# Evolution-aware VAriance (EVA) Coreset Selection for Medical Image Classification

Anonymous Authors


## ABSTRACT

In the medical field, managing high-dimensional massive medical imaging data and performing reliable medical analysis from it is a critical challenge, especially in resource-limited environments such as remote medical facilities and mobile devices. This necessitates effective dataset compression techniques to reduce storage, transmission, and computational cost. However, existing coreset selection methods are primarily designed for natural image datasets, and exhibit doubtful effectiveness when applied to medical image datasets due to challenges such as intra-class variation and inter-class similarity. In this paper, we propose a novel coreset selection strategy termed as ***Evolution-aware VAriance (EVA)***, which captures the evolutionary process of model training through a dual-window approach and reflects the fluctuation of sample importance more precisely through variance measurement. Extensive experiments on medical image datasets demonstrate the effectiveness of our strategy over previous SOTA methods, especially at high compression rates. EVA achieves 98.27% accuracy with only 10% training data, compared to 97.20% for the full training set. None of the compared baseline methods can exceed Random at 5% selection rate, while EVA outperforms Random by 5.61%, showcasing its potential for efficient medical image analysis.



## KEYWORDS

Coreset Selection, Medical Image Classification, Evolution-aware Variance


## 1 INTRODUCTION

In the medical field, data collection and processing are essential for delivering accurate and reliable diagnoses and treatment plans. Medical imaging data, typically characterized by high dimensionality and large volumes, necessitates substantial resources for storage and transmission. Moreover, training deep learning models on large-scale medical image datasets requires extensive computational resources and time. This presents challenges in resource-limited settings, such as remote medical facilities where effective medical image analysis is crucial, or on mobile devices where real-time monitoring and analysis are needed. Therefore, efficient data compression and processing techniques become imperative. In this context, coreset selection, or dataset pruning, emerges as a promising approach to mitigate these challenges. Coreset selection condenses a given large-scale dataset into a significantly smaller subset, known as the coreset. The coreset is expected to preserving the essential knowledge of the original full dataset such that the former yields a similar performance as the latter.

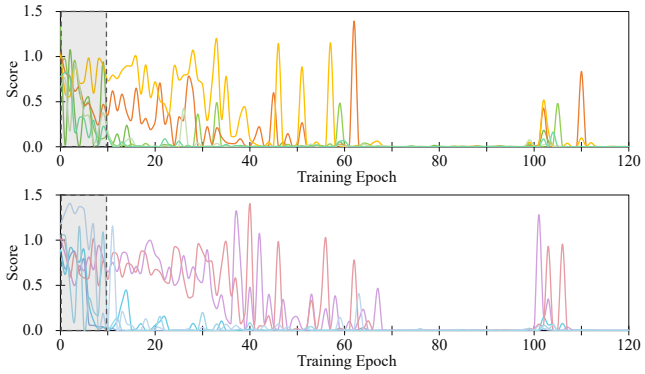


**Figure 1: Existing single-timeframe/window snapshots methods fail to capture the fluctuations of sample importance across epochs. Different samples are denoted in different colors. Here, we measure importance score using the error vector score, a snapshot-based criterion defined in [39], which considers only the first 10 epochs as indicated by the dashed box. These scores are obtained by training ResNet-18 on dataset OrganAMNIST.**

Numerous coreset selection works [19, 31, 33, 37, 40, 51, 56] have explored various criteria for identifying important data samples, including geometry distance [44, 50], uncertainty [11], loss [39, 48], decision boundary [13, 32], and gradient matching [35]. However, most of these methods have been validated mainly on natural image datasets, such as CIFAR-10, CIFAR-100 [28], and not extensively on medical datasets. The applicability of those methods for medical image datasets are under exploration, given the unique characteristics of medical images. Compared to natural image datasets, the intra-class variation and inter-class similarity of medical image datasets [46] pose specific challenges to coreset selection. On the one hand, in medical imaging, samples within the same category can exhibit significant differences, making it difficult to capture consistent features for each class. This variation largely comes from the diversity in disease manifestation across patients and discrepancies in imaging conditions. On the other hand, the challenge of inter-class similarity arises when images representing different diseases exhibit similar visual characteristics. Fig. 6 provides a more straightforward demonstration of this characteristic. These factors contribute to the complexity of medical image analysis and underscore the need for sophisticated coreset selection methods that can effectively address these challenges.

**Unpublished working draft. Not for distribution.**



*ACM MM, 2024, Melbourne, Australia*



ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

To enable the coreset to effectively approximate the model's performance on the full training set with fewer samples, it is essential to consider the training process on the original dataset. This necessitates that the selection methods should effectively capture the varying importance of samples at different training stages. Yosinski et al. [58] highlighted that shallow layers of the network learn general features, while deeper layers learn task-specific features. Han et al. [17] observed that deep models tend to memorize easy instances initially and adapt to harder instances as training progresses. These studies confirm the evolutionary nature of deep learning from simpler to more complex stages. Given these observations[17, 58], we posit that in the domain of medical imaging, the training process of deep learning models exhibits similar characteristics. For instance, in kidney images, the model initially learns the general kidney shape and gradually distinguishes more detailed features of different kidneys. Moreover, as shown in Fig. 1, the significance of samples in enhancing the model performance varies across different training stages [7, 20, 48, 60]. Specifically, certain samples may be crucial for the model's initial learning phase, while others gain importance in the later stages of training.

Most of the existing methods evaluate sample importance using a snapshot of training progress. For example, Xia et al. [51] calculate the distribution distances of features at the end of training. Zhang et al. [60] have proved that the importance scores of samples varies with epochs during training, resulting in significant variations in the constructed coresets at different snapshots. Therefore, methods reliant on single-timeframe snapshots might be inadequate for capturing the comprehensive evolution of model training, overlooking the dynamic characteristics of learning process.

Expanding the scope of the considered training dynamics is a straightforward approach to address this limitation. Previous studies have attempted to incorporate training dynamics using various methods. For example, Pleiss et al. [41] measures the probability gap between the target class and the second-largest class in each epoch; Paul et al. [39] utilize the expected value of error vector scores generated by a few epochs in early training (the first 10 epochs). While this approach partially expands the scope of the considered training dynamics, it overlooks the potential effectiveness of later stages of training, and more importantly, it focuses on samples with high expected error values, indicating that these samples are consistently predicted incorrectly over many training iterations. Such samples may just be too difficult/noisy and may degrade the model performance [7]. Toneva et al. [48] count the number of forgetting events during training, which occur when samples, previously classified correctly, are subsequently predicted incorrectly. However, this counting approach only provides the discrete probability of an event, lacking the granularity needed to reflect the variations of sample contributions throughout the training process.

To address these limitations, in this paper, we propose a novel sample importance scoring strategy called **Evolution-aware VAriance (EVA)**, aiming at achieving reasonable and effective compression of medical image datasets. Firstly, to mitigate the biases from focusing solely on a snapshot or single segment of the training process, we introduce a dual-window approach that considers training dynamics at different stages. This strategy provides a more holistic understanding of the model's learning evolution, enabling nuanced assessment of sample importance as the model evolves from learning general to specific features. Secondly, within each window, to reflect the fluctuation of sample importance during the model training process in a more precise way, we propose to measure the variance of samples' error vector. The combination of these two strategies provides a more refined and accurate evaluation of sample importance, enabling a more effective coreset selection that aligns with the dynamic nature of neural network training. This approach is particularly beneficial in high compression scenarios for medical image datasets, where maintaining accuracy and reliability is challenging but crucial.

In a nutshell, our contributions can be summarized as follows.

- We identify the limitations of existing coreset selection methods in capturing the evolutionary nature of model training and the fluctuations in sample importance within medical image datasets.
- We thereby propose a novel coreset selection strategy called **Evolution-aware VAriance (EVA)**, which features two key components. The first is a dual-window approach that captures the training dynamics by considering distinct stages of the learning process. The second is the employment of variance measurement on samples' error vectors, offering a granular and more precise evaluation of each sample's contribution to the model training.
- Extensive evaluations on the OrganAMNIST and OrganSMNIST datasets demonstrate that our EVA strategy outperforms SOTA methods at challenging low selection rates while achieving comparable accuracy at high selection rates, showcasing its potential for efficient medical image analysis.

## 2 PRELIMINARIES

In this paper, vectors and matrices are denoted by bold-faced letters. Given a large-scale dataset, we denote the full training set contains N samples as $\mathbb{T} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$, where $\boldsymbol{x}_n \in \mathbb{R}^D$ represents the input feature vector and the corresponding ground-truth label is $\boldsymbol{y}_n \in \mathbb{R}^{1\times C}$, $C$ is the number of classes. All samples are drawn i.i.d. from a underlying distribution $\mathcal{P}$. We define the neural network as $f_{\boldsymbol{\theta}}$, parameterized by the weight vector $\boldsymbol{\theta}$. The model output $f_{\boldsymbol{\theta}}(\boldsymbol{x}_n) \in \mathbb{R}^{1\times C}$ represents the predicted probabilities of each class. Coreset selection aims to construct a subset (or coreset) $\mathbb{S} = \{(\boldsymbol{x}_m, \boldsymbol{y}_m)\}_{m=1}^{M}$ ($\mathbb{S} \subset \mathbb{T}$) that captures the essential characteristics of the full dataset, enabling model $f_{\boldsymbol{\theta}^{\mathbb{S}}}$ trained on $\mathbb{S}$ to achieve comparable or even superior performance compared to model $f_{\boldsymbol{\theta}^{\mathbb{T}}}$ trained on the entire training set $\mathbb{T}$. The data selection rate $\alpha$ in constructing the coreset is then $\frac{M}{N}$. Under these definitions, following previous work [44], we formulate the objective of coreset selection as,

$$\mathop{\mathbb{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}\\ \boldsymbol{\theta}_0\sim\mathcal{G}}}\left[\ell(\boldsymbol{x},\boldsymbol{y};f^{\mathbb{S}}_{\boldsymbol{\theta}_0})\right] \simeq \mathop{\mathbb{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}\\ \boldsymbol{\theta}_0\sim\mathcal{G}}}\left[\ell(\boldsymbol{x},\boldsymbol{y};f^{\mathbb{T}}_{\boldsymbol{\theta}_0})\right] \tag{1}$$

where $f^{\mathbb{S}}_{\boldsymbol{\theta}_0}$ and $f^{\mathbb{T}}_{\boldsymbol{\theta}_0}$ represent the neural networks trained on $\mathbb{S}$ and $\mathbb{T}$ with weight $\boldsymbol{\theta}_0$ initialized from distribution $\mathcal{G}$.

## 3 METHODOLOGY

To construct a coreset that satisfies Eq. 1, the error/loss-based approaches propose to measure the contribution of each sample by considering factors such as the loss, gradient, or its influence on

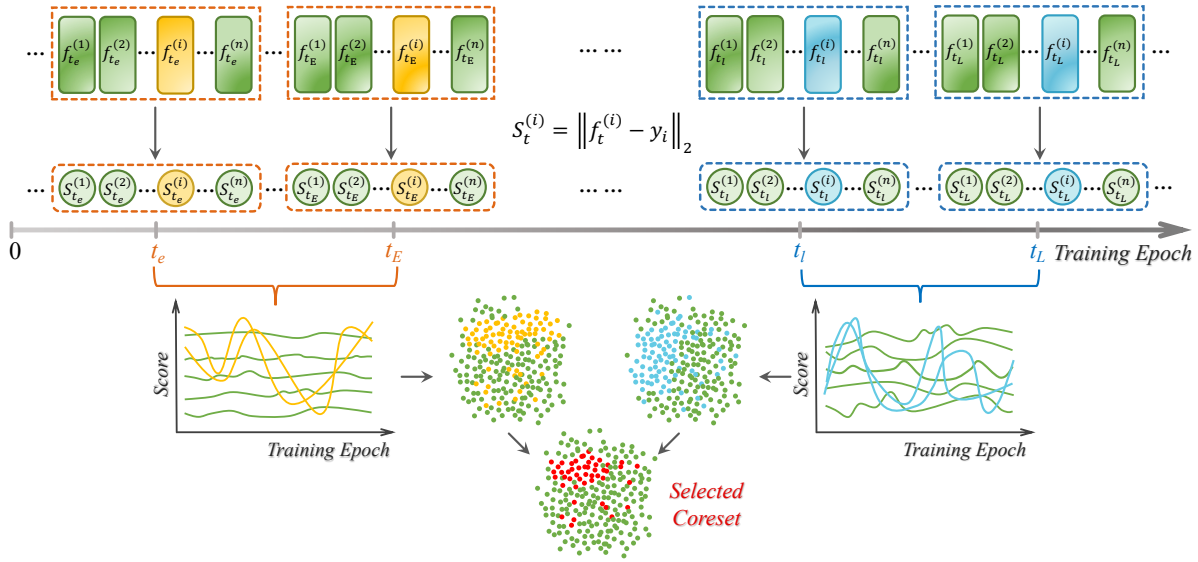


**Figure 2: The pipeline of our proposed EVA . First, we record individual predicted probabilities $f_t^{(i)} = f_{\theta_t}(x_i)$ of samples during training. Then, we measure a score $\mathcal{S}_t^{(i)}$ for each sample, i.e. the L2 norm of error vector. Next, the variance of scores within a window of epochs are calculated to reflect the fluctuation of each sample's contribution. Samples that fluctuate the most are considered important in this stage. Finally, we identify samples that exhibit high importance in dual-window.**

other samples' prediction during model training [16]. In this context, samples that contribute more to the error or loss are considered more important and are thus selected as part of the coreset.

In this section, we delve into the specifics of our proposed Evolution-aware VAriance (EVA) strategy, which comprises two key components. Firstly, we describe how EVA reflects the epoch-level fluctuation by calculating the variance of error-based scores in Sec. 3.1. Following that, we elaborate on how EVA captures the training evolution through a dual-window approach in Sec. 3.2.

## 3.1 Reflecting Epoch-Level Fluctuation via Variance

To approximate the individual contribution of each sample to the reduction in model loss, we initially calculate the variance of error-based scores over a segment of epochs. This process can be further divided into the following steps.

**Step 1. Single Epoch Scoring.** In this step, we concentrate on calculating the error score for each sample at a specific epoch across multiple independent runs. Specifically, for each sample $(x_i, y_i)$ in the training set, we first consider a single epoch $t$ and compute the total mean square error (MSE) across all categories using the equation below:

$$\mathrm{MSE}_t^{(i)} = \sum_{j=1}^{C} (\hat{y}_{ij} - y_{ij})^2, \tag{2}$$

where $\hat{\boldsymbol{y}}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$, therefore $\hat{y}_{ij}$ denotes the model output of the $i$-th sample for the $j$-th category, and $y_{ij}$ is the one-hot encoding of the ground-truth label for the $i$-th sample in the $j$-th category. Then, we take the square root of the total MSE for each sample.

Thus, for each sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ at epoch $t$, we have the L2 norm of the error vector, representing the discrepancy between model predictions and ground-truth labels:

$$\mathcal{S}_t^{(i)} = \| f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \boldsymbol{y}_i \|_2, \tag{3}$$

This process yields a sequence of error scores, providing insights into the prediction performance of the model across different training iterations.

**Step 2. Variance Across Multiple Epochs.** Having obtained the error scores for each sample at individual epochs, in this step, we proceed to assess the variability of scores across multiple epochs by calculating the variance over a segment of epochs. Specifically, for each sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, we analyze the training dynamics over a span of $K$ epochs, from $t$ to $t + K - 1$. The error-based scores for this period are represented as $\left\{\mathcal{S}_t^{(i)}, \mathcal{S}_{t+1}^{(i)}, ..., \mathcal{S}_{t+K-1}^{(i)}\right\}$. We then compute the variance of these scores within the $K$-epoch window using the following equation:

$$\mathcal{V}_t^{(i)} = \frac{1}{K} \sum_{k=t}^{t+K-1} \left( \mathcal{S}_k^{(i)} - \mathcal{E}_t^{(i)} \right)^2, \tag{4}$$

where $\mathcal{E}_t^{(i)} = \frac{1}{K} \sum \mathcal{S}_k^{(i)}$ denotes the mean value within the $K$-epoch window. This calculation provides insight into the consistency or variability of the error-based scores for each sample over a specified segment of training epochs, enabling a more precise understanding of the subtle fluctuations in a sample's impact on model performance over time.

### 3.2 Capturing Training Evolution with Dual-Window

As mentioned in Sec. 1, snapshot-based methodologies often fall short in capturing the comprehensive evolution of model training, thus warranting an expansion in the scope of considered training dynamics. One approach to broaden the scope is to sample some epochs during the training dynamics. However, random or probabilistic sampling of epochs may not effectively capture the dynamic changes in sample importance throughout the entire training process. Another method is to consider epochs within a certain window, as we did in Eq. 4. Nevertheless, this approach carries the risk of excessive bias towards specific training phases.

Therefore, we introduce a dual-window approach to capture the evolution of the training process more comprehensively. The first window focuses on the early stages of training, during which the model primarily learns general features. Samples that significantly impact the overall model performance are likely to exhibit high importance during this stage. The second window targets the later stages of training, where the model gradually learns more specific task-related features. The importance of samples that have a significant impact on the overall model performance may increase or decrease during this stage. By integrating information from dual windows, we aim to identify samples that exhibit high importance in both early and late stages. This implies that these samples contain both general features and specific task-related features. Additionally, the continuous sequence of epochs provides more temporal information, allowing for a more comprehensive assessment of sample importance throughout the entire training process. Overall, the use of two windows provides a more nuanced understanding of training dynamics and sample importance, enhancing the effectiveness of the selection process for constructing a coreset. This effectiveness has been proved in Sec. 4.3.

To maintain consistency with Sec. 3.1, in the dual-window scenario, we also consider windows spanning $K$ epochs. We define the total number of training epochs as $T$, the first window ranges from $t_e$ to $t_E = t_e + K - 1$, and the second window ranges from $t_l$ to $t_L = t_l + K - 1$. These windows are non-overlapping ($t_E < t_l$). Specifically, for each sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, we compute the scores within each window of epochs, denoted as $\left\{\mathcal{S}_k^{(i)}\right\}_{k=t_e}^{t_E}$ and $\left\{\mathcal{S}_k^{(i)}\right\}_{k=t_l}^{t_L}$. The variance of these scores in Eq. 4 can be formulated as:

$$\begin{aligned} \mathcal{V}_e^{(i)} &= \frac{1}{K}\sum_{k=t_e}^{t_E}\left(\mathcal{S}_k^{(i)} - \mathcal{E}_e^{(i)}\right)^2, \\ \mathcal{V}_l^{(i)} &= \frac{1}{K}\sum_{k=t_l}^{t_L}\left(\mathcal{S}_k^{(i)} - \mathcal{E}_l^{(i)}\right)^2, \end{aligned} \tag{5}$$

Here, $\mathcal{E}_e^{(i)}$ and $\mathcal{E}_l^{(i)}$ denote the average score of sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ in two windows, respectively.

Finally, we aggregate the variances from both windows to identify samples that demonstrate high importance in two stages. Thus the EVA score of each sample can be represented as:

$$\mathcal{V}^{(i)} = \mathcal{V}_e^{(i)} + \mathcal{V}_l^{(i)} \tag{6}$$

We then sort samples in the full training set $\mathbb{T}$ by their EVA score $\mathcal{V}^{(i)}$. Samples with higher scores are deemed more effective at reducing training loss. Given a selection rate $\alpha$, we select the top-ranked M samples to form the coreset, where $M = \lceil \alpha N \rceil$.

Algorithm 1 provides a detailed explanation of the procedure for the EVA scoring strategy.

**Algorithm 1** Evolution-aware VAriance (EVA) Scoring Strategy

**Inputs:** Full training set $\mathbb{T} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$; Selection rate $\alpha$; Network $f_{\boldsymbol{\theta}}$ with weight $\boldsymbol{\theta}$; Epochs $T$; Iteration $I$ pre epoch; Early window $(t_e, t_E)$; Late window $(t_l, t_L)$.

```
for t = 1 to T do
    for i = 1 to I, sample a mini-batch B_i ⊂ T do
        Obtain predicted probabilities f_θt(x_n), x_n ∈ B_i
        Calculate S_i^(n) by defined Eq. 3 for each x_n
        Update S_t^(n) += S_i^(n)
    end for
    if t_e ≤ t < t_E then
        Calculate V_e^(n) by defined Eq. 5 of early window, x_n ∈ T
    else if t_l ≤ t < t_L then
        Calculate V_l^(n) by defined Eq. 5 of late window, x_n ∈ T
    else if t = t_L then
        Update V^(n) by defined Eq. 6 as the EVA score of x_n
    end if
end for
Sort samples by V^(n) in descending order, x_n ∈ T
```

**Output:** Top-M samples as the coreset $\mathbb{S} = \{(\boldsymbol{x}_m, \boldsymbol{y}_m)\}_{m=1}^{M}$

## 4 EXPERIMENTS

In this section, we provide a comprehensive set of experiments and analyses to showcase the effectiveness of our proposed Evolution-aware VAriance scoring strategy in diverse scenarios. We start by empirically evaluating the performance of our EVA method by comparing it with other baselines (Sec. 4.2). Subsequently, we conduct a series of ablation studies to investigate the effectiveness of the proposed two main components: variance measurement and dual-window strategy (Sec. 4.3). Additionally, we perform cross-architecture experiments to evaluate the robustness of our coresets, assessing their performance when selected on one architecture and tested on others.

### 4.1 Experiment Setup

**Datasets.** MedMNIST is a large-scale collection of medical images comprising 10 datasets, covering multi-modal, diverse data scales (from 100 to 100,000) and classification tasks. The classification performance of this public large-scale datasets has been validated as effective in [54]. More details about MedMNIST are included in Sec. 5.1. In this work, considering the time-consuming training, the effectiveness of the proposed method is primarily evaluated on two 2D datasets from MedMNIST: OrganAMNIST and OrganSMNIST [3, 52], both derived from 3D computed tomography (CT) images from the Liver Tumor Segmentation Benchmark (LiTS). These datasets are designed for multi-class classification tasks, involving 11 body organs with labels including the bladder, femur

**Table 1: Performances of ResNet-18 using various coreset selection methods on MedMNIST medical datasets. All training is repeated 3 times with different random seeds to calculate mean accuracy with standard deviation. The first and second best results in each column are marked in red and blue, respectively.**

| | OrganAMNIST | | | | | OrganSMNIST | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 2% | 5% | 10% | 20% | 30% | 2% | 5% | 10% | 20% | 30% |
| Full dataset | | | 98.39 ±0.02 | | | | | 91.76 ±0.55 | | |
| Random | 87.63 ±0.76 | 93.43 ±0.65 | 95.68 ±0.45 | 97.30 ±0.13 | 98.14 ±0.13 | 58.74 ±0.76 | 73.10 ±1.84 | 80.95 ±0.66 | 85.77 ±1.14 | 87.64 ±0.72 |
| Forgetting [48] | 15.58 ±0.47 | 38.53 ±2.78 | 75.85 ±1.69 | 97.22 ±0.38 | 98.11 ±0.04 | 4.33 ±0.22 | 22.33 ±0.31 | 33.15 ±0.60 | 64.43 ±1.23 | 81.28 ±2.31 |
| Entropy [11] | 41.46 ±3.46 | 55.37 ±1.4 | 69.04 ±1.16 | 77.07 ±1.29 | 91.98 ±0.83 | 27.93 ±2.08 | 41.69 ±0.73 | 59.86 ±1.84 | 78.69 ±2.13 | 86.20 ±0.54 |
| EL2N [39] | 14.16 ±1.14 | 40.68 ±3.36 | 81.25 ±3.22 | 97.25 ±0.24 | 98.16 ±0.30 | 17.63 ±1.59 | 23.24 ±1.88 | 28.24 ±1.44 | 37.58 ±1.53 | 60.06 ±2.14 |
| AUM [41] | 12.81 ±2.62 | 35.10 ±3.46 | 68.44 ±0.95 | 93.76 ±1.89 | 98.12 ±0.14 | 4.56 ±0.18 | 7.01 ±1.24 | 22.13 ±1.86 | 39.87 ±2.19 | 65.93 ±1.61 |
| CCS [61] | 88.05 ±0.62 | 93.51 ±0.10 | 95.58 ±0.32 | 96.86 ±0.25 | 97.18 ±0.08 | 58.43 ±0.25 | 71.73 ±0.83 | 78.46 ±0.18 | 83.64 ±0.55 | 84.94 ±0.22 |
| EVA (Ours) | 88.83 ±0.88 | 94.43 ±1.32 | 97.20 ±0.34 | 98.27 ±0.57 | 98.63 ±0.34 | 61.23 ±0.75 | 78.71 ±0.93 | 83.11 ±0.72 | 86.38 ±1.02 | 88.77 ±0.43 |

(left and right), heart, kidney (left and right), liver, lung (left and right), pancreas, and spleen. OrganAMNIST, previously known as OrganMNIST-Axial in MedMNIST v1 [53], consists of 58,830 axial view slices of abdominal CT images, distributed into 34,561 training, 6,491 validation, and 17,778 testing images. OrganSMNIST, formerly OrganMNIST-Sagittal, includes 25,211 abdominal CT images split into 13,932 training, 2,452 validation, and 8,827 testing images.

**Baselines and Networks.** We compare our method against six representative baselines, the latter five of which are state-of-the-art (SOTA) methods: 1) **Random**; 2) **Forgetting score** [48]; 3) **Entropy** [11]; 4) **EL2N** [39]; 5) **Area under the margin AUM)** [41]; 6) **Coverage-Centric Coreset Selection (CCS)** [61]. Details of these baselines are provided in the Supplementary material due to space limitations. The effectiveness of these strategies is evaluated based on their ability to select representative samples for coreset construction using various criteria. For all baselines except CCS, coresets are formed by pruning less important examples according to the respective importance metric.

The effectiveness of our method is primarily evaluated using ResNet-18 [18]. We also conduct cross-architecture generalization experiments with ResNet-50 [18], MobileNet [42] and VGGNet [45] to validate its robustness across different models. Further details are available in the Supplementary material.

**Implementation details.** To ensure fairness in our comparisons, we adhere to the experimental setup outlined in [61]. Our method is implemented using PyTorch [38] and all models are trained on an NVIDIA 3090 GPU. Unless specified otherwise, we utilize the same network architecture, ResNet-18, for both the coreset and the surrogate network on the full dataset. We maintain consistency in all hyperparameters and experimental settings before and after coreset selection. The surrogate network is trained for 200 epochs across all datasets. Initially, we train a network on the complete dataset to establish baseline performance. Subsequently, we calculate the importance scores by assessing the variance of each sample's error vector across multiple epochs within a dual-window. As to the start epoch and end epoch of each window, we employ a grid search with a 10-step size ($K = 10$). This process helps us identify the most effective window combinations, denote as $(t_e, t_E)$+$(t_l, t_L)$ for different datasets and selecting rate $\alpha$.

## 4.2 Benchmark Evaluation Results

Our systematic comparison of EVA against other baselines, as detailed in Sec. 4.1, reveals its superior performance on the OrganSMNIST and OrganAMNIST medical datasets, particularly at more challenging selection rates. As shown in Tab. 1, our Evolution-aware VArianceapproach consistently achieves top-ranking performance, underscoring its robustness in coreset selection. In addition, on the OrganAMNIST dataset, EVA nearly matches the full dataset's performance at a 20% selection rate and surpasses it at 30%, highlighting its efficiency in utilizing smaller datasets. Notably, at extremely low selection rate of 2% and 5% on the OrganSMNIST dataset, EVA surpasses the Random baseline by a margin of 2.49% and 5.61%, respectively, illustrating its effectiveness even with severely limited data, establishing the method's capability to discern and retain the most influential samples for model training.

The baselines, including well-established SOTA methods, do not exhibit the same level of performance at these lower selection rates, often failing to exceed the benchmark set by random selection. This trend highlights the limitations of traditional coreset selection methods when dealing with the complexities of medical datasets.

Here, our experiments focus on low selection rates scenarios, but EVA also maintains competitive performance at high selection rates. Moreover, our methodology's effectiveness is not confined to medical imaging datasets alone. Preliminary experiments on widely recognized natural image datasets, such as CIFAR, corroborate that EVA stands out by surpassing most SOTA methods. Detailed

results from these additional experiments are documented in the Supplementary materials due to space constraints.

### 4.3 Ablation Study and Analysis

We delve into ablation studies to dissect the contributions of the variance and dual-window components in our method. By systematically removing each component and evaluating their impact on performance, we elucidate their individual roles in enhancing coreset selection accuracy. In this context, we partition our experiments into four conditions: VAR-S, EXP-S, VAR-D, and EXP-D. Here, VAR-S denotes calculating variance in a single window, EXP-S represents computing expectation in a single window; VAR-D indicates variance calculation in dual-window, and EXP-D signifies expectation computation in dual-window.

**Effectiveness of Variance.** In this section, to demonstrate the effectiveness of variance measurement, we display the test accuracy results of calculating the expectation and variance of the samples' error vectors within a single window or dual windows on different datasets. As shown in Fig. 3, these results were obtained under varied selection rates from 2% to 30%.

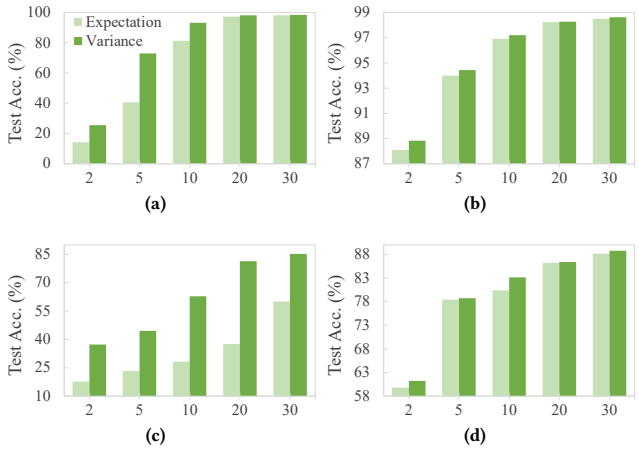


**Figure 3: Ablation study on the summary statistics. We validated the effectiveness of variance measurement under single-window and dual-window settings on OrganAMNIST (a)(b) and OrganSMNIST (c)(d). In (a) and (c), we contrast the EXP-S and VAR-S strategies within an early 10-epoch window. (b) and (d) explore the EXP-D and VAR-D strategies in dual-window setting.**

The first thing we notice is that, on both datasets, as the selection rate increases, the effectiveness of the models trained on the samples selected by both statistics tends to increase on the test set. This is intuitive because as the number of samples selected increases, the information richness of the selected samples are effectively preserved.

Besides, we can observe that at each selection rate, the variance measurement has better performance in coreset selection compared to the expectation measurement, and this advantage is especially significant at low selection rates. For example, in Fig. 3c, the test accuracy under VAR-S is at least 20% higher than under EXP-S for all compared selection rates. The consistent superiority of variance (VAR-S and VAR-D) suggests its robustness as a measure, further proving our previous points that (1) *Expectation* may mask variability within the data by averaging contributions, thereby potentially underrepresenting the underlying fluctuations. Samples with large expectation values may be consistently predicted incorrectly over many training iterations, indicating them too noisy/difficult and detrimental to the model's performance; (2) *Variance* captures the degree to which sample contributions fluctuate over training iterations. High variance in sample errors suggests that their influence on the model is not consistent but varies significantly, potentially due to their informative nature or because they are challenging for the model to learn. At low selection rates, samples with higher variance are indicative of a greater potential to contribute to the model's generalization ability, as they embody the critical challenges within the learning task.

**Effectiveness of dual-window.** In this section, we demonstrate the effectiveness of the dual window setting and analyze the results for different window combinations. First, we compare the performance of using single-window and dual-window on different datasets (as shown in Fig. 4). Similar to the former part, we utilized the variance and expectation of errors within single and dual windows as importance metrics. The results consistently demonstrate the advantage of dual windows over single window across all selection rates. This advantage, akin to the findings from the variance ablation experiments, is more pronounced at lower selection rates. For instance, on dataset OrganSMNIST, at selection rates of 2%, the variance calculated within dual windows exhibited an improvement of 2.29%, compared to the single-window approach, suggesting that employing dual-window calculation for scores enables more effective capturing of the diversity and variability of sample importance.

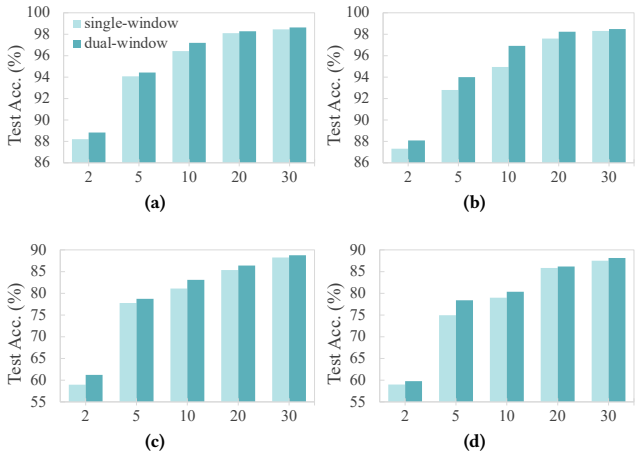


**Figure 4: Ablation study on the window setting. The results are obtained on OrganAMNIST (top row) and OrganSMNIST (bottom row). Performance of the VAR-S versus VAR-D strategies is illustrated in (a) and (c), while (b) and (d) show comparisons between EXP-S and EXP-D strategies.**

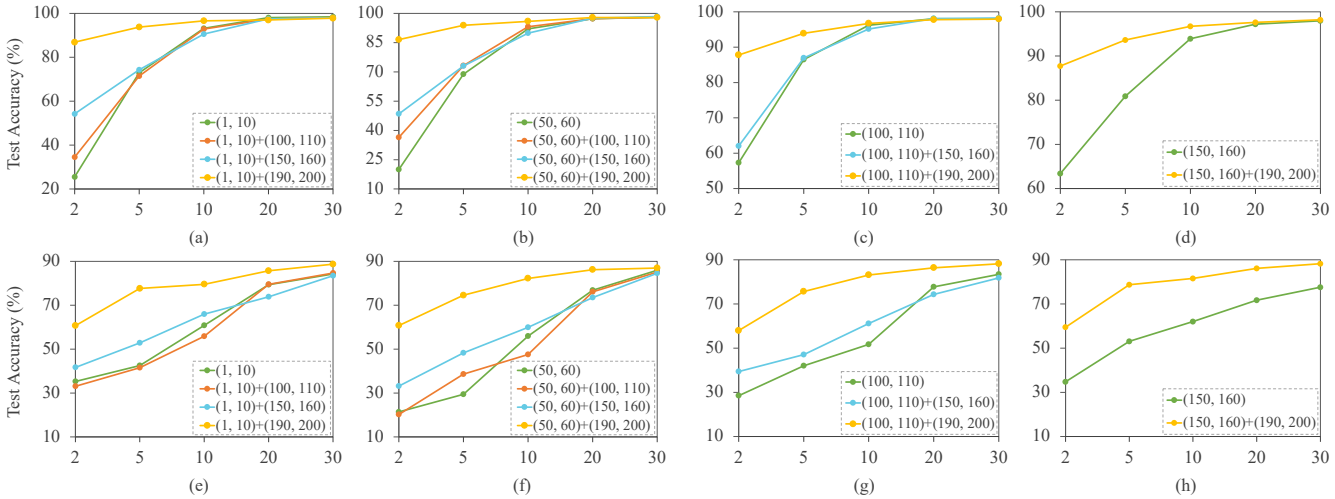


**Figure 5: Comparison of different window combinations. These windows represent different training phases. (a)-(d) show experimental results for OrganSMNIST, and (e)-(h) for OrganAMNIST, with each line depicting a unique window combination (single or dual windows).**

Moreover, in the dual-window setting, we further explore the effect of the combination of windows at different periods on model performance. Fig. 5 reveals two critical insights: (1) At a high compression rate of 2%, dual-window combinations show a definitive advantage over single-window ones on both datasets. This can be attributed to the dual-window's ability to encapsulate more diverse information from different stages of the training process, providing a broader perspective for coreset selection. (2) As the selection rate increases, allowing for larger data budgets, corresponding to the need of capturing a wider range of training dynamics. The implication here is that the windows selected for the dual-window setting should ideally come from a later stage in the training process, when the model has begun to stabilize and the samples are more reflective of the generalization capabilities required for the test. The results on OrganAMNIST suggests that the early dual-window stage may not be sufficient for selecting a more representative coreset.

# 5 RELATED WORKS

## 5.1 Medical Imaging

**Challenges in Medical Imaging with Deep Learning.** Medical imaging technology has brought transformative advancements to the diagnosis of a variety of diseases in the past few decades, enabling earlier detection and the development of more personalized treatment plans. Deep learning (DL), in particular, has been widely used in various medical imaging tasks and has achieved remarkable success in many medical imaging applications [8, 36, 43, 62, 63], enhancing the accuracy of diagnoses through the innovative use of historical data [29].

Despite the substantial progress, integrating deep learning into medical imaging is fraught with challenges [22]. The effectiveness of DL in this context is largely dependent on the availability of large, well-annotated datasets tailored for specific tasks and reliant on advances in high-performance computing. The necessity for vast complex datasets introduces complications such as inconsistencies in data quality, arising from variations in imaging equipment and protocols. Moreover, the extensive volume of medical data demands significant computational resources, posing logistical challenges for efficient processing [62]. Additionally, the inherent heterogeneity of medical images, characterized by a multimodal probability distribution, complicates the model training process by requiring algorithms capable of handling diverse visual features and patterns within the data. Another issue is the inter-class similarity and intra-class variation, as depicted in Fig. 6, where different diseases may appear similar, and the same disease may present differently across patients.

**MedMNIST: A Standardized Dataset for Biomedical Imaging.** To address some of these challenges, MedMNIST, a large-scale MNIST-like collection of standardized biomedical images, provides a comprehensive dataset for research and application. This dataset includes 12 datasets for 2D imaging and 6 for 3D, all pre-processed into 28x28 or 28x28x28 pixels with corresponding classification labels. MedMNIST encompasses primary data modalities in biomedical imaging, including abdominal CT, chest X-ray, breast ultrasound, and blood cell microscopy, making it an ideal choice for multi-modal machine learning in medical image analysis. Additionally, it supports various classification tasks such as binary/multi-class, ordinal regression, and multi-label classification, further establishing its utility for developing and testing deep learning models in medical imaging.

## 5.2 Dataset Compression

The proliferation of large-scale datasets in deep learning necessitates the compression of data size to meet specific requirements, such as computational efficiency and storage constraints. Therefore, the identification of key samples serves a fundamental role not only in dataset pruning but also across a spectrum of machine learning tasks, such as active learning [1, 4, 14], where the model is trained iteratively on a subset of the dataset, and only the most informative

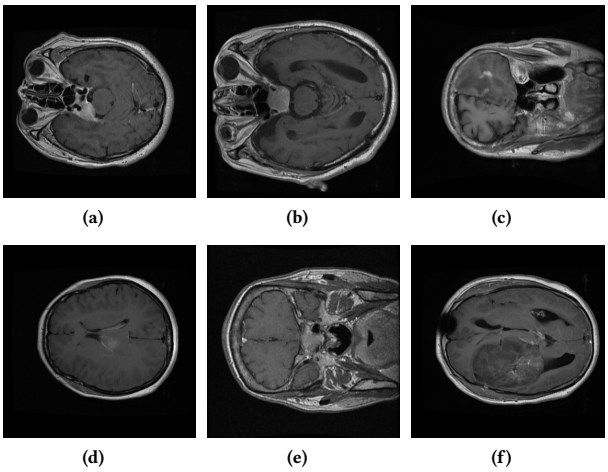


**Figure 6: Examples of the intra-class variation and inter-class similarity in medical image classification. These axial brain tumor images come from the public dataset provided by Jun Cheng et al.[9]. Each column respectively represents a brain tumor category: meningioma (a)(d), pituitary (b)(e), and glioma (c)(f). The variation within the same category can be noticed by observing the two instances in each column. Furthermore, the similarity between different classes is illustrated by comparing (a)(b), (c)(e), and (d)(f).**

samples are selected for inclusion in subsequent training rounds. Techniques such as uncertainty sampling and query-by-committee have been proposed to select data samples that are most beneficial for model improvement. Continual learning [57], where a memory buffer is maintained to store informative training samples from previous tasks for rehearsal in future tasks. And other problems like noisy learning [35], clustering [2], semi-supervised learning [5], and unsupervised learning [10].

Dataset pruning, also known as coreset selection, can generally be categorized into several groups: Score-based techniques [11, 15, 34, 39, 47, 48], methods driven by coverage or diversity considerations [44, 50, 51], and strategies grounded in optimization [21, 23–25, 27, 35, 49, 55]. Specifically, score-based techniques first assign an importance score to each training sample based on its influence over a specific permanence metric during model training. The samples are then sorted by their scores, and those within a certain range are selected to construct the coreset.

Besides, in the sphere of data-efficient deep learning, associated topics include techniques like data distillation [6, 30] and data condensation [12, 26, 30], which seeks to condense the knowledge contained in a large dataset into a smaller, distilled dataset. This technique often involves training a smaller "student" model to mimic the behavior of a larger "teacher" model, effectively transferring the knowledge from the larger dataset to the distilled one. Similarly, most distillation methods are evaluated on natural image datasets and their effectiveness lack comprehensive verification on medical datasets. To the best of our knowledge, a recent work [59] propose a comprehensive benchmark to evaluate the medical image dataset distillation.

## 6 LIMITATION & FUTURE WORK

While our EVA coreset selection strategy demonstrates superior performance in high compression scenarios, as evidenced by the comparative analysis presented in Tab. 1, it's important to acknowledge the limitations that the level of accuracy achieved in scenarios demanding extreme compression may not fully meet the rigorous standards necessary for medical diagnostics. Medical imaging tasks often require the highest degree of precision due to their direct impact on patient care, and there remains room for improvement in ensuring that the selected coresets are not only statistically representative but also clinically relevant.

Additionally, our current approach does not incorporate data from different modalities, which is essential in smart healthcare diagnostic systems. Such systems typically combine various types of data, including medical images, electronic health records (EHRs), patient interview descriptions, and pathology reports, for holistic analysis to enhance diagnostic accuracy.

Future research could focus on exploring different compression limits for various datasets to find the optimal balance between accuracy and efficiency. This would involve systematically determining how much data can be pruned while still maintaining sufficient performance levels for clinical applications. Moreover, there is a promising avenue to extend our work by integrating multimodal data, which would align well with the ongoing trends in applying large language models (LLMs) and other advanced AI techniques in healthcare. Such integration could enhance the robustness and applicability of our coreset selection strategy, particularly in systems where diverse types of data need to be synthesized for effective decision-making.

## 7 CONCLUSION

In this paper, we identify the limitations of existing coreset selection methods in capturing the evolutionary nature of model training and the fluctuations in sample importance within medical image datasets. To address this challenge, we introduced a novel sample scoring strategy, Evolution-aware VAriance (EVA), which incorporates a dual-window method to consider the training dynamics at different stages and employs a variance measurement of samples' error vectors for a more precise evaluation of sample importance. Extensive evaluations on various datasets and networks demonstrate the superior performance of our proposed EVA strategy.

## REFERENCES

[1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671* (2019).

[2] MohammadHossein Bateni, Aditya Bhaskara, Silvio Lattanzi, and Vahab Mirrokni. 2014. Distributed balanced clustering via mapping coresets. *Advances in Neural Information Processing Systems* 27 (2014).

[3] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. 2023. The liver tumor segmentation benchmark (lits). *Medical Image Analysis* 84 (2023), 102680.

[4] Antoine Bordes, Seyda Ertekin, Jason Weston, Léon Botton, and Nello Cristianini. 2005. Fast kernel classifiers with online and active learning. *Journal of machine learning research* 6, 9 (2005).

[5] Zalán Borsos, Marco Tagliasacchi, and Andreas Krause. 2021. Semi-supervised batch active learning via bilevel optimization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3495–3499.

[6] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. 2022. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4750–4759.

[7] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems* 30 (2017).

[8] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. 2018. DRINet for medical image segmentation. *IEEE transactions on medical imaging* 37, 11 (2018), 2453–2462.

[9] Jun Cheng. 2017. Brain tumor dataset. *figshare. Dataset* 1512427, 5 (2017).

[10] Rachit Chhaya, Anirban Dasgupta, and Supratim Shit. 2020. On coresets for regularized regression. In *International conference on machine learning*. PMLR, 1866–1876.

[11] 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).

[12] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. 2022. DC-BENCH: Dataset condensation benchmark. *Advances in Neural Information Processing Systems* 35 (2022), 810–822.

[13] Melanie Ducoffe and Frederic Precioso. 2018. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841* (2018).

[14] Zeyad Ali Sami Emam, Hong-Min Chu, Ping-Yeh Chiang, Wojciech Czaja, Richard Leapman, Micah Goldblum, and Tom Goldstein. 2021. Active learning at the imagenet scale. *arXiv preprint arXiv:2111.12880* (2021).

[15] Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems* 33 (2020), 2881–2891.

[16] Chengcheng Guo, Bo Zhao, and Yanbing Bai. 2022. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*. Springer, 181–195.

[17] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems* 31 (2018).

[18] 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.

[19] Xijie Huang, Zechun Liu, Shih-Yang Liu, and Kwang-Ting Cheng. 2023. Efficient quantization-aware training with adaptive coreset selection. *arXiv preprint arXiv:2306.07215* (2023).

[20] Angelos Katharopoulos and François Fleuret. 2018. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*. PMLR, 2525–2534.

[21] Vishal Kaushal, Suraj Kothawade, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. 2021. Prism: A unified framework of parameterized submodular information measures for targeted data subset selection and summarization. *arXiv preprint arXiv:2103.00128* (2021).

[22] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. 2017. Deep learning applications in medical image analysis. *Ieee Access* 6 (2017), 9375–9389.

[23] Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. 2021. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*. PMLR, 5464–5474.

[24] Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. 2021. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8110–8118.

[25] Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, and Rishabh Iyer. 2021. Retrieve: Coreset selection for efficient and robust semi-supervised learning. *Advances in neural information processing systems* 34 (2021), 14488–14501.

[26] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. 2022. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning*. PMLR, 11102–11118.

[27] Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. 2021. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems* 34 (2021), 18685–18697.

[28] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[29] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.

[30] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. 2022. Dataset distillation via factorization. *Advances in neural information processing systems* 35 (2022), 1100–1113.

[31] Ilya Loshchilov and Frank Hutter. 2015. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343* (2015).

[32] Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. *arXiv preprint arXiv:2109.03764* (2021).

[33] Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564* (2023).

[34] Kristof Meding, Luca M Schulze Buschoff, Robert Geirhos, and Felix A Wichmann. 2021. Trivial or impossible–dichotomous data difficulty masks model differences (on ImageNet and beyond). *arXiv preprint arXiv:2110.05922* (2021).

[35] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. 2020. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*. PMLR, 6950–6960.

[36] Andreas S Panayides, Amir Amini, Nenad D Filipovic, Ashish Sharma, Sotirios A Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, et al. 2020. AI in medical imaging informatics: current challenges and future directions. *IEEE journal of biomedical and health informatics* 24, 7 (2020), 1837–1857.

[37] Dongmin Park, Seola Choi, Doyoung Kim, Hwanjun Song, and Jae-Gil Lee. 2024. Robust data pruning under label noise via maximizing re-labeling accuracy. *Advances in Neural Information Processing Systems* 36 (2024).

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[39] 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.

[40] Mansheej Paul, Brett Larsen, Surya Ganguli, Jonathan Frankle, and Gintare Karolina Dziugaite. 2022. Lottery tickets on a data diet: Finding initializations with sparse trainable networks. *Advances in Neural Information Processing Systems* 35 (2022), 18916–18928.

[41] 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.

[42] 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.

[43] Vaibhav Saraf, Pallavi Chavan, and Ashish Jadhav. 2020. Deep learning challenges in medical imaging. In *Advanced Computing Technologies and Applications: Proceedings of 2nd International Conference on Advanced Computing Technologies and Applications—ICACTA 2020*. Springer, 293–301.

[44] Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489* (2017).

[45] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[46] Yang Song, Weidong Cai, Heng Huang, Yun Zhou, David Dagan Feng, Yue Wang, Michael J Fulham, and Mei Chen. 2015. Large margin local estimate with applications to medical image classification. *IEEE transactions on medical imaging* 34, 6 (2015), 1362–1377.

[47] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems* 35 (2022), 19523–19536.

[48] 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).

[49] Kai Wei, Rishabh Iyer, and Jeff Bilmes. 2015. Submodularity in data subset selection and active learning. In *International conference on machine learning*. PMLR, 1954–1963.

[50] Max Welling. 2009. Herding dynamical weights to learn. In *Proceedings of the 26th annual international conference on machine learning*. 1121–1128.

[51] Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. 2022. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *The Eleventh International Conference on Learning Representations*.

[52] Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai. 2019. Efficient multiple organ localization in CT image using 3D region proposal network. *IEEE transactions on medical imaging* 38, 8 (2019), 1885–1898.

[53] Jiancheng Yang, Rui Shi, and Bingbing Ni. 2021. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 191–195.

[54] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data* 10, 1 (2023), 41.

[55] S Yang, Z Xie, H Peng, M Xu, M Sun, and P Li. [n. d.]. Dataset pruning: Reducing training data by examining generalization influence. arXiv 2022. *arXiv preprint arXiv:2205.09329* ([n. d.]).
[56] Yu Yang, Hao Kang, and Baharan Mirzasoleiman. 2023. Towards sustainable learning: Coresets for data-efficient deep learning. In *International Conference on Machine Learning*. PMLR, 39314–39330.
[57] Jaehong Yoon, Divyam Madaan, Eunho Yang, and Sung Ju Hwang. 2021. Online coreset selection for rehearsal-based continual learning. *arXiv preprint arXiv:2106.01085* (2021).
[58] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems* 27 (2014).
[59] Zhen Yu, Yang Liu, and Qingchao Chen. 2024. Progressive trajectory matching for medical dataset distillation. *arXiv preprint arXiv:2403.13469* (2024).
[60] Xin Zhang, Jiawei Du, Yunsong Li, Weiying Xie, and Joey Tianyi Zhou. 2023. Spanning Training Progress: Temporal Dual-Depth Scoring (TDDS) for Enhanced Dataset Pruning. *arXiv preprint arXiv:2311.13613* (2023).
[61] Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. 2022. Coverage-centric coreset selection for high pruning rates. *arXiv preprint arXiv:2210.15809* (2022).
[62] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. 2021. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE* 109, 5 (2021), 820–838.
[63] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 3–11.