RANDOM SAMPLING FOR ENERGY-EFFICIENT TRAINING OF MACHINE LEARNING MODELS

Presented by:

Oumayma Haddaji

Directed by :

Mr.Balakrishna Prabhu

Mr.Olivier Brun

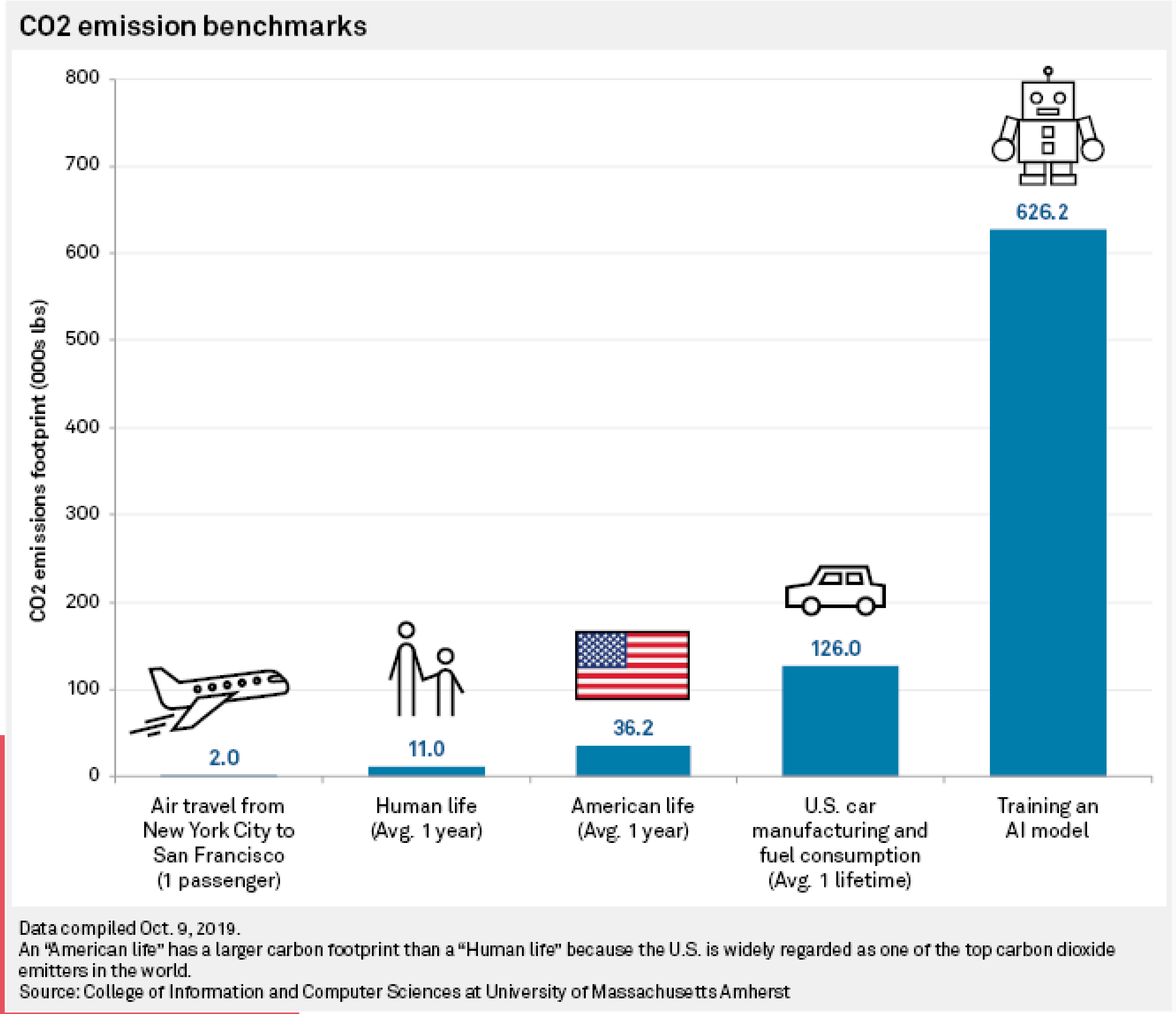
Financed by:

ANR DELIGHT

Start Date I 2 January 2024

PLAN

- INTRODUCTION
- LITERATURE REVIEW
- EXPERIMENTAL SETUP
- RESULTS & ANALYSIS
- CONCLUSION



How to achieve significant energy reductions in training ML models without compromising their accuracy?

Data Summarization Techniques

Sample Similarity : Similarity-Based Data Subset Selection

Gradient Similarity (CRAIG): Gradient-based subset selection.

Stratified Random Sampling: Class-balanced sampling.

Sample Similarity

- **Objective**: Determine the optimal subset for effective training by solving the **Facility Location Problem**.
- **Facility Location Problem**:
  - Select a subset of data points that best represent the entire dataset.
  - Maximize a utility function $f(S)$ that quantifies how well the subset S covers the full dataset.
- **Formal Definition**:

$$f(S) = \sum_{x_i \in \mathcal{X}} \max_{x_j \in S} s(x_i, x_j)$$

where $s(x_i, x_j)$ measures similarity between data points.

Gradient Similarity (CRAIG)

Goal: Select a weighted subset S minimizing the difference between full and subset gradients:

$$\min_{S \subseteq V, \mathbf{w}} \|\nabla \mathcal{L} - \nabla \mathcal{L}_S\|^2$$

$$\nabla \mathcal{L}_S = \sum_{i \in S} w_i \nabla \mathcal{L}_i$$

- V: Full dataset, S: Selected subset, $\mathbf{w}$: Weights
- $\nabla \mathcal{L}$: Full gradient, $\nabla \mathcal{L}_S$: Subset gradient

Facility Location Problem: Maximize submodular function $F(S)$:

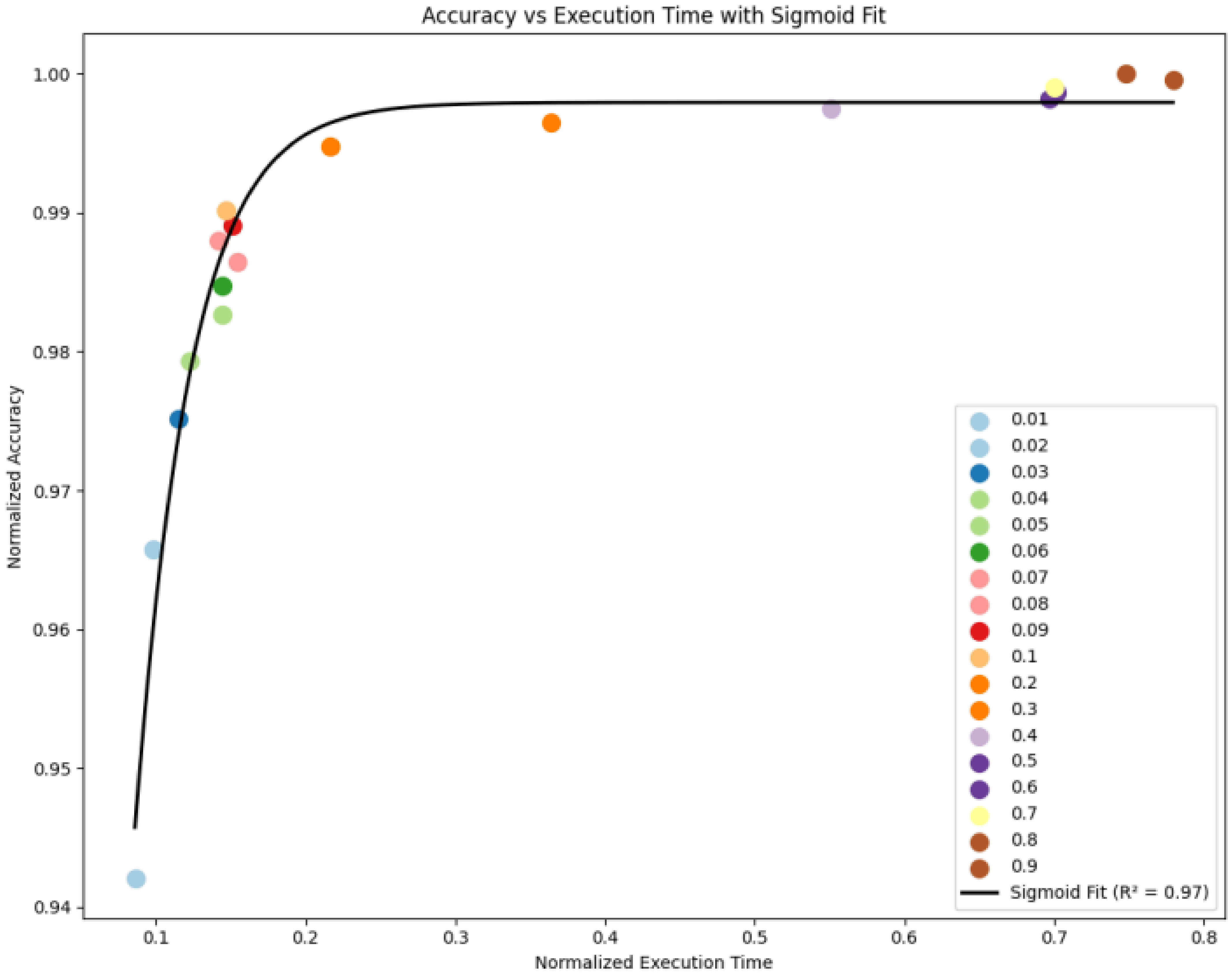
$$F(S) = \sum_{j \in V} \max_{i \in S} \text{sim}(\nabla \mathcal{L}_i, \nabla \mathcal{L}_j)$$

where $\nabla \mathcal{L}_i$,$\nabla \mathcal{L}_j$ are the gradients of data points i and j.

Experimental Setup

- **Centralized Setup**
- **Datasets:** MNIST, Fashion MNIST, CIFAR-10.
- **Energy Measurement:**
  - Grid'5000 platform for energy monitoring provided by the Delight project.
  - Mojitos, an open-source tool for monitoring system energy and network usage at the operating system level on GNU/Linux.

MNIST DATASET



- Normalized Mean of execution time & accuracy
- 100 repetitions
- Maximum of 100 rounds
- Early Stopping: accuracy stabilized over five consecutive rounds.

→ Small subsets (1–5%): Fast but less accurate.

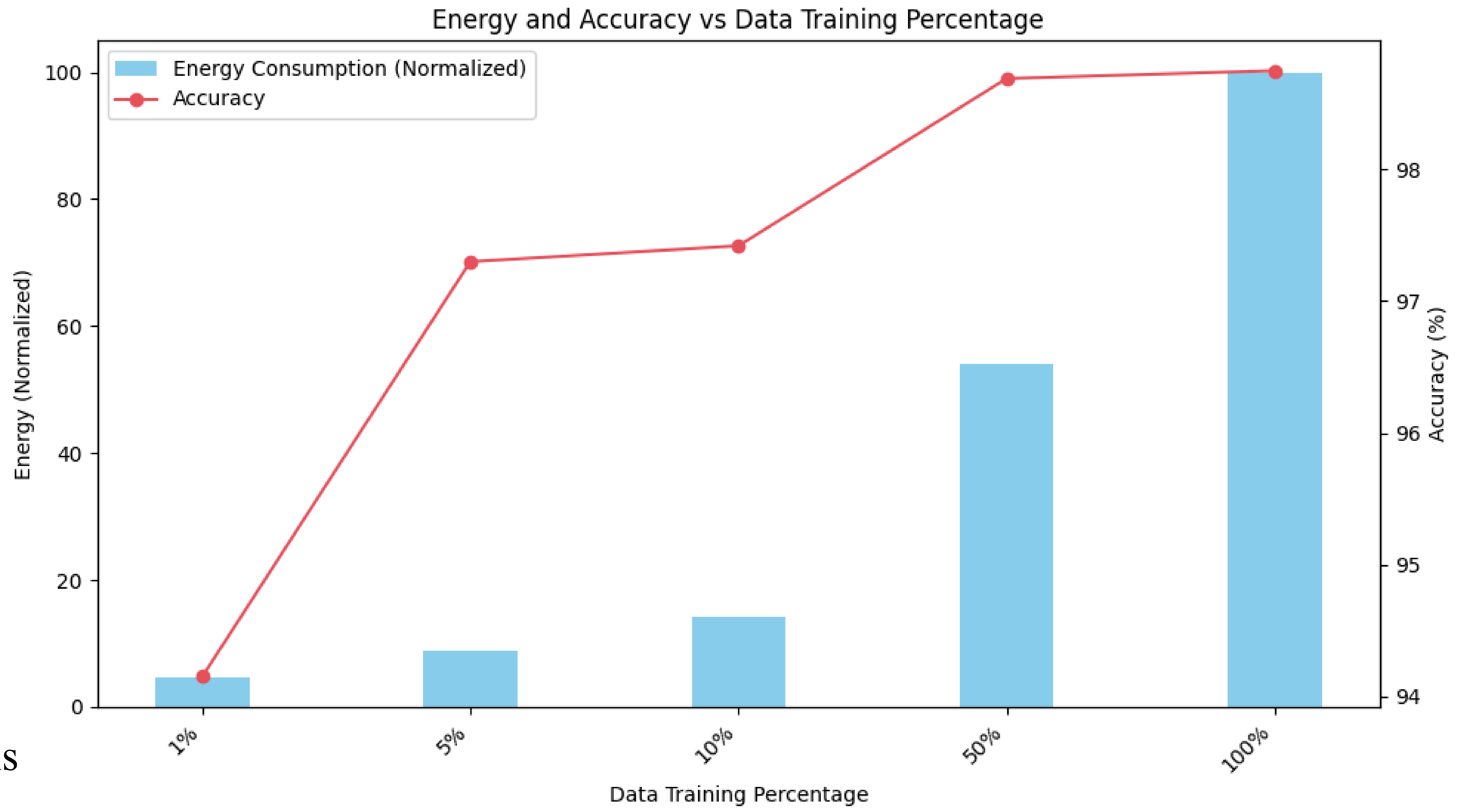
→ Large subsets (80–90%) : Accurate but slow.

→ **Optimal range: 10–20%**

Energy & Accuracy Analysis For MNIST

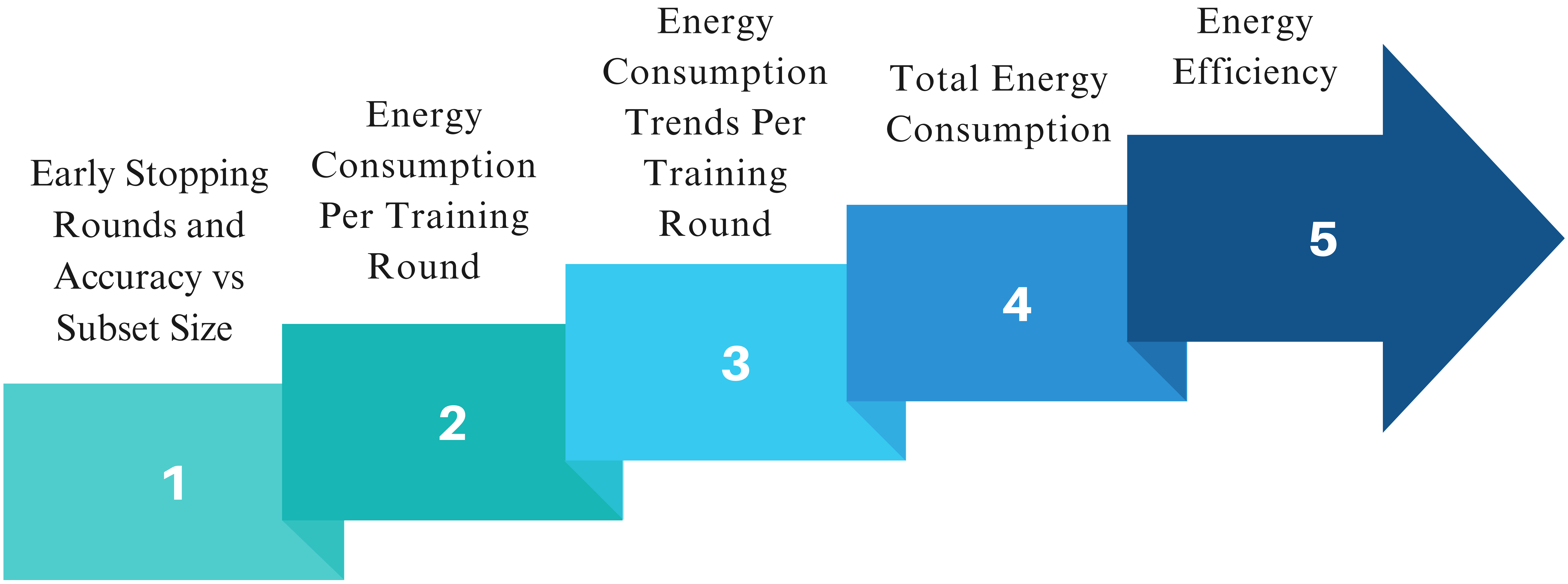
- Energy consumption increases exponentially with dataset size.
- Accuracy gains diminish after 10%.
- Optimal range: 1–10% for energy efficiency.

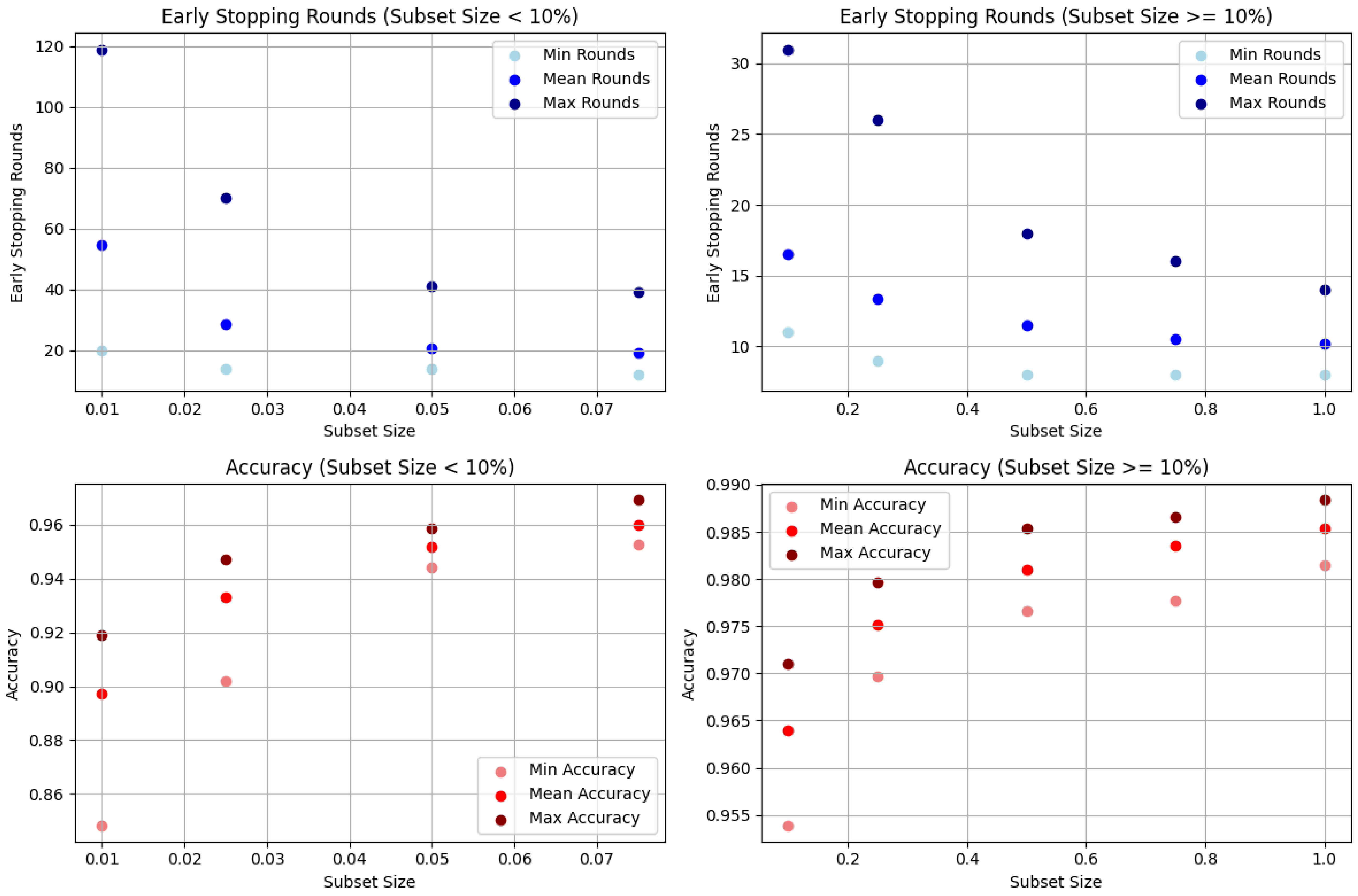
Issue: Resource limitations on Grid’5000.



ROUND-BASED TRAINING ANALYSIS

5-Step Process





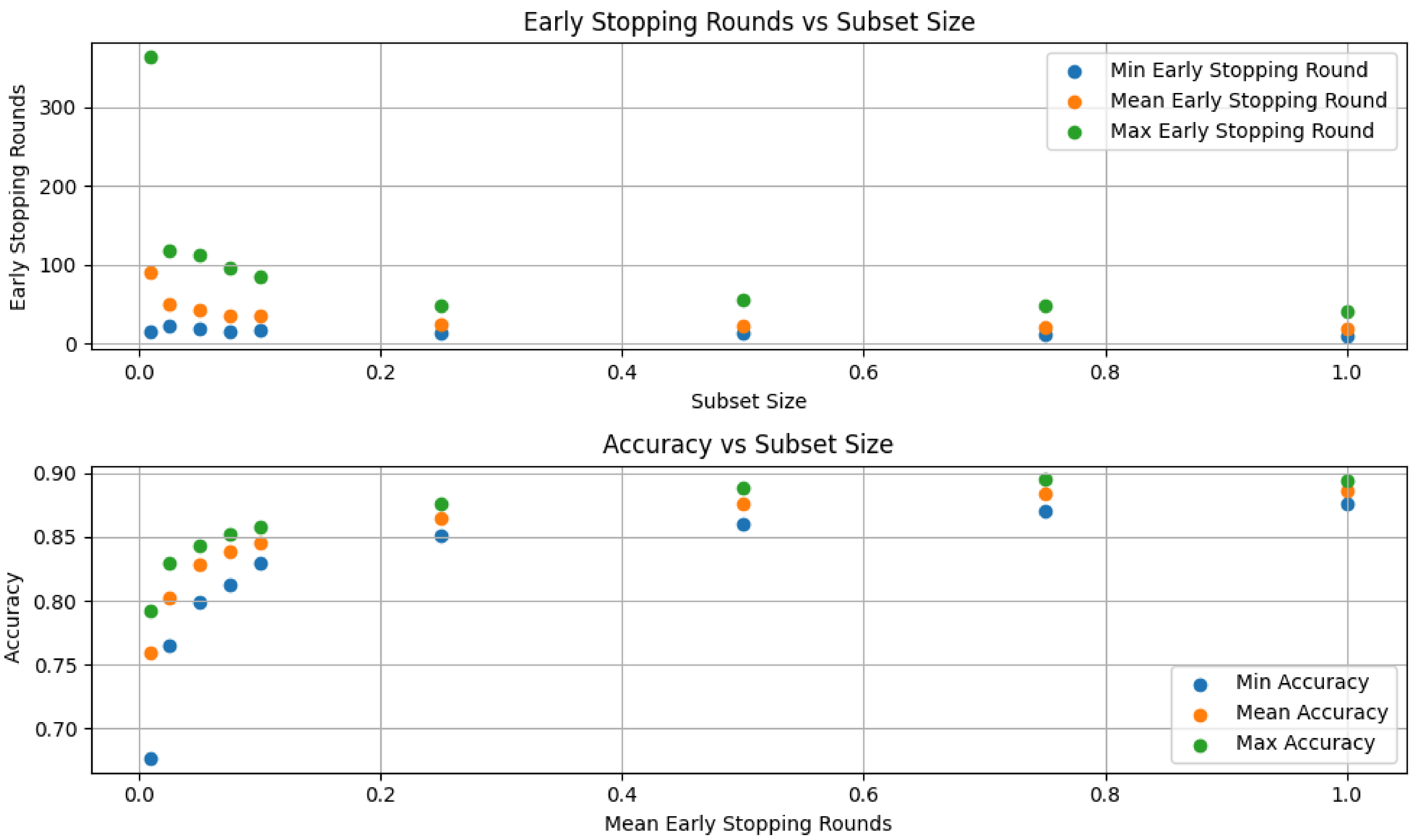
MNIST DATA: Early Stopping Rounds vs Subset Size (Top) and Accuracy vs Subset Size (Bottom)

- 100 repetitions

Training Setup:

- Early stopping criteria: Patience = 5, Min Delta = 0.005.
- Maximum rounds = 500.

- **Training rounds decrease significantly as subset size increases.**



FashionMNIST DATA: Early Stopping Rounds vs Subset Size (Top) and Accuracy vs Subset Size (Bottom)

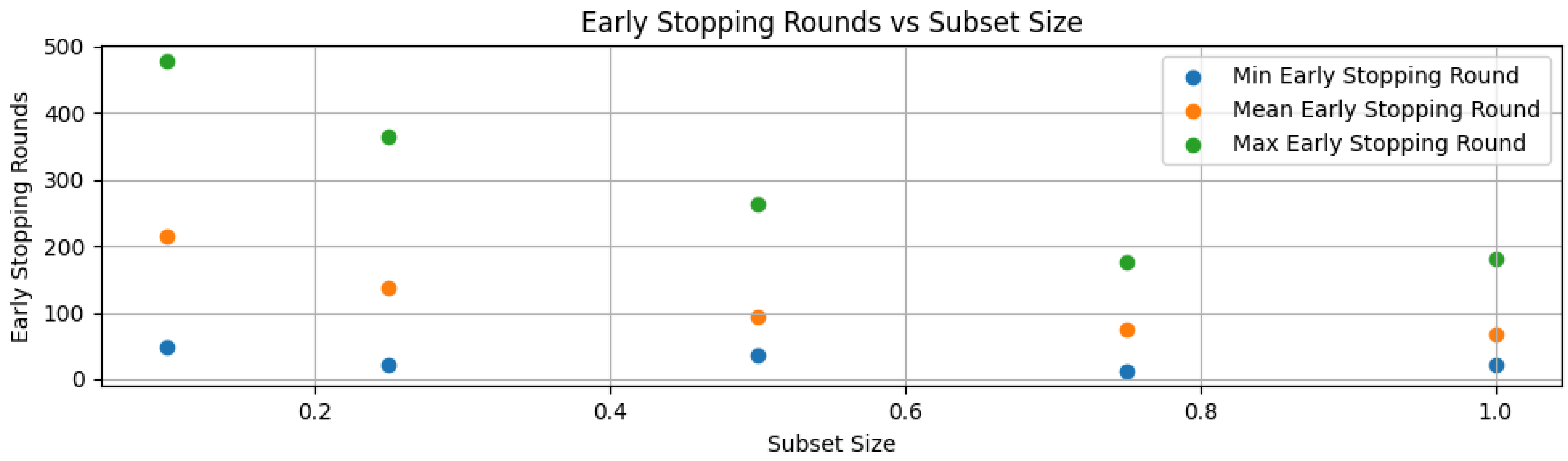
- 100 repetitions

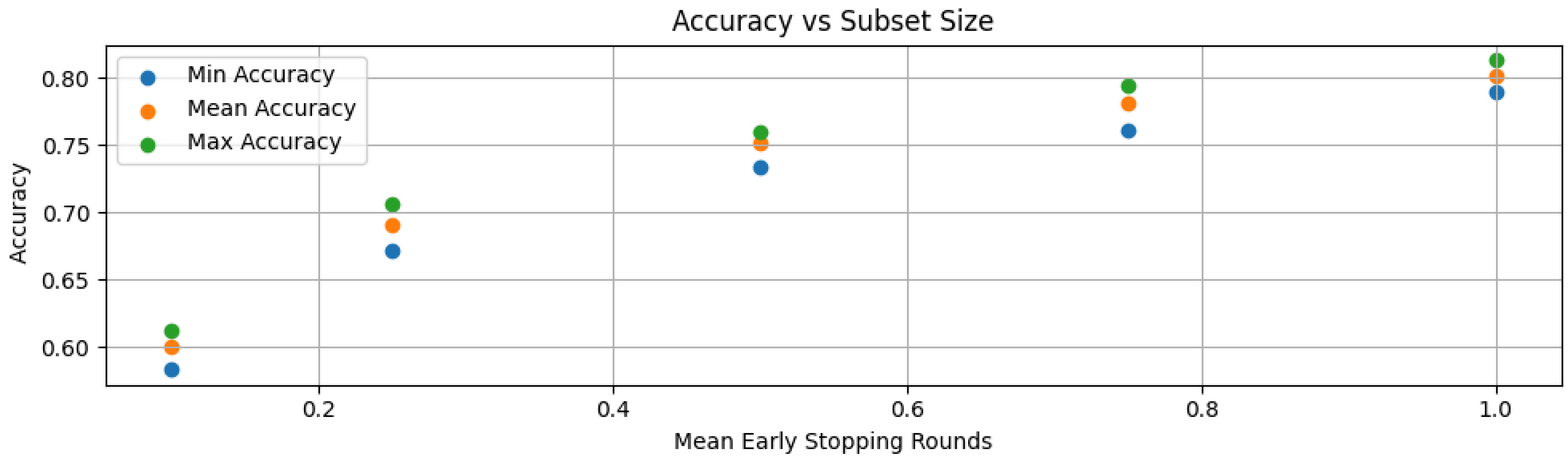
Training Setup:

- Early stopping criteria: Patience = 5, Min Delta = 0.005.
- Maximum rounds = 500.

- **Training rounds decrease significantly as subset size increases.**

CIFAR10 DATA: Early Stopping Rounds vs Subset Size (Top) and Accuracy vs Subset Size (Bottom)



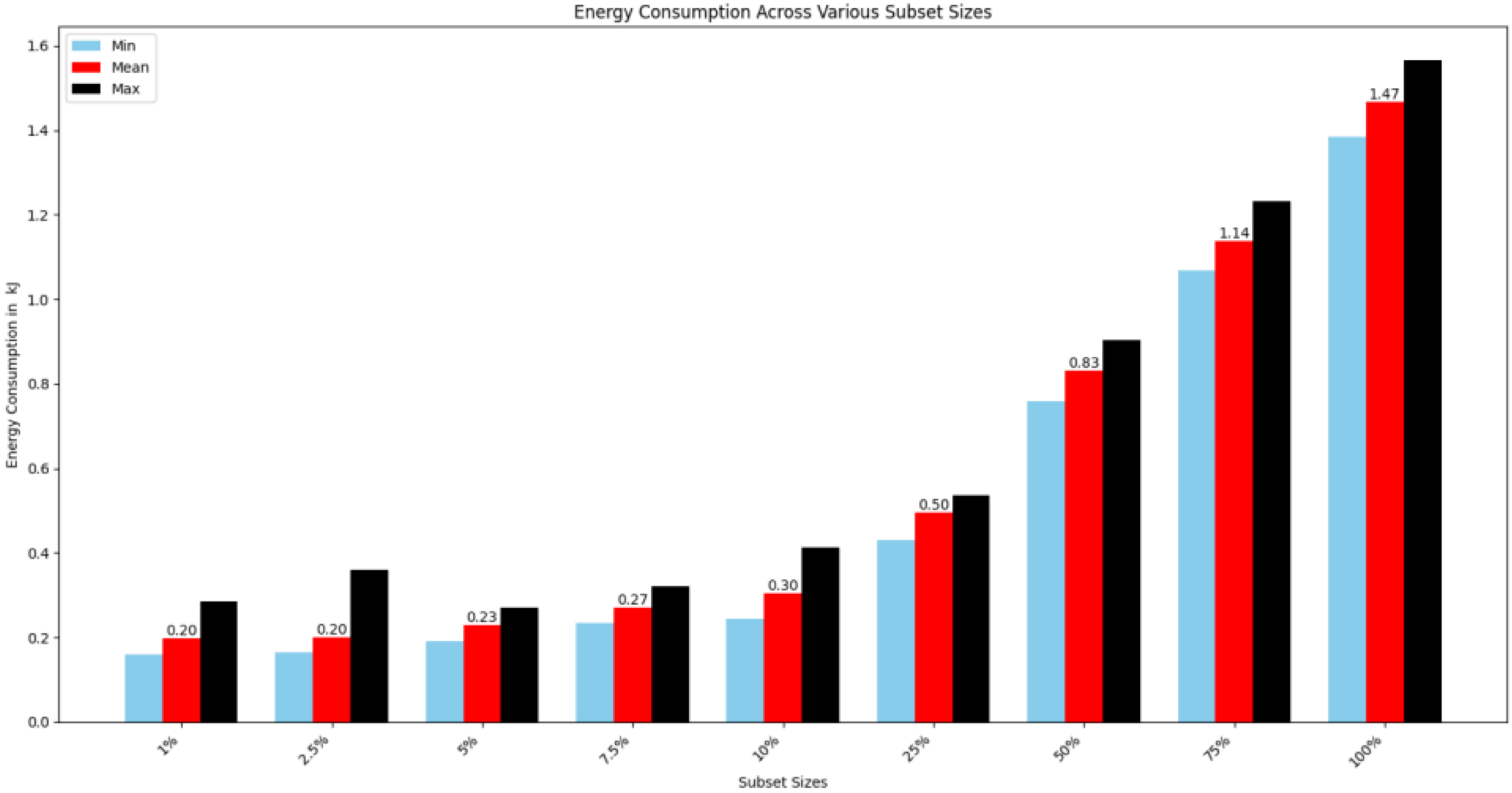


- 100 repetitions

Training Setup:

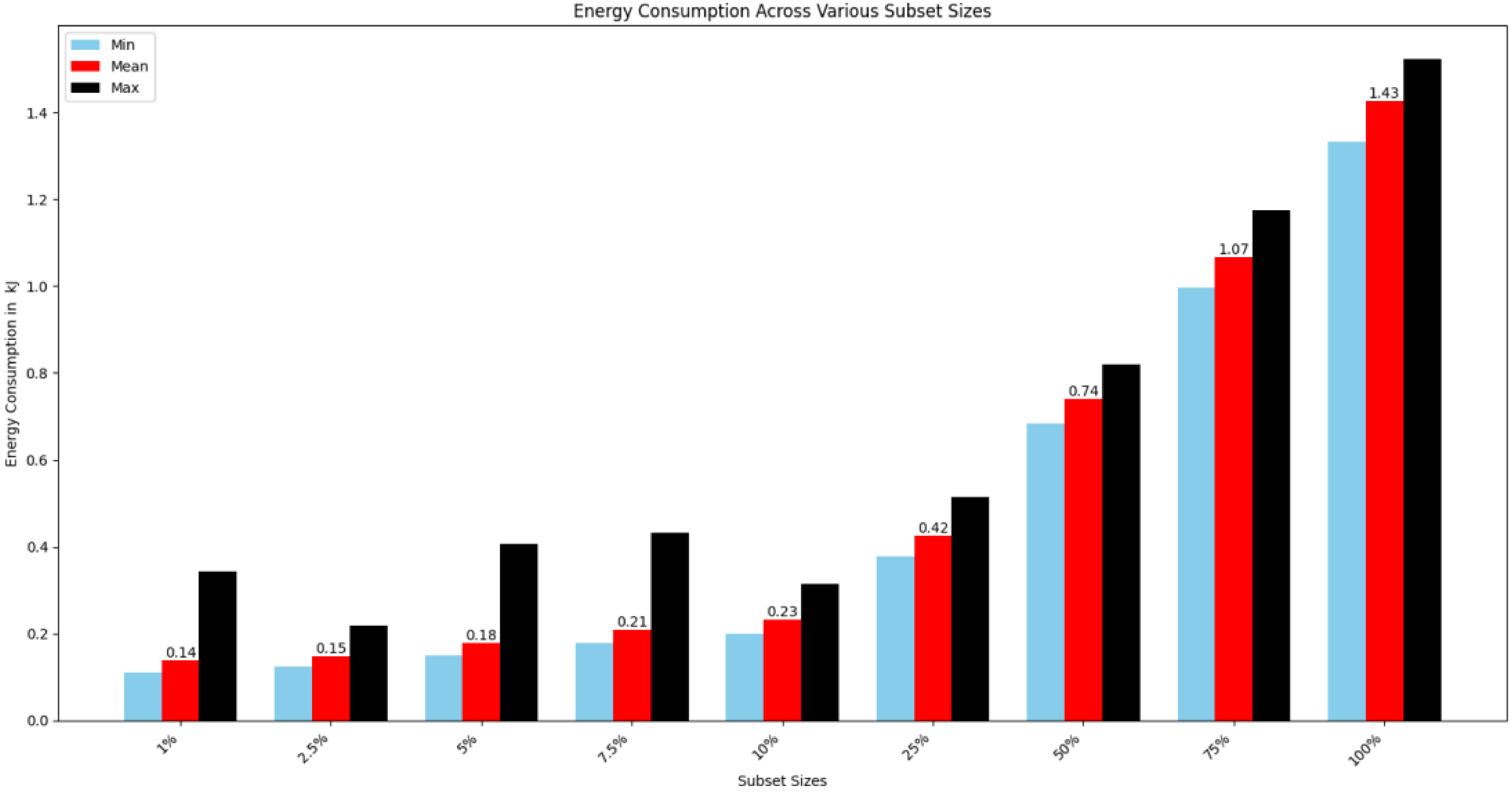
- Early stopping criteria: Patience = 5, Min Delta = 0.005.
- Maximum rounds = 500.

- **Training rounds decrease significantly as subset size increases.**



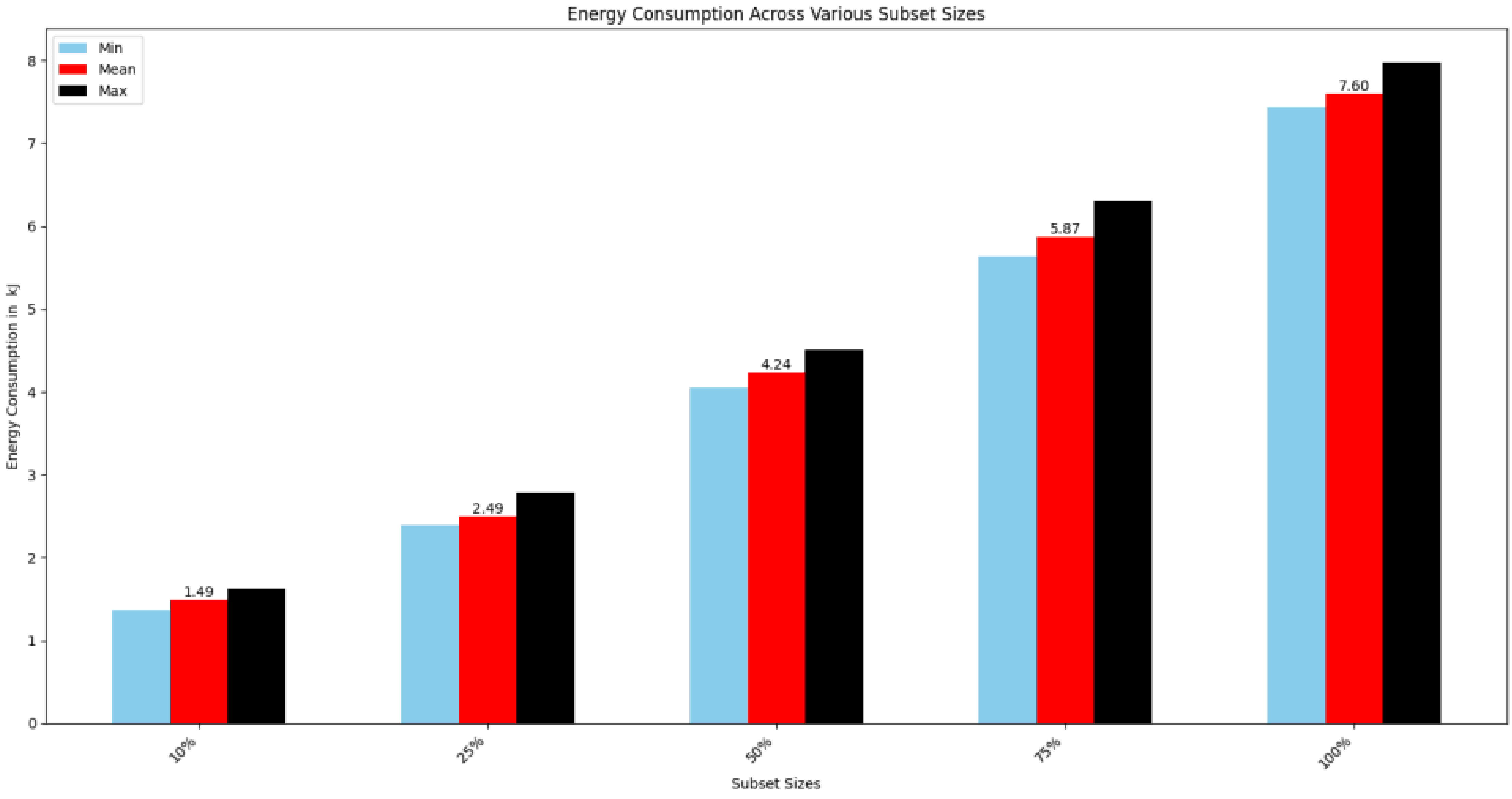
Energy Consumption of a Training Round - MNIST

- 100 repetitions
- Rennes cluster designated as paradoxe-[1,3-4,12,21].
- Energy increases with subset size.



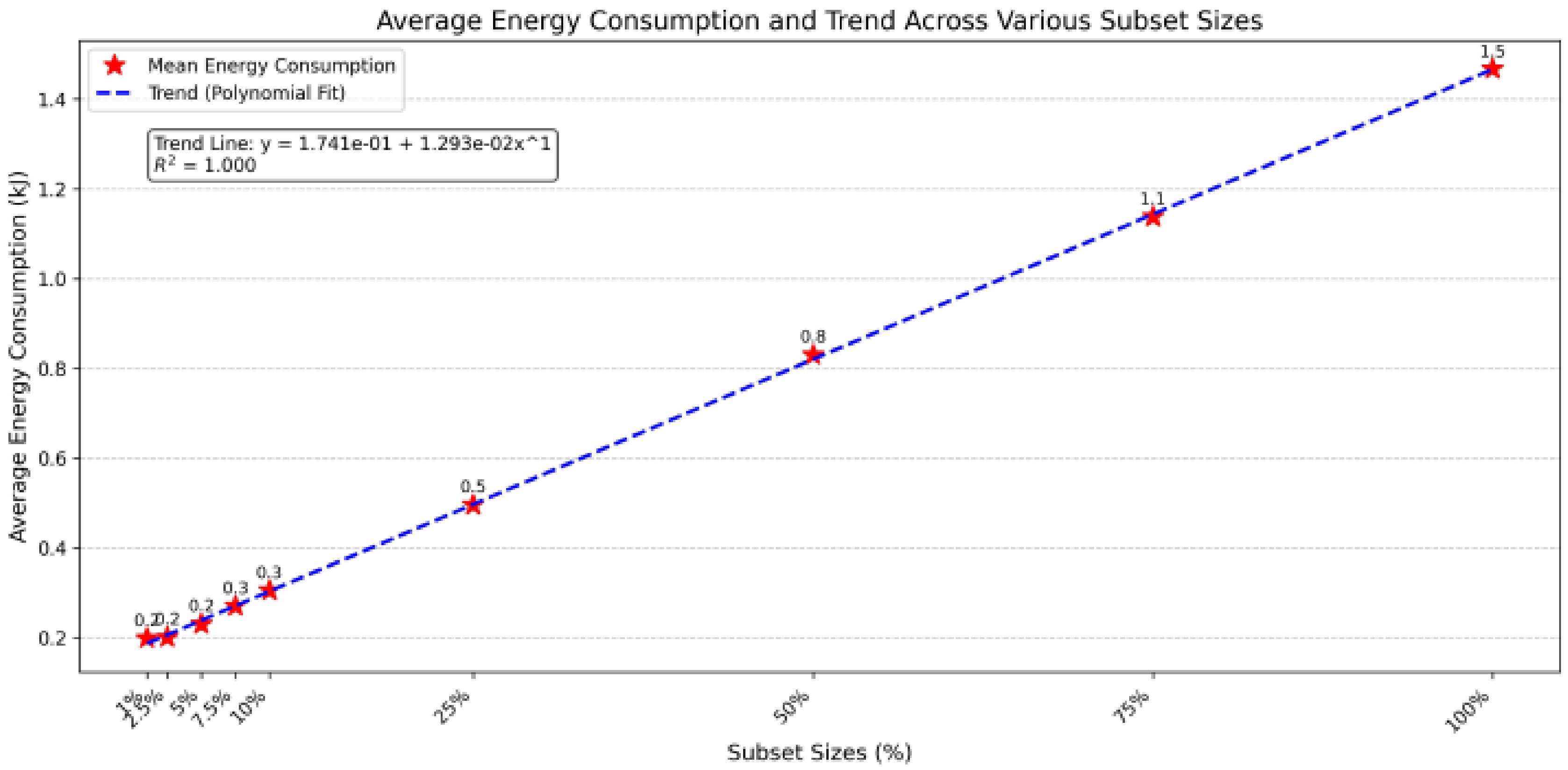
Energy Consumption of a Training Round - Fashion MNIST

- 100 repetitions

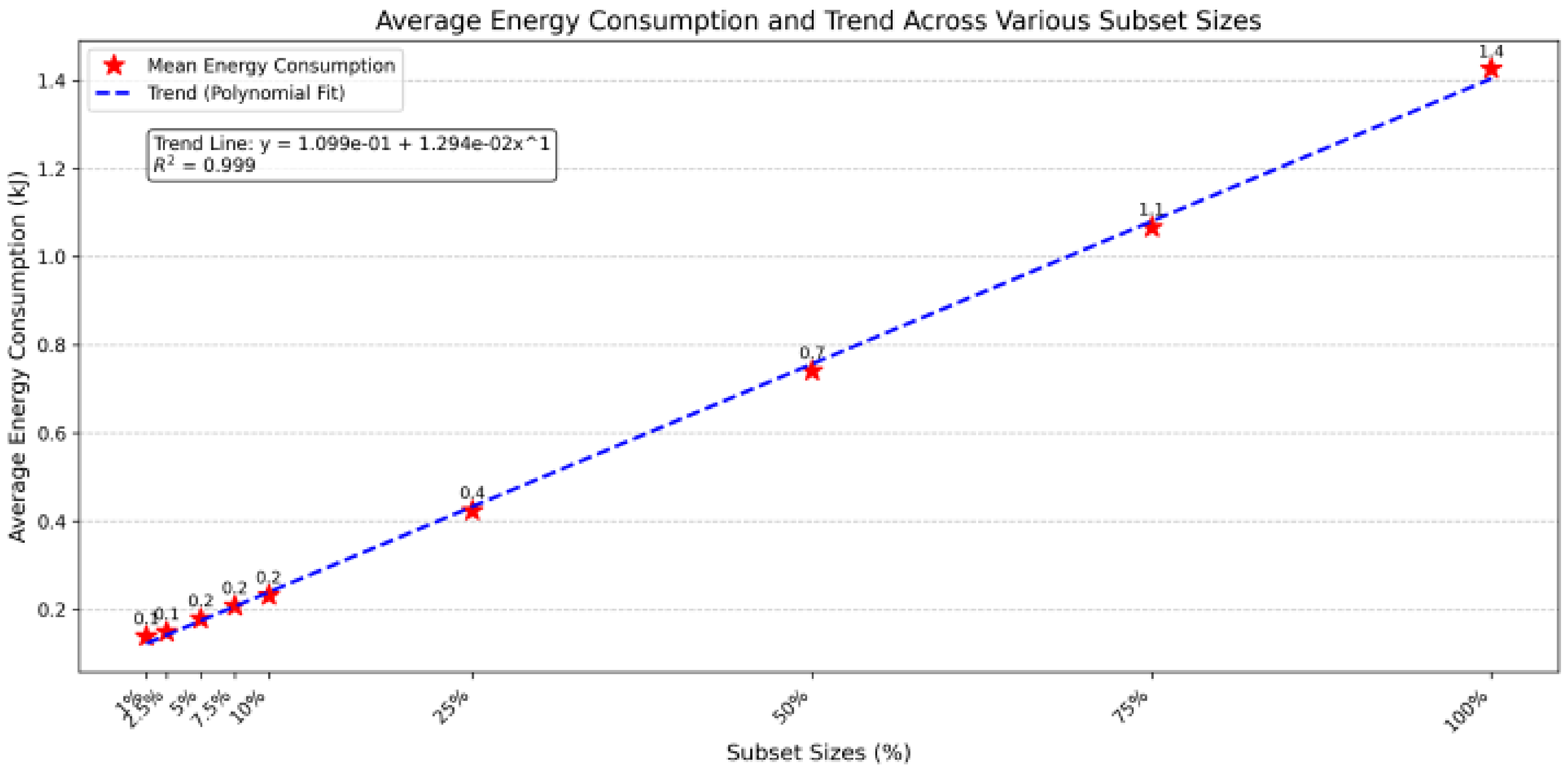


Energy Consumption of a Training Round - CIFAR10

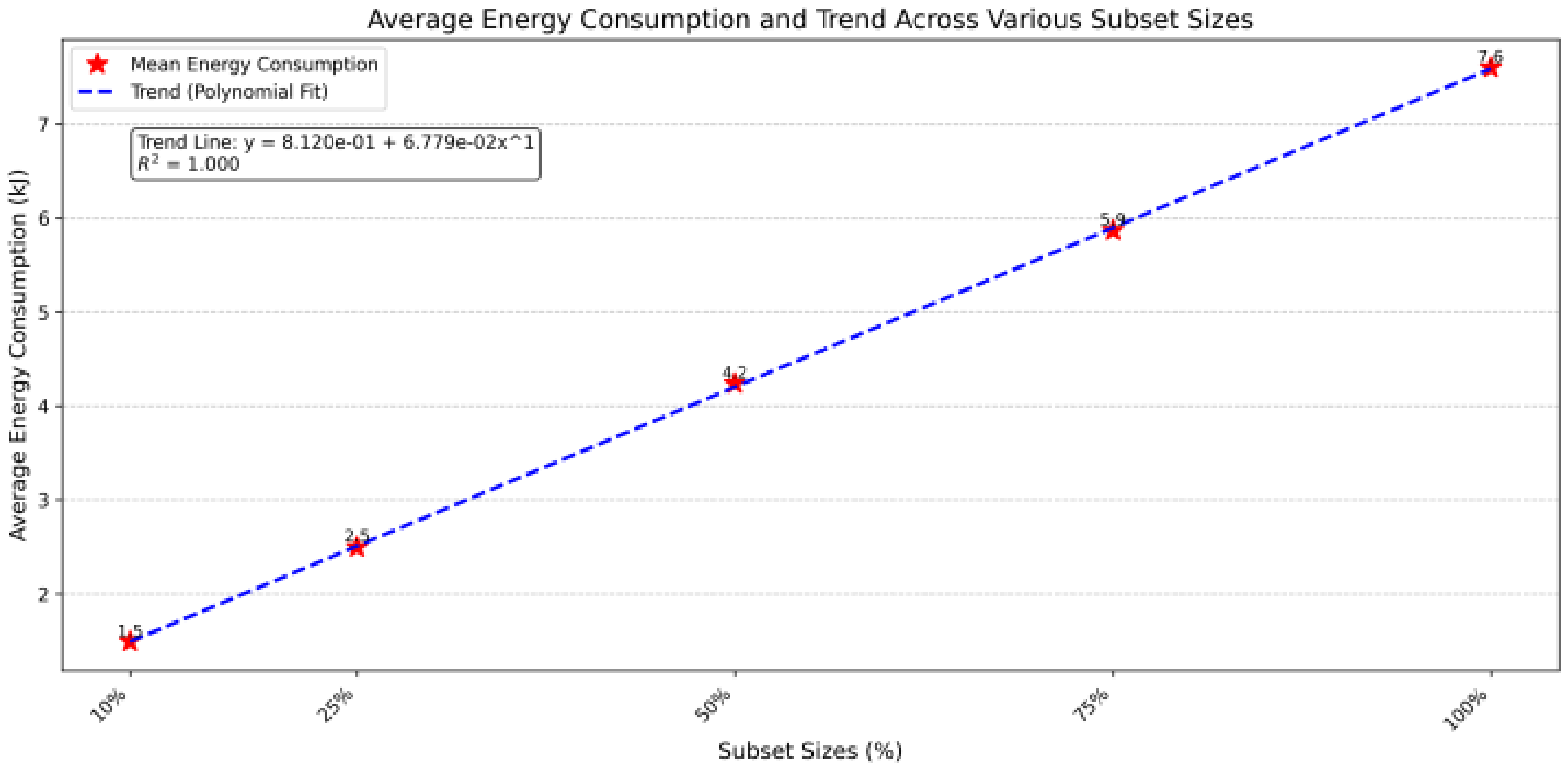
- 100 repetitions
- CIFAR10 requires significantly more energy due to its complexity.



(A) MNIST



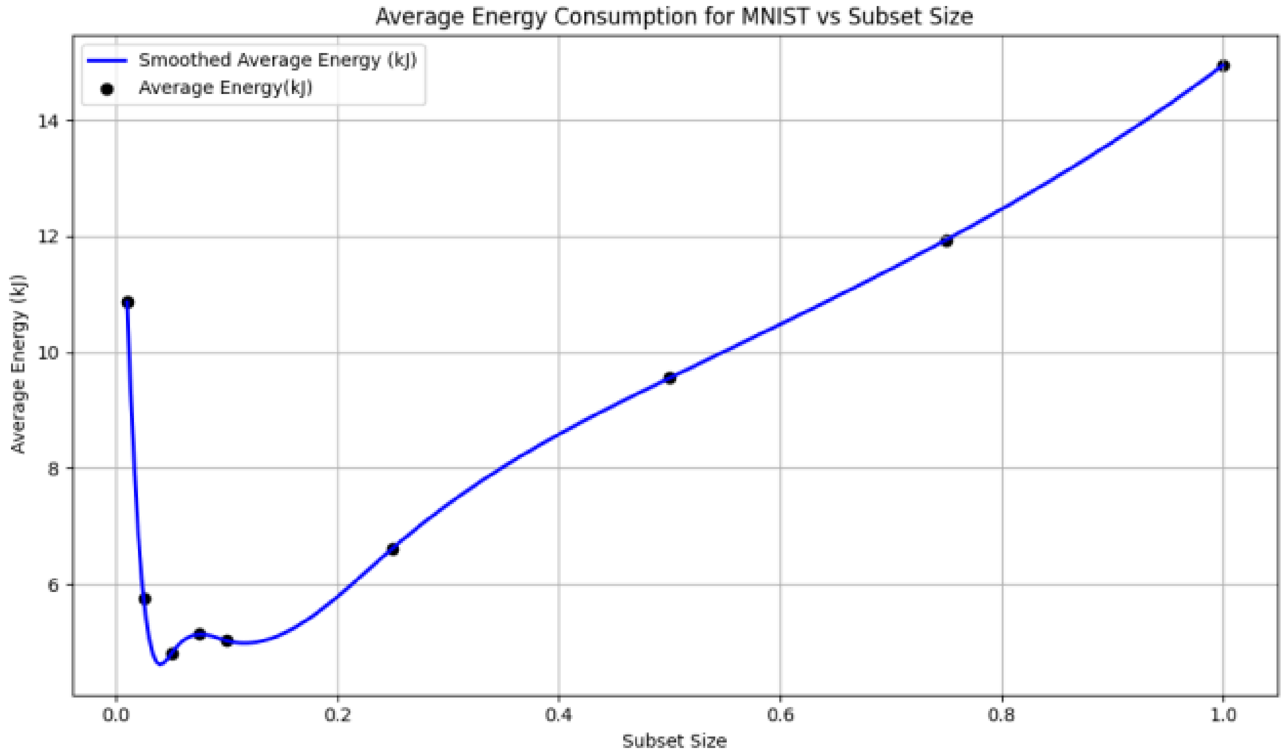
(B) Fashion MNIST



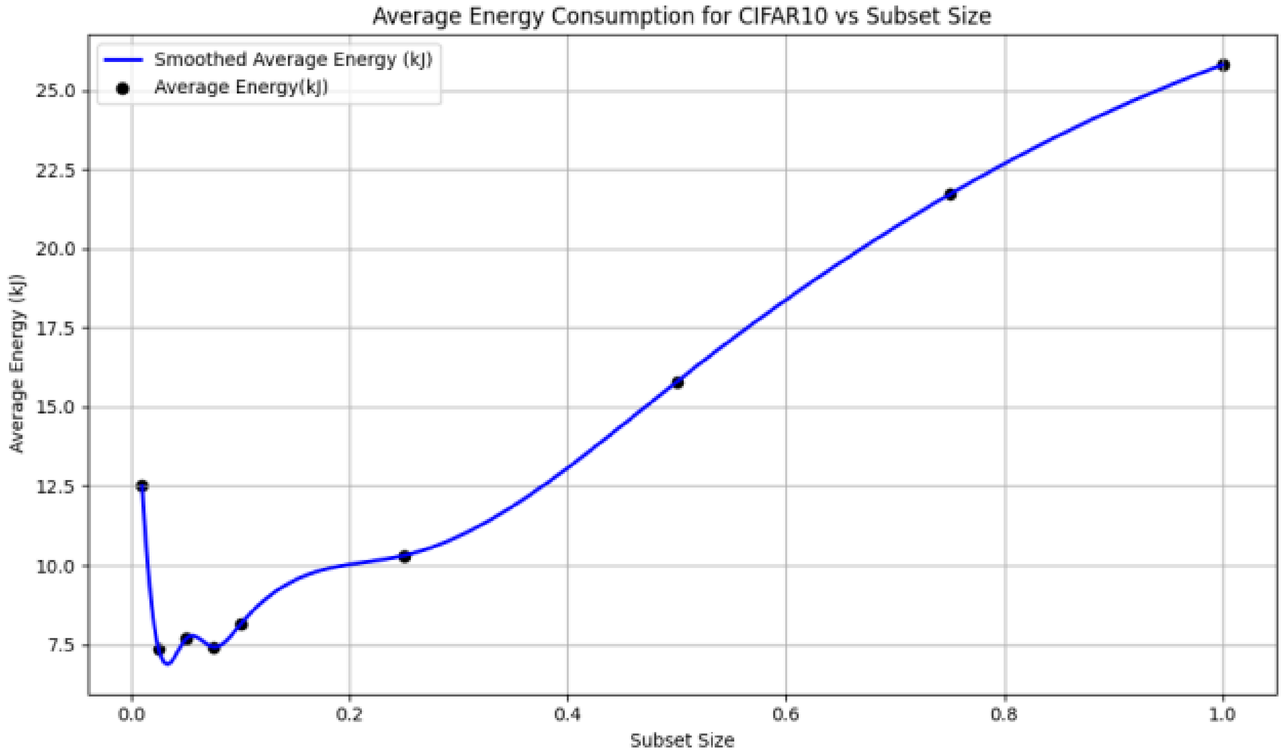
(C) CIFAR-10

- MNIST and Fashion MNIST have similar energy trends.

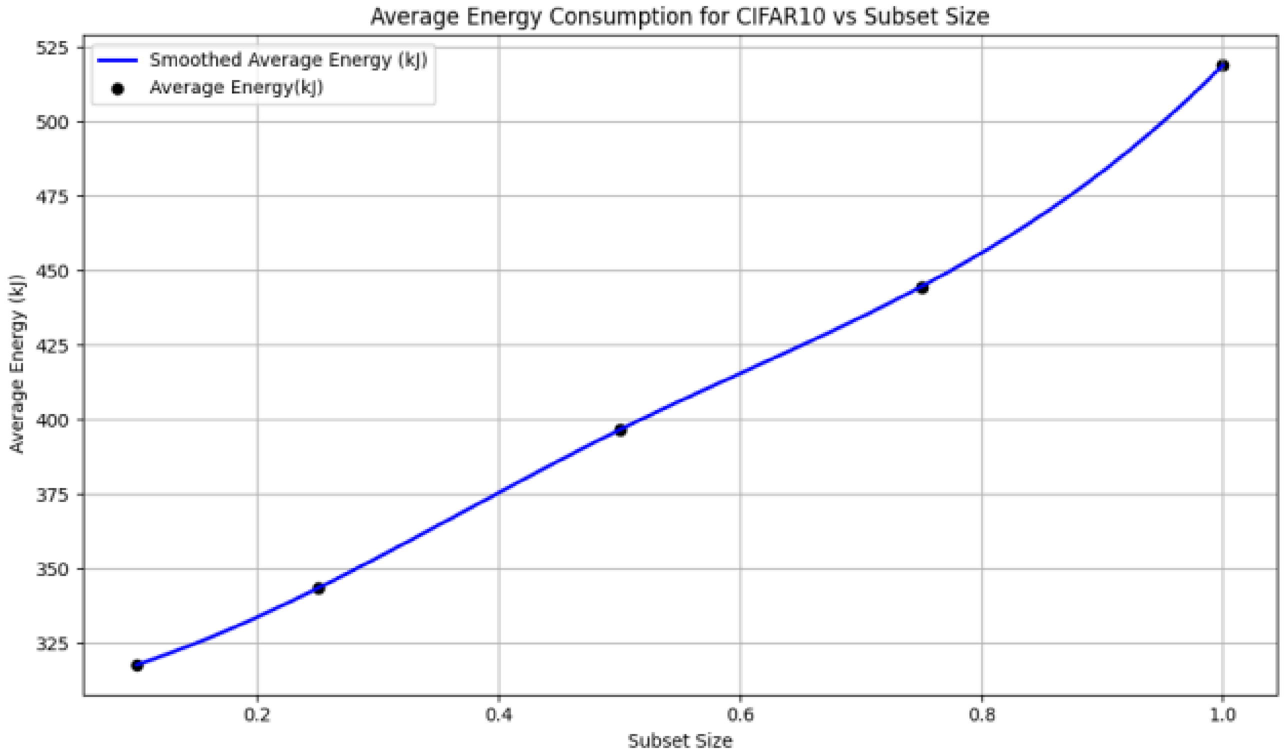
Total Energy Consumption = Average number of training rounds× Average energy per round



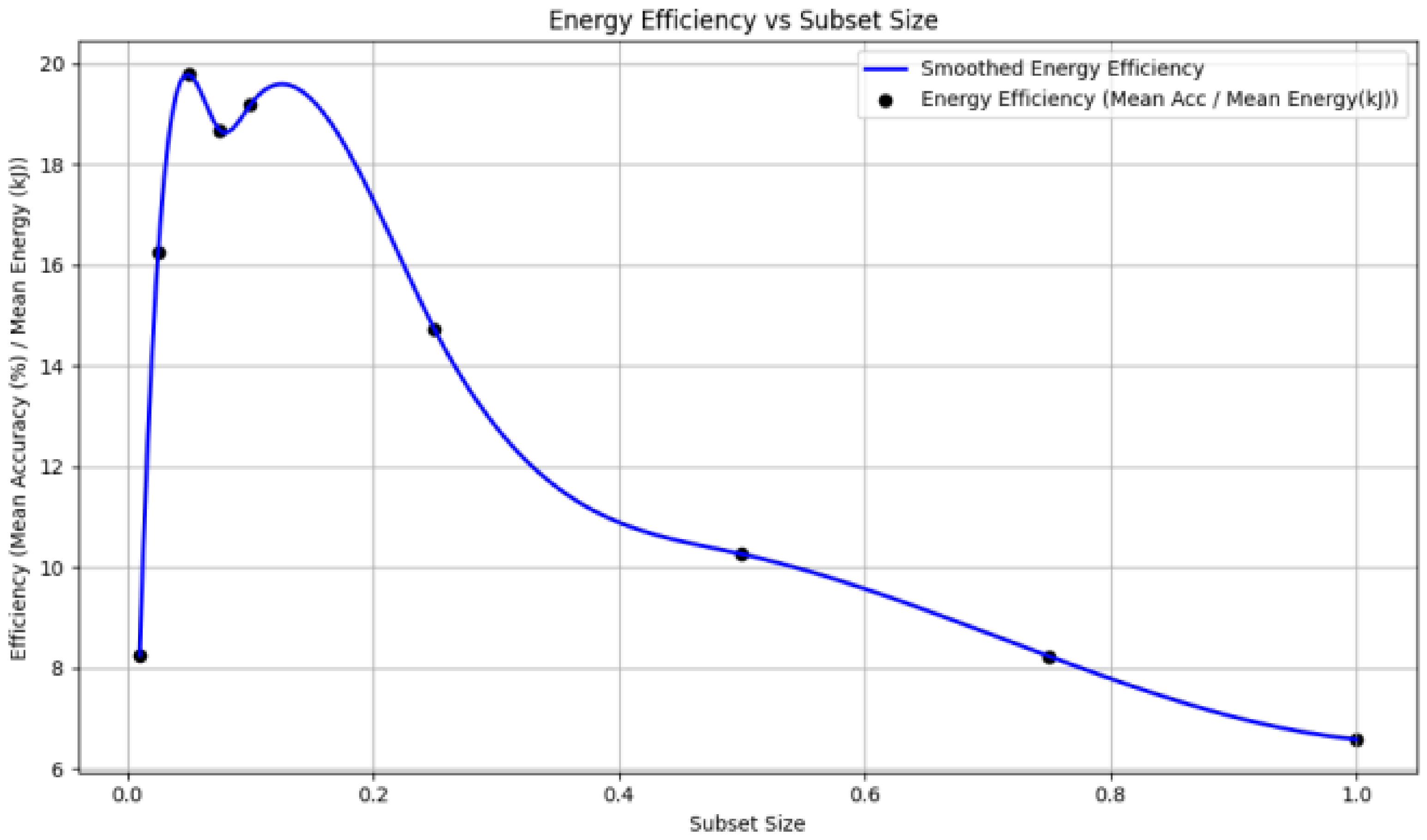
(A) Energy Consumption for Training - MNIST



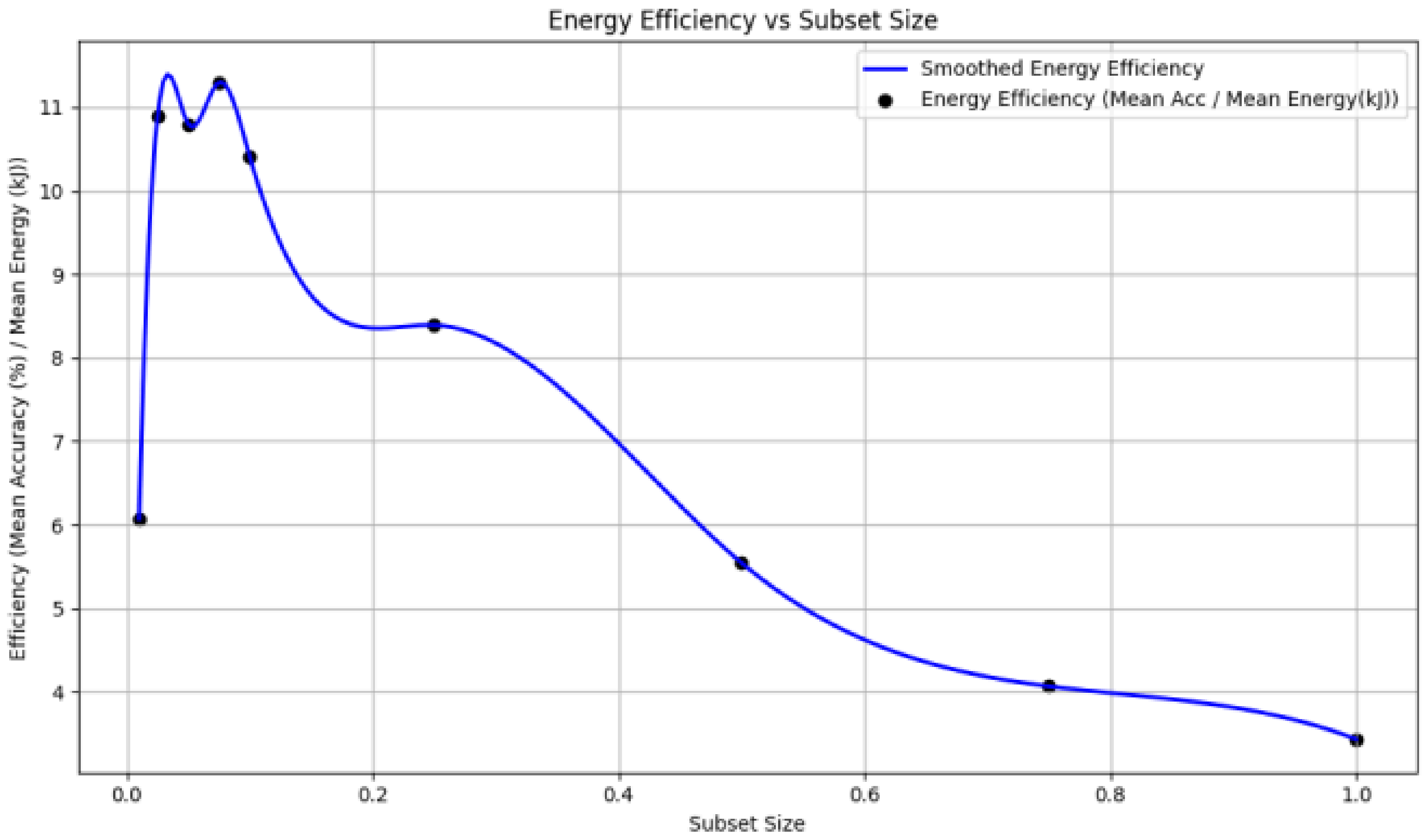
(B) Energy Consumption for Training - Fashion MNIST



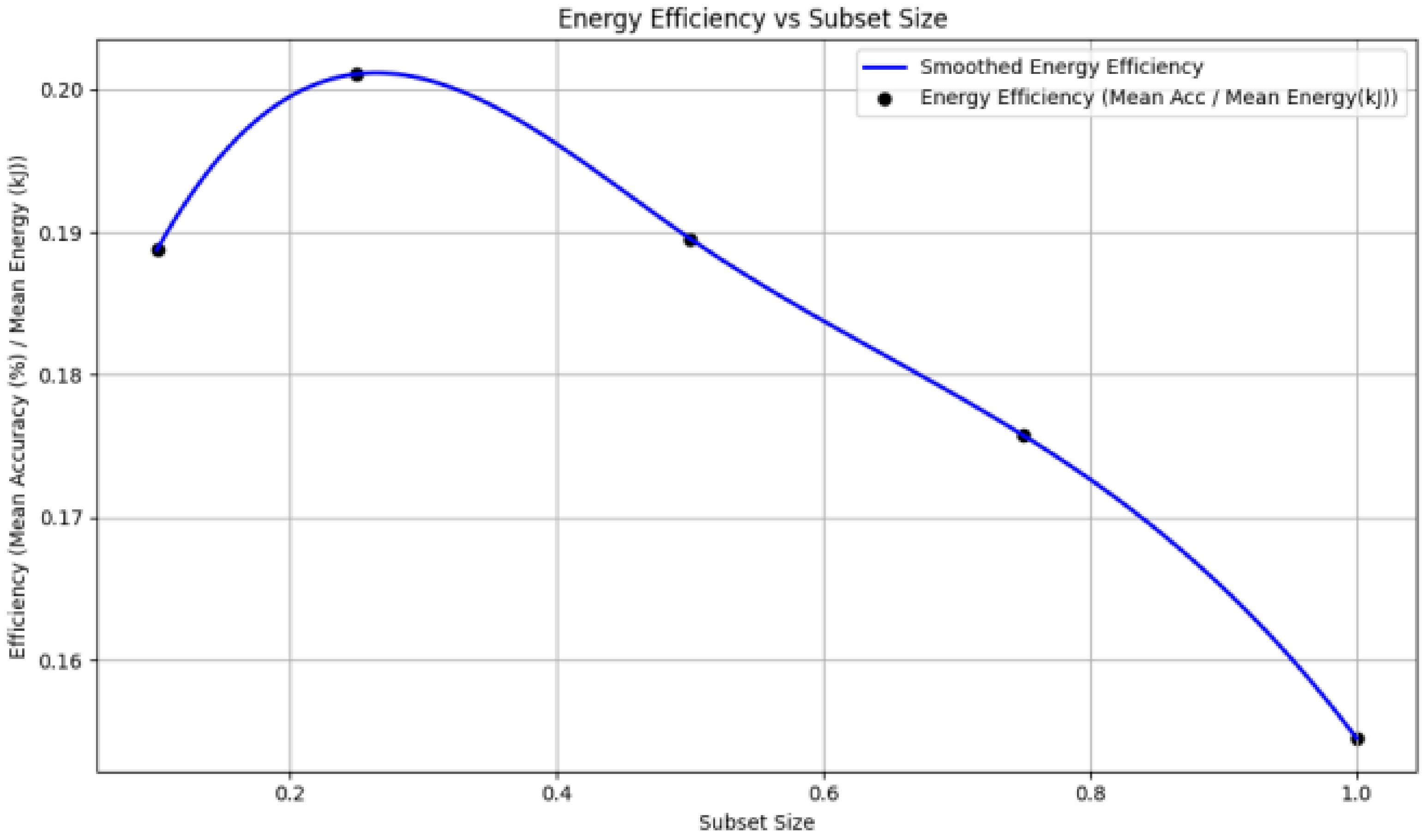
(C) Energy Consumption for Training - CIFAR10



(A) Energy Efficiency - MNIST



(B) Energy Efficiency - Fashion MNIST



(C) Energy Efficiency - CIFAR10

$$\text{Energy Efficiency} = \frac{\text{Mean Accuracy}}{\text{Mean Energy Consumption}}$$

- CIFAR-10: Peak efficiency at 25%.
- MNIST: Peaks at 5% and 10%.
- Fashion MNIST: Peaks at 3% and 7.5%.

Sample Similarity Results

Subset Size	Subset Selection Time (s)	Training Time (s)	Train Accuracy	Percentage Time	Final Test Accuracy	Stopping Round
Full Dataset	-	45.94	0.9782		0.9788	5
10	35.38	0.24	1.0000	77.54%	0.8537	12
60	39.87	0.70	0.9883	88.32%	0.9157	7
150	41.41	3.47	0.9973	97.71%	0.9495	13
300	43.30	5.21	0.9930	105.60%	0.9494	10
450	44.48	6.06	0.9876	110.03%	0.9484	8
600	44.88	9.31	0.9748	117.94%	0.9709	9
1500	47.88	14.34	0.9725	135.44%	0.9701	6
3000	50.22	23.02	0.9690	159.43%	0.9779	5

Training and Test performance on different subset sizes: MNIST

Subset Size	Subset Selection Time (s)	Training Time (s)	Train Accuracy	Percentage Time (%)	Final Test Accuracy	Stopping Round
Full Dataset	-	81.41	0.8710	-	0.8597	8
10	35.37	0.83	0.9900	44.47	0.7049	39
60	37.51	4.30	0.9917	51.36	0.7736	37
150	38.66	3.84	0.8880	52.19	0.7933	14
300	40.27	11.10	0.8593	63.09	0.7899	20
450	41.05	9.67	0.8449	62.30	0.7965	12
600	41.67	16.35	0.8503	71.27	0.8137	15
1500	43.57	26.08	0.8130	85.56	0.8342	10
3000	46.15	29.06	0.8236	92.39	0.8327	6

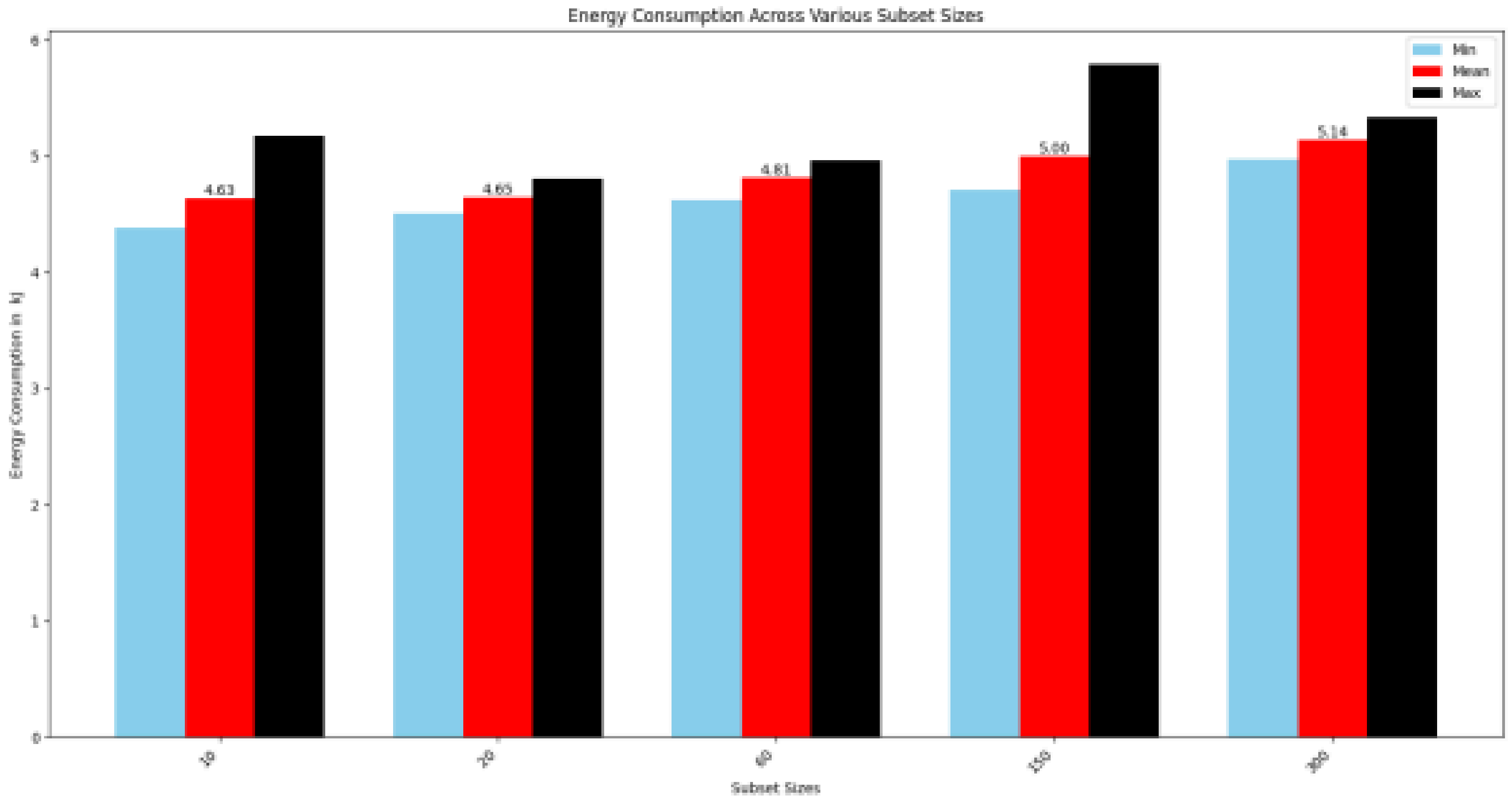
Training and Test performance on different subset sizes: Fashion MNIST

Subset Size	Subset Selection Time (s)	Training Time (s)	Train Accuracy	Percentage Time (%)	Final Test Accuracy	Stopping Round
Full Dataset	-	6002.77	0.9298	-	0.8105	23
10	49.35	20.55	0.6800	1.16	0.2729	41
100	50.20	306.73	0.9370	5.95	0.3778	64
500	51.74	807.56	0.9230	14.32	0.5594	34
750	52.53	945.26	0.9021	16.62	0.6257	27
1250	53.38	1739.58	0.9314	29.87	0.6681	29
2500	57.82	3668.50	0.9404	62.08	0.7581	29

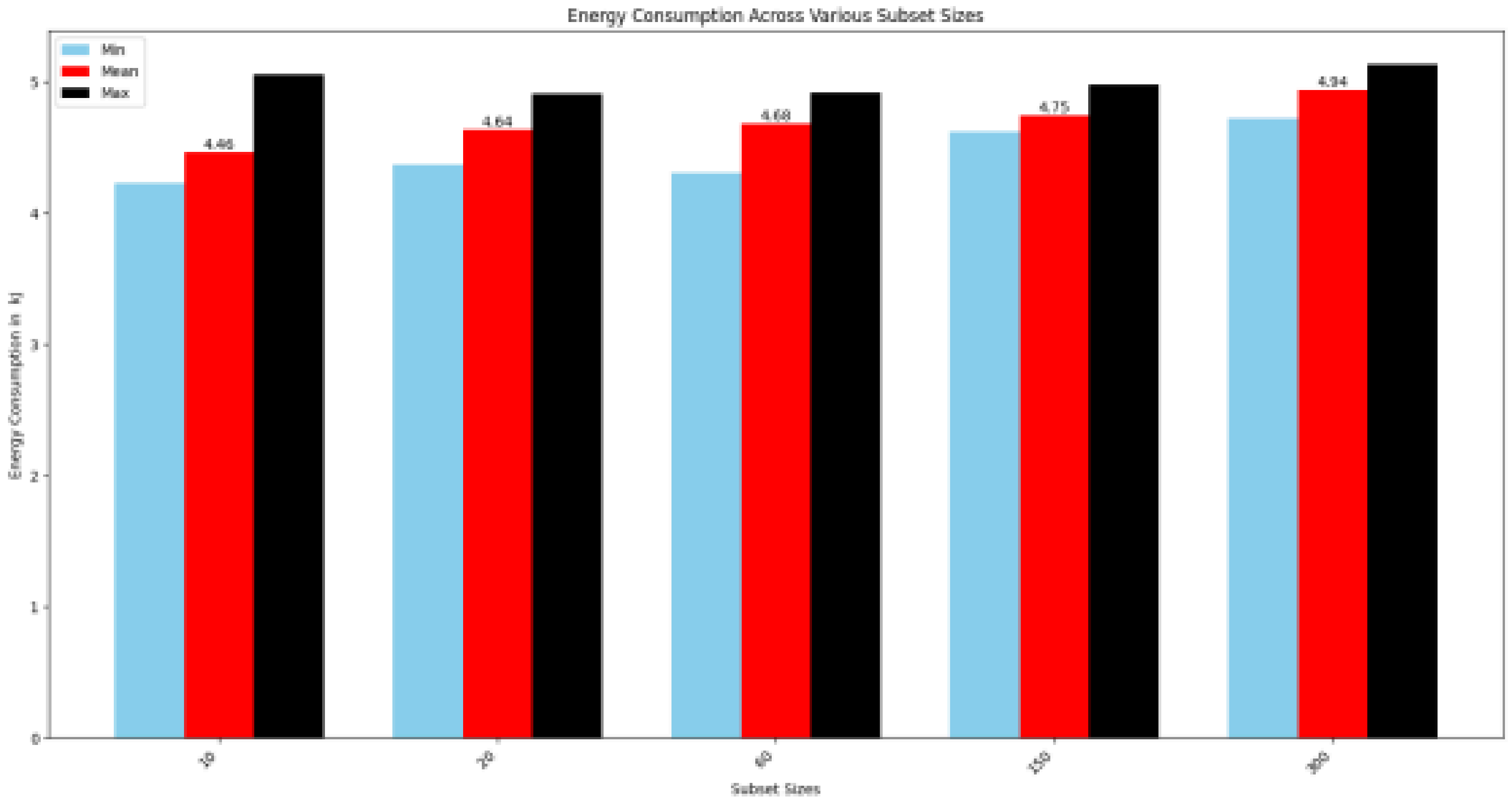
Training and Test performance on different subset sizes: CIFAR10

Early stopping criteria:

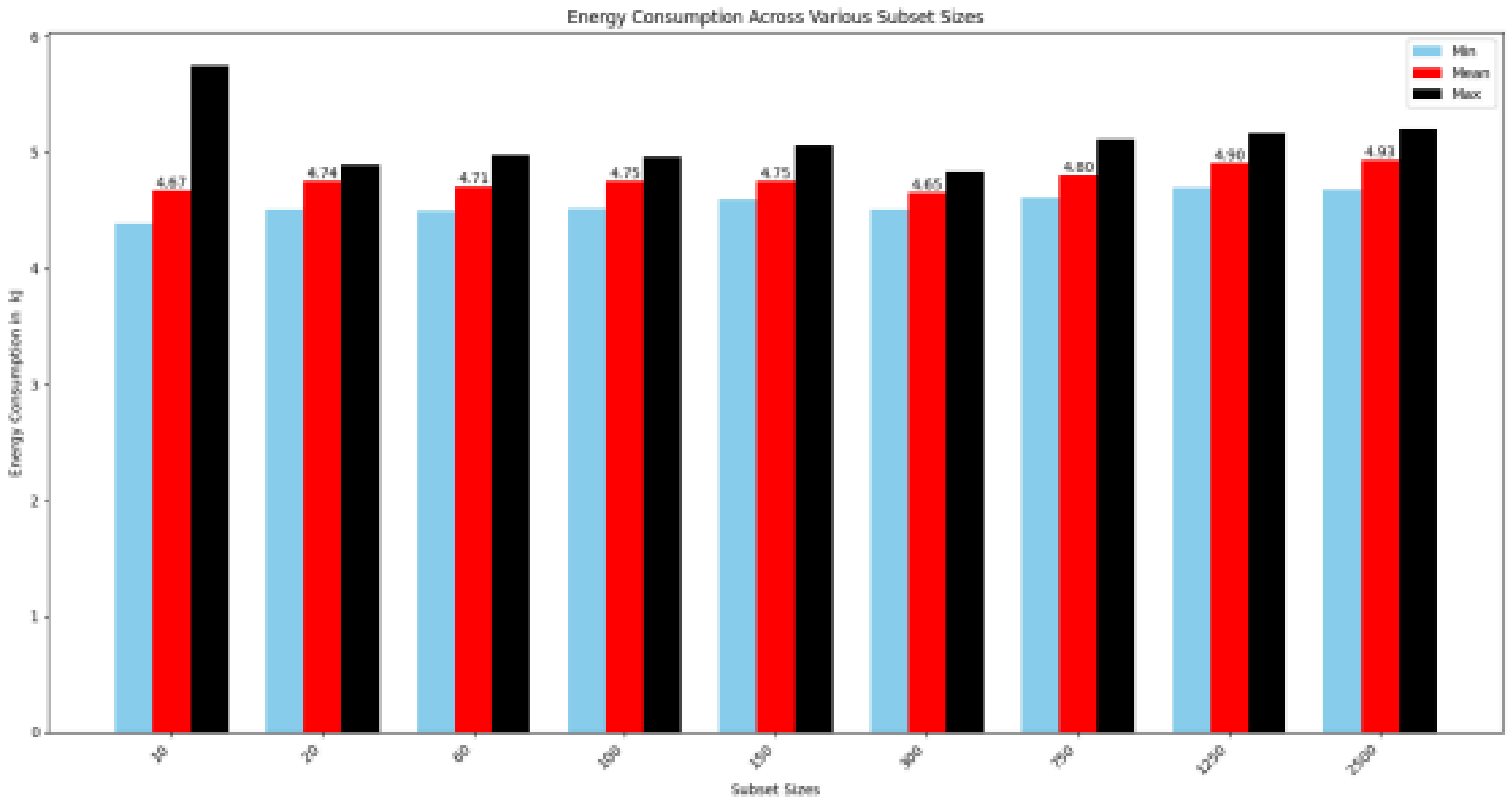
- Patience = 3
- Min Delta = 0.01
- Training accuracy instead of test accuracy.



(A) Energy consumption for MNIST



(B) Energy consumption for Fashion MNIST



(C) Energy consumption for CIFAR10

Subset Size	Test Accuracy	Stopped at Round	Training Time (s)	Data Summary Time (s)
100	0.7642	20	16.930	19.348
200	0.8615	31	26.060	19.725
600	0.8979	21	17.567	20.298
1500	0.9336	14	13.692	21.599
3000	0.9522	13	14.122	21.757
4500	0.9542	10	13.027	22.359
6000	0.9760	13	19.826	23.111
15000	0.9848	10	25.127	23.892
30000	0.9881	9	38.927	41.657

Summary of Results for Different Subset Sizes

Subset Size	Test Accuracy	Stopped at Round	Total Training Time (s)	Total Data Summary Time (s)
100	0.2806	111	90.548	116.237
200	0.2743	92	117.050	117.162
1500	0.4981	45	348.050	127.601
5000	0.5713	29	731.088	130.941
7500	0.6245	31	1163.417	132.763
12500	0.6958	28	1009.136	84.372
25000	0.7587	28	1907.726	74.874

Summary of Results for CIFAR-10 Dataset with Different Subset Sizes

Subset Size	Test Accuracy	Stopped at Round	Total Training Time (s)	Total Data Summary Time (s)
100	0.7175	174	131.216	25.452
200	0.7518	131	80.304	25.234
600	0.7603	40	33.237	26.329
1500	0.7693	24	24.147	27.428
3000	0.8263	18	18.251	27.975
4500	0.8345	16	20.224	28.468
6000	0.838	17	26.755	28.659
15000	0.8523	17	55.027	30.511

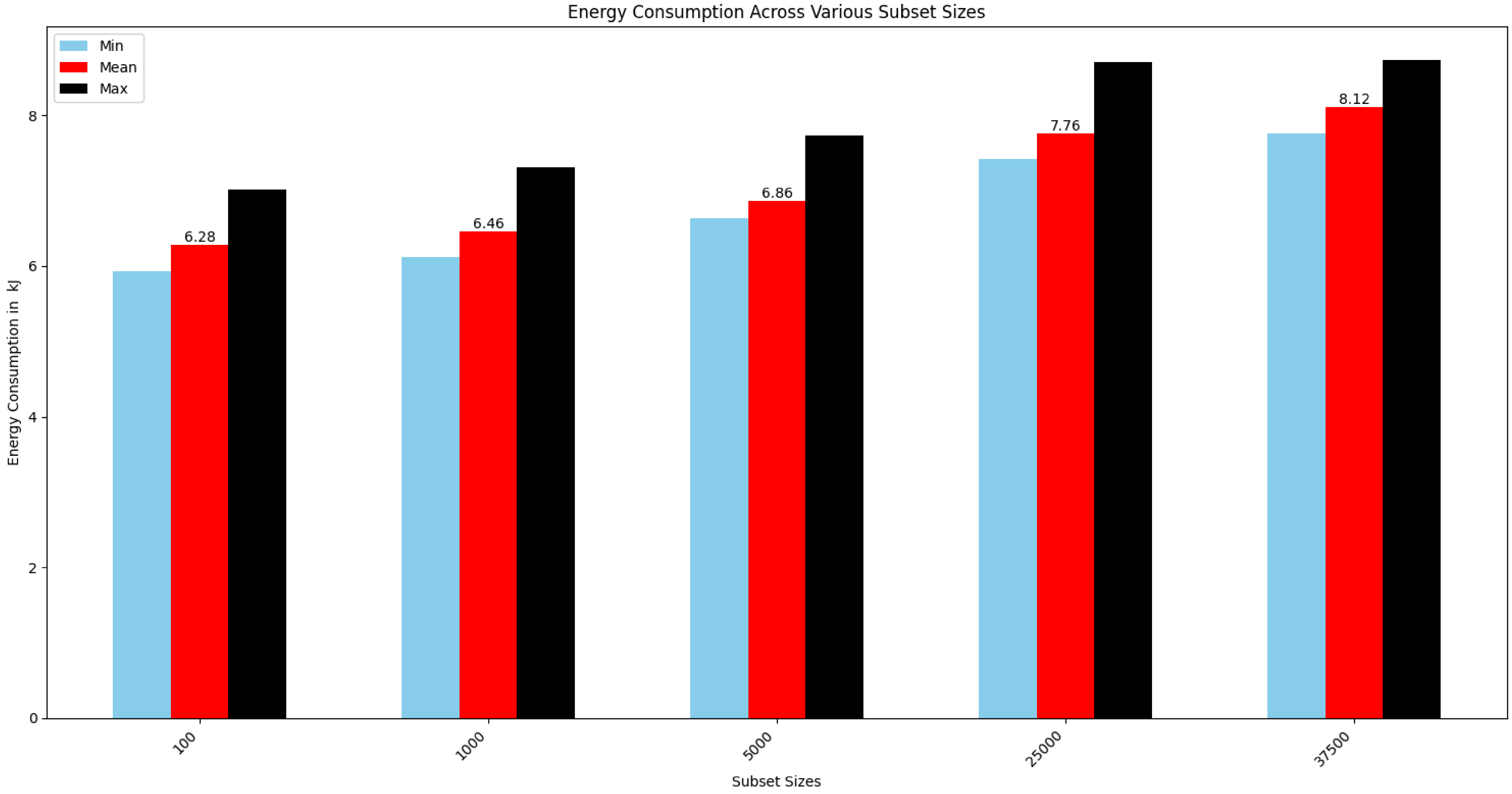
Summary of Results for Fashion-MNIST Dataset with Different Subset

CRAIG Results

Early stopping criteria:

- Patience = 3
- Min Delta = 0.01

- Training accuracy instead of test accuracy.



Energy Consumption of CRAIG Sampling - CIFAR10

- 100 repetitions

CONCLUSION

- Optimal Subset Sizes: For energy-efficient training, subset sizes between 1-10% are optimal for simpler datasets, while more complex datasets may require larger subsets (up to 25%).

- Sample similarity sampling and gradient similarity sampling can be effective in some cases, but they are still significantly less effective at reducing energy compared to random sampling.

- The results we obtained are highly sensitive to the early stopping techniques we implemented.

THANK YOU !