12%	26.18 ± 0.61%	38.25 ± 0.69%
Entropy [1]	8.92 ± 0.40%	14.64 ± 0.50%	25.01 ± 0.46%	40.33 ± 0.24%	48.95 ± 0.46%
EL2N [4]	3.63 ± 0.02%	5.16 ± 0.22%	7.26 ± 0.22%	14.65 ± 0.87%	34.83 ± 0.50%
AUM [5]	3.92 ± 0.03%	5.25 ± 0.04%	8.38 ± 0.29%	16.64 ± 0.07%	31.34 ± 0.49%
CCS [8]	13.50 ± 0.47%	23.84 ± 1.07%	36.39 ± 1.94%	53.14 ± 1.34%	64.72 ± 0.21%
EVA (Ours)	**13.28 ± 0.33%**	**24.38 ± 0.87%**	**39.60 ± 0.46%**	**55.86 ± 0.92%**	**62.93 ± 0.37%**

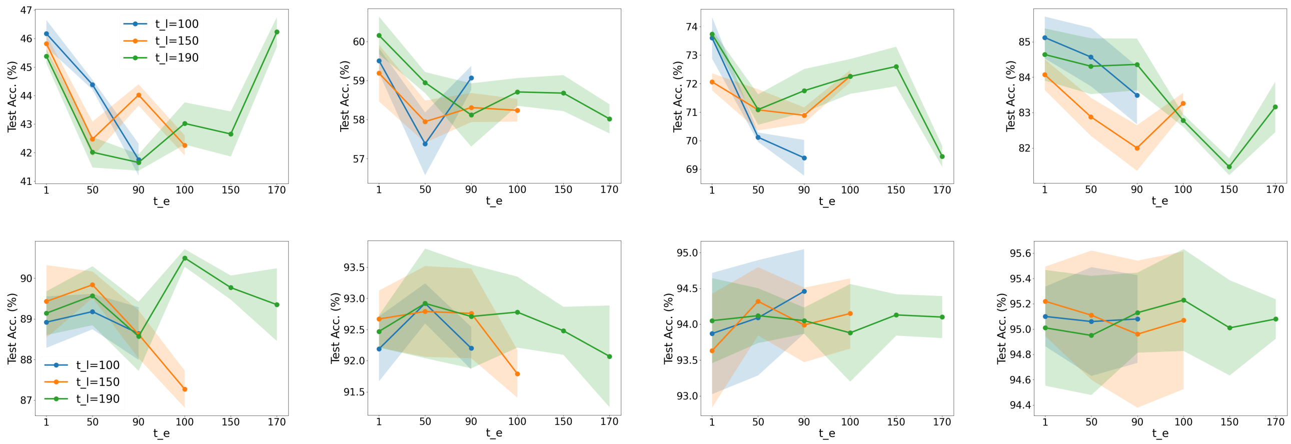

Figure 1: Window combinations on CIFAR-10. Different colors indicate the start epoch of different late windows t_l, and x-axis represents the start epoch of the early window t_e. From left to right and top to bottom, the corresponding selection rates are 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, and 0.9.

As reported in Fig. 1, we analyze the influence of window combination on performance. For a smaller selection rate, we should select samples earlier in training, and as the selection rate increase, the data budgets also increase, therefore the optimal window combination gradually slides from early to later stage.

Table 4: Cross-architecture generalization performance. We train ResNet-50, MobileNet-v2 and LeNet models with coresets of OrganSMNIST selected by scores calculated on training dynamics with ResNet-18.

	ResNet-50				**MobileNet-v2**				**LeNet**			
α	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
Random	23.73	76.81	82.13	83.79	48.10	77.64	83.45	86.91	49.95	62.99	67.09	79.35
Forgetting [7]	4.64	35.79	59.42	68.21	4.10	28.66	53.47	76.61	4.59	24.76	33.59	63.62
Entropy [1]	26.71	48.29	76.90	81.05	24.80	54.15	73.00	84.91	21.58	47.56	65.92	72.56
EL2N [4]	15.33	56.64	67.24	78.71	20.56	52.34	70.56	79.98	14.60	46.92	61.08	68.26
AUM [5]	4.10	22.36	37.16	53.81	4.00	21.44	37.55	58.01	4.30	13.38	32.08	42.82
CCS [8]	43.90	71.88	78.27	82.52	45.65	77.98	81.40	84.91	52.15	69.48	71.44	77.10
EVA (Ours)	**52.83**	**77.00**	**82.96**	**85.84**	**51.17**	**79.79**	**84.38**	**88.43**	**59.67**	**70.90**	**77.00**	**80.32**

E GENERALIZATION ACROSS ARCHITECTURE

In this section, we investigate the generalization ability across architectures. Specifically, We train a ResNet-18 model with the entire dataset and use various scores to select coresets with different selection rates. Then we train three representative architectures including ResNet-50 [2], MobileNet-v2 [6] and LeNet [3] models with these coresets. The evaluation results in Tab. 4 demonstrate that the coresets selected by the proposed EVA outperform the compared SOTA baselines and have good transferability across architectures.

REFERENCES

[1] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. 2019. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829* (2019).

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[3] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[4] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems* 34 (2021), 20596–20607.

[5] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems* 33 (2020), 17044–17056.

[6] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[7] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159* (2018).

[8] Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. 2022. Coverage-centric coreset selection for high pruning rates. *arXiv preprint arXiv:2210.15809* (2022).