Hierarchical Dataset Selection for High-Quality Data Sharing

Xiaona Zhou[1], **Yingyan Zeng**[2], **Ran Jin**[3], **Ismini Lourentzou**[1],
[1]University of Illinois Urbana-Champaign
[2]University of Cincinnati
[3]Virginia Polytechnic Institute and State University
xiaonaz2@illinois.edu, zengyy@ucmail.uc.edu, jran5@vt.edu, lourent2@illinois.edu

Abstract

The success of modern machine learning hinges on access to high-quality training data. In many real-world scenarios, such as acquiring data from public repositories or sharing across institutions, data is naturally organized into discrete datasets that vary in relevance, quality, and utility. Selecting which repositories or institutions to search for useful datasets, and which datasets to incorporate into model training, are therefore critical decisions, yet most existing methods select individual samples and treat all data as equally relevant, ignoring differences between datasets and their sources. In this work, we formalize the task of dataset selection: selecting entire datasets from a large, heterogeneous pool to improve downstream performance under resource constraints. We propose Dataset Selection via Hierarchies (**DaSH**), a dataset selection method that models utility at both dataset and group levels (*e.g.*, collections, institutions), enabling efficient generalization from limited observations. Across two public benchmarks (DIGIT-FIVE and DOMAINNET), DaSH outperforms state-of-the-art data selection baselines by up to 26.2% in accuracy, while requiring significantly fewer exploration steps. Ablations show DaSH is robust to low-resource settings and lack of relevant datasets, making it suitable for scalable and adaptive dataset selection in practical multi-source learning workflows.

Project Page — https://plan-lab.github.io/projects/dash

1 Introduction

Deep learning models have achieved impressive performance across a wide range of supervised learning tasks, largely due to their ability to leverage large, high-quality datasets (Alzubaidi et al. 2023; Sun et al. 2017; Mohammed et al. 2025). In many real-world scenarios, however, available data is distributed across multiple heterogeneous sources, such as publicly available dataset repositories or collaborating institutions, with varying degrees of relevance to a target task. A key challenge in such settings is determining which external datasets, if any, can meaningfully improve model performance (Zhou et al. 2022; Zhang et al. 2022).

While practitioners often rely on intuition, domain expertise, or coarse metadata to guide dataset selection, there is little formal understanding of how to model such decisions algorithmically. Most existing approaches to data selection, *e.g.*, active learning (Sener and Savarese 2018; Gal, Islam, and Ghahramani 2017; Christen, Christen, and Rahm 2020; Paul, Bappy, and Roy-Chowdhury 2017; Zeng, Chen, and Jin 2023), data valuation (Ghorbani and Zou 2019; Pandl et al. 2021; Tang et al. 2021; Schoch, Xu, and Ji 2022; Kwon and Zou 2022), *etc.* , operate at the instance level, selecting individual data samples and assuming that all datasets and data sources in the selection pool are uniformly relevant to the task. This assumption fails in multi-source settings, where data is naturally organized into datasets and repositories that vary in relevance, redundancy, and quality. In practice, datasets are typically acquired, licensed, or shared in discrete units, and often originate from common sources such as institutions, simulation pipelines, or web-scale repositories, which induce a hierarchical structure over the dataset pool.

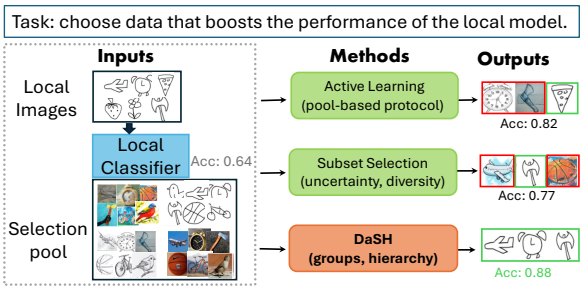


Figure 1: Dataset selection aims to select entire datasets from external sources to improve local model performance. Instance-level methods, such as active learning and subset selection, ignore dataset structure and often select irrelevant or misleading samples. In contrast, **DaSH** leverages hierarchical grouping to efficiently identify relevant datasets, avoiding noisy sources and achieving higher downstream accuracy.

To address this gap, in this work, we formalize the task of *dataset selection*: given a pool of datasets with unknown relevance to a target task, how can we efficiently identify a subset of datasets that will improve model performance, without having to exhaustively evaluate all candidates? This setting, illustrated in Figure 1, reflects many real-world constraints, where data is acquired, licensed, or shared in dataset-level units and must be selected under resource, bandwidth, or labeling constraints from multiple sources such as web-scale

repositories or partnering institutions.

To solve this new task, we propose Dataset Selection via Hierarchies (**DaSH**), a hierarchical Bayesian method that models dataset utility at both the group and dataset levels. Given a large pool of candidate datasets, grouped based on dataset origin (*e.g.*, institution or collection), DaSH performs structured exploration to infer both group-level relevance and individual dataset utility via posterior inference over observed model performance. This hierarchical modeling allows DaSH to prioritize informative groups and avoid wasted evaluation on unrelated or harmful sources. Experiments on two benchmarks demonstrate DaSH significantly outperforms state-of-the-art baselines by up to 26.2% in accuracy under low-resource settings. The contributions of this work are:

(1) We formalize the task of dataset selection from a heterogeneous pool of external datasets, a setting common in real-world workflows such as public data acquisition and cross-institutional collaboration, where data is organized into discrete, variably relevant sources.

(2) We propose DaSH, the first dataset selection method that models dataset utility through hierarchical inference over groups and datasets, enabling efficient and robust selection under limited feedback.

(3) We benchmark DaSH against four state-of-the-art data selection methods across two public datasets, demonstrating consistent performance gains, improving accuracy by up to 26.2% DIGIT-FIVE and 10.8% on DOMAINNET. Ablation studies show DaSH remains robust to grouping noise and scales effectively to large dataset pools, whereas existing methods frequently select irrelevant or low-utility data samples.

2 Related Work

Data Selection. Improving model performance through strategic data selection has been extensively explored across various paradigms. In active learning, methods aim to minimize labeling costs by iteratively selecting the most informative unlabeled instances (Sener and Savarese 2018; Gal, Islam, and Ghahramani 2017; Christen, Christen, and Rahm 2020; Paul, Bappy, and Roy-Chowdhury 2017; Zeng, Chen, and Jin 2023; Wang et al. 2023; Coleman et al. 2020). Batch active learning extends this by selecting diverse subsets in each iteration to improve efficiency (Kirsch, Van Amersfoort, and Gal 2019; Kaushal et al. 2018). Beyond active learning, data valuation techniques assess the contribution of individual points to model performance. Approaches like Data Shapley (Ghorbani and Zou 2019) and its adaptations (Pandl et al. 2021; Tang et al. 2021; Schoch, Xu, and Ji 2022; Kwon and Zou 2022; Liu et al. 2023; Courtnage and Smirnov 2021; Wang and Jia 2023; Just et al. 2023; Yoon, Arik, and Pfister 2020; Kwon and Zou 2023) quantify data utility, guiding the selection of valuable training instances. Additionally, subset selection methods (Killamsetty et al. 2021; Coleman et al. 2020) focus on constructing representative subsets to expedite learning without compromising accuracy.

However, existing methods largely operate at the instance level and overlook the hierarchical structure often present in real-world settings, where datasets are naturally grouped into repositories, *e.g.*, by source or collection. In contrast, DaSH targets dataset selection, *i.e.*, identify groups of datasets that jointly maximize downstream performance. Empirical results demonstrate that incorporating hierarchical information improves selection efficiency and model robustness.

Hierarchical Bandits. Hierarchical bandit algorithms address decision-making problems where actions are structured in a hierarchy, enabling efficient exploration and exploitation across multiple levels (Hong et al. 2022; Munos et al. 2014). In recommendation systems, hierarchical bandits have been employed to model user preferences (Yue, Hong, and Guestrin 2012) and item categories (Wang et al. 2018; Zuo et al. 2022), enabling personalized content delivery under resource constraints through adaptive frameworks (Yang et al. 2020; Santana et al. 2020). Beyond recommendation, hierarchical bandits have been applied to intelligent tutoring, decentralized reinforcement learning, and multi-task off-policy learning (Castleman, Macar, and Salleb-Aouissi 2024; Hong et al. 2023; Kao, Wei, and Subramanian 2022). These applications highlight the flexibility of hierarchical formulations in structuring complex decision processes across domains. Concurrently, theoretical advancements have focused on regret minimization and generalization across tasks using hierarchical Bayesian models (Kveton et al. 2021; Hong et al. 2022; Guan and Xiong 2024), offering principled frameworks for exploration under structured priors. Inspired by works in this space, our method tackles the unique setting of dataset selection by introducing a hierarchical Bayesian formulation that propagates dataset utility estimates across groups, enabling efficient amortization of training feedback via structured priors, and improving robustness to irrelevant or redundant datasets. To our knowledge, this is the first work to employ hierarchical bandits for dataset selection, with empirical evidence showing large gains in both accuracy and efficiency over non-hierarchical alternatives.

3 Method

Problem Definition. Consider n data groups $\mathbf{g} = \{g_1, g_2, \ldots, g_n\} = \{g_i\}_{i=1}^{n}$, where each group g_i contains one or more datasets. Let the set of datasets in group g_i be denoted $\mathbf{d}_i = \{d_{i,j}\}_{j=1}^{m_i}$, where $d_{i,j}$ is the j-th dataset in group i. Each dataset may contain an arbitrary number of data points. The full dataset pool is thus $\mathcal{D} = \bigcup_{i=1}^{n} \mathbf{d}_i = \{d_{i,j}\}$. Given a local model M_k, the goal is to select a subset $\tilde{\mathcal{D}}_k \subseteq \mathcal{D}$ from external sources that maximizes the performance gain over training on the local data d_k alone. Formally, we define:

$$\Delta\text{Acc}_k = \max_{\tilde{\mathcal{D}}_k \subseteq \mathcal{D}} \left(\text{Acc}(M_k, \tilde{\mathcal{D}}_k) - \text{Acc}(M_k, d_k) \right), \quad (1)$$

where $Acc(M_k, d_k)$ is the performance of local model M_k trained on local data d_k, $Acc(M_k, \tilde{\mathcal{D}}_k)$ is the performance of M_k after training on selected datasets $\tilde{\mathcal{D}}_k$, and ΔAcc_k is the performance gain for model $\mathcal{M}_k$.

DaSH Initialization

To address this selection objective, we introduce DaSH, a bi-level hierarchical Bayesian model that captures structured uncertainty across data groups and individual datasets. As depicted in Figure 2, each data group g_i is modeled with a latent

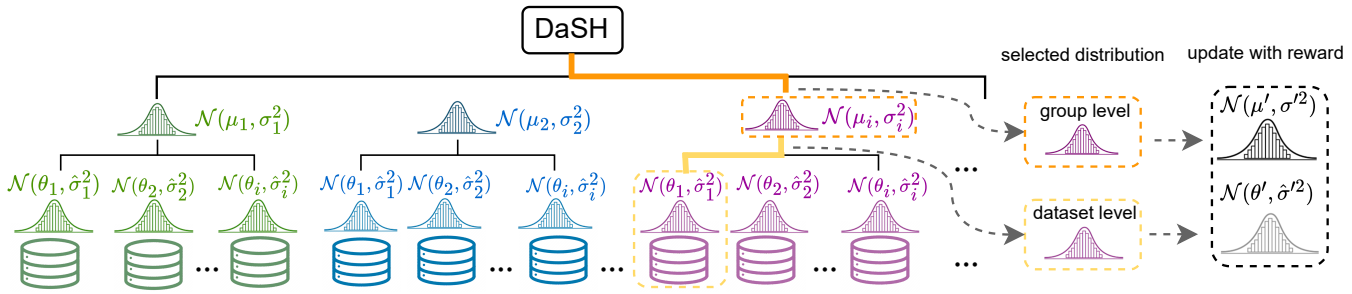


Figure 2: Overview of the DaSH dataset selection method. Each dataset and its corresponding group are modeled using Gaussian distributions $\mathcal{N}(\theta_i, \hat{\sigma}_i^2)$ and $\mathcal{N}(\mu_i, \sigma_i^2)$ for datasets and dataset groups, respectively. The selection process involves choosing a dataset group, followed by a specific dataset within that group. Upon receiving a reward, the posterior distributions for the dataset and the dataset group are updated to $\mathcal{N}(\mu', \sigma'^2)$ and $\mathcal{N}(\theta', \hat{\sigma}'^2)$ respectively. After training, dataset groups and datasets with higher posterior means are selected as described in Section 3.

parameter θ_i encoding its expected utility, and each dataset $d_{i,j}$ is governed by a local parameter $\theta_{i,j}$, with corresponding reward observations $r_{i,j}(t)$ at timestep t. We assume normal distributions for both the priors and the reward models, with unknown means and fixed variances. Conditional on $\theta_{i,j}$, the reward $r_{i,j}(t)$ is independent of the group-level parameter θ_i. The generative process is:

$$\begin{aligned} \theta_i &\sim \mathcal{N}(\mu_i, \sigma_i^2), \forall i \in [n] \\ \theta_{i,j}|\theta_i &\sim \mathcal{N}(\theta_i, \hat{\sigma}_i^2), \forall j \in [m] \\ r_{i,j}(t)|\theta_{i,j} &\sim \mathcal{N}(\theta_{i,j}, \sigma_r^2), \forall D(t) = d_{i,j}, \end{aligned} \tag{2}$$

where μ_i is the mean of the prior distribution for data group g_i, σ_i^2 is the variance of the group prior, $\hat{\sigma}_i^2$ is the variance of the dataset prior $\theta_{i,j}$, and σ_r^2 is the variance of the reward observation model. The goal is to iteratively update the posterior distribution of θ_i and $\theta_{i,j}$ by incorporating all observed reward values accumulated up to the current time step t. Through this continual update process, DaSH converges towards accurate estimations of the true distributions for both θ_i and $\theta_{i,j}$ after a number of iterations, as described in Algorithm 1 in the Appendix. Initialization begins with all dataset groups $\mathbf{g}$ sharing a common prior $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{N}(\theta_0, \hat{\sigma}_0^2)$. At each time step t, $\hat{\theta}_i$ is drawn from the normal distributions associated with each dataset group $\hat{\theta}_i \sim P(\theta_i|r_i)$ and the dataset group g_i with the largest value is chosen. Given dataset group selection g_i, DaSH then draws $\hat{\theta}_{i,j}$ from the distributions associated with the datasets within the chosen dataset group, *i.e.*, $\hat{\theta}_{i,j} \sim P(\theta_{i,j}|r_{i,j})$, and selects the dataset with the largest values, denoted as $D(t) = d_{i,j}$.

DaSH Posterior Computation

DaSH receives a reward from the chosen dataset and updates the distribution associated with the chosen dataset group and dataset using Eqs. (4) and (7). The posterior distribution of θ_i after observing reward values $r_i = \{r_{i,j}\}, j \in [m]$, where $r_{i,j} = \{r_{i,j}(t), \forall D(t) = d_{i,j}\}$, is given by:

$$\int_{\theta_{i,j}} \left(\prod_{j=1}^{m} \mathcal{N}(r_{i,j}; \theta_{i,j}, \sigma_r^2) \right) \mathcal{N}(\theta_{i,j}; \theta_i, \hat{\sigma}_i^2) d\theta_{i,j} \mathcal{N}(\theta_i; \mu_i, \sigma_i^2). \tag{3}$$

From Eq.(3), this yields the closed-form posterior:

$$P(\theta_i|r_i) = \mathcal{N}\left(\lambda_i^2 \left(\frac{\mu_i}{\sigma_i^2} + \frac{\bar{s}_i}{\hat{\sigma_i}^2 + \frac{\sigma_r^2}{n_i}} \right), \lambda_i^2 \right) \tag{4}$$

where

$$\lambda_i^2 = \left(\frac{1}{\sigma_i^2} + \frac{1}{\hat{\sigma_i}^2 + \frac{\sigma_r^2}{n_i}} \right)^{-1}, \quad \bar{s}_i = \frac{\sum_{j=1}^{m} r_{i,j}}{n_i}. \tag{5}$$

Here, n_i is the total number of selections for group g_i, and $\bar{s}_i$ is the aggregated mean reward across datasets in group i. The posterior mean is a precision-weighted average of the prior mean μ_i and the empirical group mean $\bar{s}_i$. The influence of the prior decays with more observations as λ_i^2 decreases. Since the reward $r_{i,j}(t)$ is conditionally independent of the data group parameter θ_i, the posterior density of $\theta_{i,j}$, after observing rewards $r_{i,j}(t)$ at time step t, is computed by:

$$P(\theta_{i,j} \mid r_{i,j}) \propto P(\theta_{i,j}) \prod_{t:D(t)=d_{i,j}} \mathcal{N}(r_{i,j}(t); \theta_{i,j}, \sigma_r^2), \tag{6}$$

resulting in the posterior:

$$P(\theta_{i,j} \mid r_{i,j}) = \mathcal{N}\left(\lambda_{i,j}^2 \left(\frac{\theta_i}{\hat{\sigma}_i^2} + \frac{\bar{s}_{i,j} \cdot n_{i,j}}{\sigma_r^2} \right), \lambda_{i,j}^2 \right) \tag{7}$$

where

$$\lambda_{i,j}^2 = \left(\frac{1}{\hat{\sigma}_i^2} + \frac{n_{i,j}}{\sigma_r^2} \right)^{-1}, \quad \bar{s}_{i,j} = \frac{r_{i,j}}{n_{i,j}} \tag{8}$$

Here, $n_{i,j}$ is the number of times dataset $d_{i,j}$ has been selected and $\bar{s}_{i,j}$ empirical mean of $r_{i,j}$.

Different from the dataset group posterior, the dataset posterior only depends on the rewards received by the dataset. Similar to the dataset group prior mean μ_i, θ_i is a bias term that influences the decay of the dataset posterior mean. As $n_{i,j} \to \infty$, the dataset posterior variance goes to zero, and the dataset posterior mean approaches $\bar{s}_{i,j}$.

Dataset Selection Based on Posterior Distributions

We formalize dataset selection using posterior means in a two-step process: first selecting a dataset group, then a dataset

within that group. A dataset or group is selected if its posterior mean μ exceeds a percentile-based threshold, *i.e.*, if $\mu > F^{-1}(x)$, where F^{-1} is the inverse cumulative distribution function (CDF) over the posterior means, setting the threshold at the x-th percentile. The selection threshold x is adaptively chosen based on the specific needs and constraints of the training environment. For example, a high percentile (*e.g.*, 90th) indicates a stringent criterion, suitable for scenarios with high training costs or where poor data quality significantly impacts model performance. Conversely, a lower percentile may be used in exploratory settings or when additional data inclusion costs are minimal. Alternatively, based on the use case, the selection of top-x datasets or dataset groups may be more appropriate.

Algorithmic Complexity

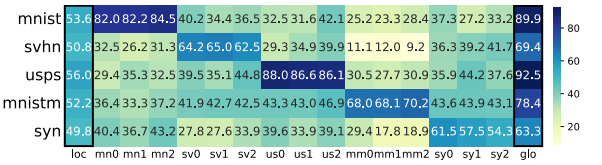
At each selection step, DaSH performs two sequential operations: (1) inter-group sampling by drawing $\hat{\theta}_i \sim P(\theta_i \mid r_i)$ for all n groups, and (2) intra-group sampling by drawing $\hat{\theta}_{i,j} \sim P(\theta_{i,j} \mid r_{i,j})$ for the m_i datasets in the chosen group. This yields a per-step computational cost of $O(n + m_i)$. Posterior updates for the chosen dataset and group require constant time per step, as the closed-form updates in Eqs. (4) and (7) avoid iterative optimization.

By contrast, a flat selection strategy must evaluate all $|D| = \sum_{i=1}^{n} m_i$ datasets at each step, incurring $O(|D|)$ cost. When groups are large, the hierarchical formulation amortizes exploration: feedback from a single dataset selection updates both its dataset-level and group-level posteriors, effectively sharing information across datasets in the same group. This reduces the total number of dataset evaluations required to achieve a fixed target accuracy, as consistently demonstrated in our experiments.

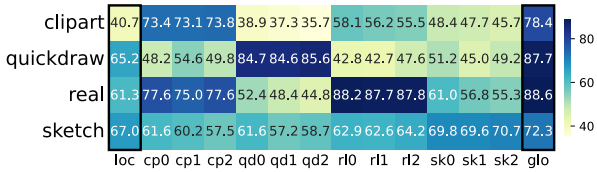
4 Experiments

Datasets. We validate DaSH on two widely used benchmarks in domain adaptation: **DIGIT-FIVE** and **DOMAINNET** (Peng et al. 2019). Each dataset contains multiple domain-specific subsets for a shared classification task. DIGIT-FIVE includes digit images from five domains (MNIST, MNIST-M, USPS, SVHN, and SYN), while DOMAINNET comprises object recognition images across different styles (CLIPART, QUICKDRAW, REAL, and SKETCH). Each domain is divided into three disjoint subsets to simulate distributed or federated settings. We use preprocessed versions of these datasets from Schrod et al. (2023), where fixed-size feature vectors are extracted from images for training and evaluation.

To evaluate the robustness of DaSH across varying dataset compositions, we examine two grouping strategies. In the **perfect group** setting, each group contains three subsets from the same domain (*e.g.*, mn0, mn1, mn2 from MNIST), modeling cases where repositories or institutions curate domain-specific datasets. In the **mixed group** setting, subsets from different domains are combined into groups (*e.g.*, mn1, mn2, mm0), modeling cases where datasets from multiple sources or domains are aggregated for a shared task and group assignments are noisy or imperfect. Preprocessing steps, group definitions, and dataset statistics are provided in the Appendix.



(a) DIGIT-FIVE



(b) DOMAINNET

Figure 3: Accuracy heatmaps of local classifiers after training on different DIGIT-FIVE and DOMAINNET subsets. The first column shows local test accuracy for each subset. The last column indicates the optimal accuracy achievable when training on all available relevant same-domain datasets. Middle columns depict accuracy after augmenting training data with additional subsets from same and different domains.

Implementation Details. For DIGIT-FIVE, each local model is a lightweight CNN trained on its respective domain-specific subsets (*e.g.*, MNIST, SVHN), while for DOMAINNET, local models are three-layer multilayer perceptrons (MLPs). Local accuracy refers to model performance on its own domain without any additional training. Additional implementation details are provided in the Appendix.

Figure 3 summarizes the empirical results obtained by training local models on different external datasets. These ground-truth results serve as a reference for evaluating the potential benefit of dataset selection. In DIGIT-FIVE, models trained on external datasets consistently underperform compared to their local baselines, indicating strong domain-specific bias. In contrast, DOMAINNET exhibits more favorable cross-domain transfer; for example, training the REAL classifier on subsets from CLIPART yields noticeable performance gains. This distinction underscores the practical relevance of dataset selection in heterogeneous sharing scenarios.

Baselines. We compare against existing methods to assess: (1) DaSH's effectiveness in dataset selection relative to state-of-the-art data selection approaches, and (2) its ability to capture dependencies among datasets.

Core-sets (Sener and Savarese 2018), which selects representative samples via geometric coverage, such that models learned only on the selected subset are as competitive.

FreeSel (Xie et al. 2023a), uses a pretrained vision transformer to perform one-pass, supervision-free data selection, with a time efficiency close to random selection.

ActiveFT (Xie et al. 2023b), which optimizes selection to match the data distribution while preserving diversity.

BiLAF (Lu et al. 2024), extends ActiveFT by introducing

Method	Hierarchical	MNIST	SVHN	USPS	MNIST-M	SYN	AVG
Local	✗	52.7±6.5	50.9±3.4	52.2±3.2	49.4±2.4	50.9±5.1	51.2±4.1
Global	✗	89.3±1.1	69.7±1.4	92.2±0.7	80.2±1.1	62.8±2.8	78.8±1.4
Core-sets (Sener and Savarese 2018)	✗	75.7±2.3 ↓13.8	52.8±2.7 ↓16.4	74.0±3.6 ↓17.2	60.8±2.1 ↓18.1	40.8±1.9 ↓22.1	60.8±2.5 ↓17.5
FreeSel (Xie et al. 2023a)	✗	87.6±1.2 ↓1.9	39.3±4.0 ↓29.9	29.3±3.1 ↓61.9	65.4±2.2 ↓13.5	40.7±2.9 ↓22.2	52.5±2.7 ↓25.8
ActiveFT (Xie et al. 2023b)	✗	58.2±1.6↓31.3	53.6±1.6↓15.6	59.2±1.3↓32.0	48.3±0.9↓30.6	41.4±1.5↓21.5	52.1±1.4↓26.2
BiLAF (Lu et al. 2024)	✗	62.6±0.5↓26.9	56.8±0.4↓12.4	67.3±0.5↓23.9	50.1±0.5↓28.8	52.6±1.0↓10.3	57.9±0.6↓20.4
DaSH	✓	**89.5±0.6**	**69.2±3.4**	**91.2±0.9**	**78.9±0.5**	**62.9±1.6**	**78.3±1.4**

Table 1: Performance comparison on DIGIT-FIVE against baselines (averaged over 5 runs) Best performance is **bold**. Red downward arrows (↓) indicate absolute drops in accuracy relative to the best-performing method.

Method	Hierarchical	CLIPART	QUICKDRAW	REAL	SKETCH	AVG
Local	✗	40.0±2.4	64.0±2.1	61.0±1.1	67.5±1.1	58.1±1.7
Global	✗	78.5±0.6	86.7±0.5	88.4±0.6	72.3±0.8	81.6±1.1
Core-sets (Sener and Savarese 2018)	✗	59.1±0.9 ↓18.2	74.1±0.3 ↓12.3	80.1±0.6 ↓8.3	67.6±0.4 ↓4.2	70.2±0.6 ↓10.8
FreeSel (Xie et al. 2023a)	✗	70.1±2.1 ↓7.2	81.7±0.8 ↓4.6	85.6±0.7 ↓2.8	67.2±1.3 ↓4.6	77.7±1.2 ↓3.3
ActiveFT (Xie et al. 2023b)	✗	67.6±1.8↓9.7	78.0±1.0↓8.3	83.8±1.1↓4.6	67.8±1.1↓4.0	74.3±1.3↓6.7
BiLAF (Lu et al. 2024)	✗	69.0±1.6↓8.3	81.3±0.5↓5.0	85.8±0.5↓2.6	67.8±0.7↓4.0	76.0±0.8↓5.0
DaSH	✓	**77.3±0.8**	**86.3±1.1**	**88.4±0.8**	**71.8±0.9**	**81.0±0.9**

Table 2: Performance comparison on DOMAINNET against baselines (averaged over 5 runs). Best performance is **bold**. Red downward arrows (↓) indicate absolute drops in accuracy relative to the best-performing method.

boundary uncertainty to enable one-shot label-free selection through pseudo-class estimation and iterative refinement.
In addition, we include two baselines for reference: **Local**, trained only on local data, and **Global**, trained on all datasets from the same domain, representing lower and upper bounds.

Experimental Results

Table 1 reports mean and standard deviation over five independent runs on DIGIT-FIVE subdomains, where we compare DaSH to local and global baselines as well as the four state-of-the-art data selection baselines. Across all five domains, DaSH matches the global model, achieving an average accuracy of 78.3%, which is only 0.5% below the global upper bound (78.8%) and significantly higher than the local lower bound (51.2%). These results indicate that our method is capable of effectively leveraging heterogeneous data sources.

Compared to competitive baselines, DaSH exhibits substantial gains. For instance, FreeSel underperforms by over 25.8% on average, and notably degrades performance on SVHN, USPS, and SYN, suggesting that its model-free selection policy does not work well under our problem setting where the selection pool contains irrelevant data. Similarly, ActiveFT and BiLAF fall behind by 26.2% and 20.4%, respectively. Notably, these methods exhibit particularly low accuracy on MNIST-M and SYN, which represent domains with significant distributional divergence from the rest of the datasets. This performance drop suggests that baselines struggle to generalize when the target domain is poorly aligned with the source distribution, highlighting their limitations in handling high domain shift scenarios. In contrast, DaSH consistently maintains top performance with low variance, highlighting its robustness across target domains.

Table 2 shows results on DOMAINNET. While performance margins are narrower than in DIGIT-FIVE, DaSH still outperforms all baselines by 3.3–10.8%. This is likely because all models use features extracted from a ResNet-18 backbone that was pretrained on the combined dataset. The shared feature extractor reduces the distributional differences between domains, making the task inherently easier for all methods and diminishing relative gains. Nevertheless, DaSH maintains its advantage, underscoring its effectiveness even when inter-domain variation is minimized.

5 Ablation Studies

To better understand the contributions of individual components in DaSH and the conditions under which it is most effective, we conduct a series of ablation studies. These experiments are designed to (1) isolate the effect of hierarchical modeling, (2) assess robustness to imperfect group definitions, (3) evaluate the role of Bayesian posterior updates, and (4) examine sensitivity to the exploration–exploitation trade-off. We also examine (5) the impact of selection granularity and (6) quantify efficiency gains from each design choice.

Impact of Hierarchical Grouping

To understand the importance of hierarchical grouping, we compare DaSH against two baseline variants: DaS (flat), a non-hierarchical counterpart, and DaSH (mixed), which uses imperfect group assignments. Figure 4 presents Pareto frontiers of accuracy versus selection cost (exploration steps) for each domain in DIGIT-FIVE and DOMAINNET, with marker shapes indicating domains and colors indicating methods. Compared to the non-hierarchical DaS (flat), DaSH consistently delivers equal or higher accuracy at substantially lower selection cost. On DIGIT-FIVE, this translates to savings of 20–60 steps per domain without sacrificing accuracy. When compared to DaSH (mixed), the gap is small in most domains, with the mixed variant often lying on or near the Pareto frontier achieved by perfect grouping. This indicates that DaSH

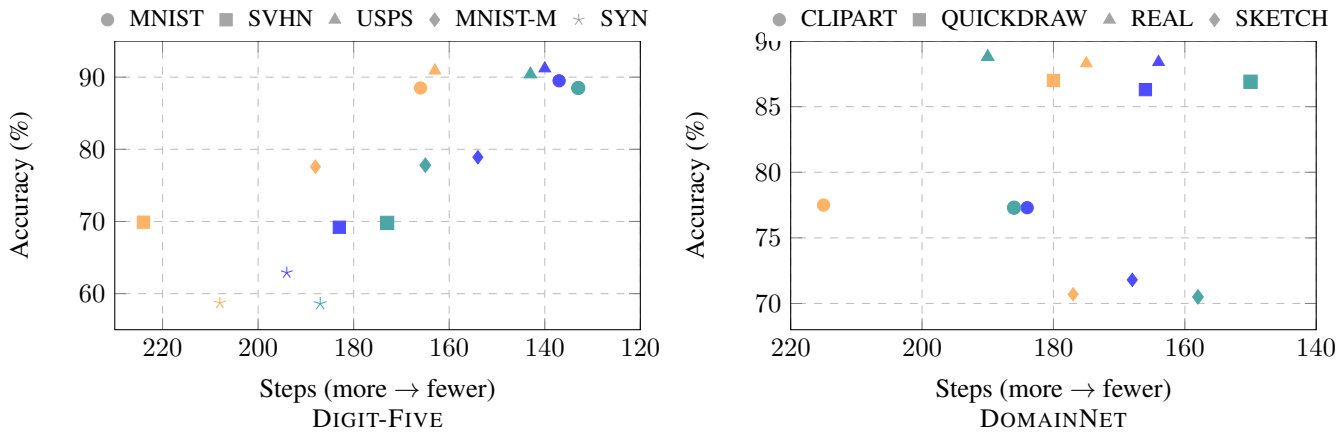


Figure 4: Pareto trade-offs between accuracy and selection cost. Each point is a method–domain result (DIGIT-FIVE left, DOMAINNET right). Marker shape encodes the domain, while color distinguishes the methods: **DaS (flat)**, **DaSH (mixed)**, and **DaSH**. Points toward the upper-right represent better trade-offs (higher accuracy, fewer steps). Across both benchmarks, the upper-right region is occupied by hierarchical variants with **DaSH** contributing most of the frontier on DIGIT-FIVE and sharing the frontier with **DaSH (mixed)** on DOMAINNET.

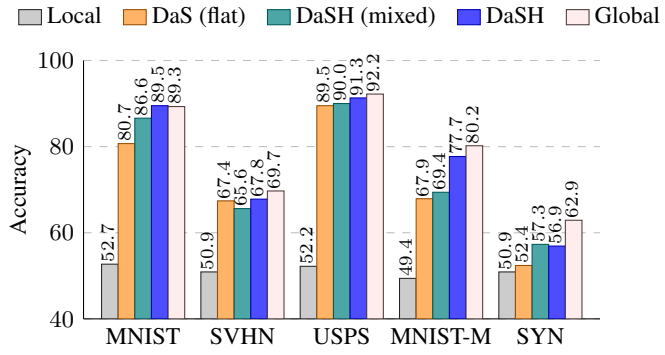


Figure 5: Performance under budget constraints. Under limited exploration (15 steps), DaSH and DaSH (mixed) outperform DaS (flat) on 4 out of 5 datasets. *Local* and *Global* denote the lower and upper bounds, respectively.

is robust to imperfect group assignments, with only modest performance drops in more challenging domains like SYN, QUICKDRAW, and REAL. Overall, these results show that hierarchical grouping not only improves efficiency and accuracy but also maintains strong performance under noisy or partially incorrect group structures.

Comparison Under Limited Exploration

We evaluate the ability of each method to identify useful datasets under stringent exploration budgets. Specifically, each method explores each dataset only once, totaling 15 steps across the 15 datasets in DIGIT-FIVE. Figure 5 reports the resulting accuracy for each domain. Under this extreme budget constraint, both DaSH and DaSH (mixed) outperform the non-hierarchical DaS (flat) in 4 out of 5 domains. The gains over DaS (flat) are substantial: +8.8% on MNIST, +1.8% on USPS, +9.8% on MNIST-M, and +4.5% on SYN. Even with imperfect grouping, DaSH (mixed) closely tracks the performance of perfect grouping, with accuracy differences within 1–2% in most domains. The Local and Global baselines show that hierarchical variants close more than half the gap to the global optimum despite operating under a 15-step budget. These results confirm that hierarchical grouping enables efficient, high-quality dataset selection even under severe exploration limits.

% Train	**10%**		**20%**		**50%**	
Name	**Init.**	**DaSH**	**Init.**	**DaSH**	**Init.**	**DaSH**
MNIST	17.6	31.5	23.6	89.6	36.6	89.6
SVHN	12.8	24.2	21.2	21.5	35.6	66.7
USPS	9.6	13.5	12.8	28.6	31.2	91.4
MNIST-M	20.6	55.1	28.8	57.6	44.2	79.3
SYN	26.6	37.6	21.4	24.9	27.4	41.0

Table 3: DaSH improves performance even with a weak initial model with low accuracy. This table reports accuracy on DIGIT-FIVE when initially trained on 10%, 20%, and 50% of the local training data (**Init.**), and after using DaSH to select additional datasets for training (**DaSH**).

Effectiveness Under Weak Initialization

We additionally investigate whether DaSH can enhance performance when initial local model accuracy is very low. We train initial local classifiers using 10%, 20%, and 50% of the available training data. Table 3 shows consistent accuracy gains across all conditions, even when initial accuracy is as low as 9.6% (USPS), demonstrating DaSH's robustness to significant variations in initial performance before selection.

Robustness under Cross-Domain Grouping

We evaluate DaSH in an extreme cross-domain grouping scenario, where each group is constructed to contain exactly one

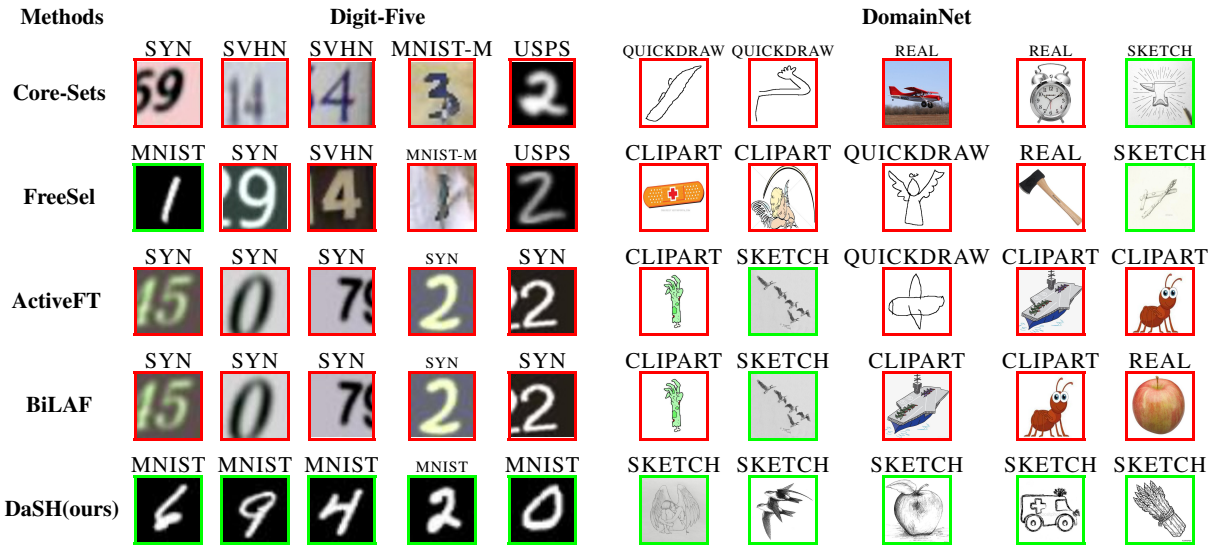


Figure 6: Qualitative comparisons on DIGIT-FIVE (target: MNIST) and DOMAINNET (target: SKETCH). Each selected image is labeled by its source domain (above), with green borders indicating a correct domain match to the target and red borders indicating a mismatch. Unlike prior methods, which frequently select subsets from mismatched domains in the first exploration step, DaSH consistently identifies subsets from the correct domain, even in challenging settings with visually similar categories.

Method	# Steps	Accuracy
DaS (flat)	163	90.9±2.0
DaSH	140	91.2±0.9
DaSH (cross-domain grouping)	154	92.2±0.7

Table 4: Robustness of DaSH under cross-domain grouping. Performance on USPS with cross-domain groups, where each group contains exactly one dataset from each domain, removing opportunities to select multiple same-domain datasets. DaSH achieves the robust accuracy while requiring fewer steps than the non-hierarchical variant DaS (flat).

dataset from each domain. This setup eliminates the possibility of selecting multiple same-domain datasets within a single group, stress-testing the ability of DaSH to perform effective selection when group structure does not align with domain semantics and offers no within-domain redundancy to exploit. As shown in Table 4, DaSH delivers robust accuracy and outperforms the non-hierarchical baseline, DaS (flat), while also requiring fewer selection steps. Our ablation results consistently show that, under different settings, DaSH remains effective, maintaining strong performance with minimal computational overhead.

6 Qualitative Analysis

Figure 6 illustrates clear qualitative differences in the selection behavior of each method. Green borders indicate that the selected data instance belongs to the target domain, while red borders indicate domain mismatches. Across both benchmarks, baseline methods such as Core-Sets, FreeSel, ActiveFT, and BiLAF often select subsets from visually similar but incorrect domains. For example, when MNIST is used as the local dataset, most baselines retrieve images that are visually distinct from the target domain. Only FreeSel selects a sample from MNIST, which is consistent with its relatively better quantitative performance (Table 1). The rest of the baselines fail to retrieve meaningful samples. In contrast, DaSH effectively selects relevant data. This behavior extends to DOMAINNET, where DaSH maintains domain-consistent selection across diverse categories. These results suggest that DaSH internalizes domain structure more effectively than prior methods, allowing it to identify relevant datasets even under distribution shift and candidate noise, an essential capability for transferability in collaborative data-sharing settings.

7 Conclusion

This work addresses a key bottleneck in machine learning: selecting training datasets from diverse sources such as institutions, repositories, or collections. We introduce DaSH, a dataset selection framework that models the hierarchical relationship among datasets and data sources to improve selection efficiency and downstream performance. Experimental results demonstrate that DaSH consistently outperforms non-hierarchical and existing instance-level data selection baselines, and remains robust under realistic constraints such as imperfect grouping and limited exploration budgets. These findings underscore the importance of effectively automating practical data curation as machine learning models increasingly depend on large-scale heterogeneous data sources from various online repositories. Future directions include incorporating multi-objective selection criteria such as utility, fairness, and domain coverage, and applying DaSH to large-scale, multi-institutional data sharing platforms, where group membership and dataset availability evolve over time.

Acknowledgments

This research is based on work partially supported by the National Science Foundation under award number CMMI-2331985, the U.S. Defense Advanced Research Projects Agency (DARPA) under award number HR001125C0303, and U.S. Army DEVCOM under award number W5170125CA160. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of NSF, DARPA, the U.S. Army, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Alzubaidi, L.; Bai, J.; Al-Sabaawi, A.; Santamaría, J.; Albahri, A. S.; Al-Dabbagh, B. S. N.; Fadhel, M. A.; Manoufali, M.; Zhang, J.; Al-Timemy, A. H.; et al. 2023. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*.

Castleman, B.; Macar, U.; and Salleb-Aouissi, A. 2024. Hierarchical Multi-Armed Bandits for the Concurrent Intelligent Tutoring of Concepts and Problems of Varying Difficulty Levels. In *Deployable RL: From Research to Practice@ Reinforcement Learning Conference*.

Christen, V.; Christen, P.; and Rahm, E. 2020. Informativeness-Based Active Learning for Entity Resolution. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD*. Springer.

Coleman, C.; Yeh, C.; Mussmann, S.; Mirzasoleiman, B.; Bailis, P.; Liang, P.; Leskovec, J.; and Zaharia, M. 2020. Selection via Proxy: Efficient Data Selection for Deep Learning. In *International Conference on Learning Representations*.

Courtnage, C.; and Smirnov, E. 2021. Shapley-value data valuation for semi-supervised learning. In *Discovery Science: 24th International Conference*. Springer.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep Bayesian Active Learning with Image Data. In *International Conference on Machine Learning*.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*.

Ghorbani, A.; and Zou, J. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *International Conference on Machine Learning*.

Guan, J.; and Xiong, H. 2024. Improved Bayes Regret Bounds for Multi-Task Hierarchical Bayesian Bandit Algorithms. *Advances in Neural Information Processing Systems (NeurIPS)*.

Hong, J.; Kveton, B.; Zaheer, M.; and Ghavamzadeh, M. 2022. Hierarchical Bayesian Bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Hong, J.; Kveton, B.; Zaheer, M.; Katariya, S.; and Ghavamzadeh, M. 2023. Multi-task off-policy learning from bandit feedback. In *International Conference on Machine Learning*. PMLR.

Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jin, X.; Lan, C.; Zeng, W.; and Chen, Z. 2021. Re-energizing domain discriminator with sample relabeling for adversarial domain adaptation. In *IEEE/CVF International Conference on Computer Vision*.

Just, H. A.; Kang, F.; Wang, T.; Zeng, Y.; Ko, M.; Jin, M.; and Jia, R. 2023. LAVA: Data Valuation Without Pre-Specified Learning Algorithms. In *International Conference on Learning Representations*.

Kao, H.; Wei, C.-Y.; and Subramanian, V. 2022. Decentralized cooperative reinforcement learning with hierarchical information structure. In *International Conference on Algorithmic Learning Theory*. PMLR.

Kaushal, V.; Sahoo, A.; Doctor, K.; Raju, N.; Shetty, S.; Singh, P.; Iyer, R.; and Ramakrishnan, G. 2018. Learning from Less Data: Diversified Subset Selection and Active Learning in Image Classification Tasks. *arXiv Preprint arXiv:1805.11191*.

Killamsetty, K.; Sivasubramanian, D.; Ramakrishnan, G.; and Iyer, R. 2021. Glister: Generalization based data subset selection for efficient and robust learning. In *AAAI Conference on Artificial Intelligence*.

Kirsch, A.; Van Amersfoort, J.; and Gal, Y. 2019. Batchbald: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Komatsu, T.; Matsui, T.; and Gao, J. 2021. Multi-source domain adaptation with sinkhorn barycenter. In *European Signal Processing Conference (EUSIPCO)*. IEEE.

Kveton, B.; Konobeev, M.; Zaheer, M.; Hsu, C.-w.; Mladenov, M.; Boutilier, C.; and Szepesvari, C. 2021. Meta-Thompson Sampling. In *International Conference on Machine Learning*.

Kwon, Y.; and Zou, J. 2022. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Kwon, Y.; and Zou, J. 2023. Data-OOB: Out-of-Bag Estimate as a Simple and Efficient Data Value. In *International Conference on Machine Learning*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *IEEE*.

Li, Y.; Yuan, L.; Chen, Y.; Wang, P.; and Vasconcelos, N. 2021. Dynamic transfer for multi-source domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, Z.; Just, H. A.; Chang, X.; Chen, X.; and Jia, R. 2023. 2D-shapley: a framework for fragmented data valuation. In *International Conference on Machine Learning*.

Lu, H.; Xie, Y.; Yang, X.; and Yan, J. 2024. Boundary Matters: A Bi-Level Active Finetuning Method. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Luo, S.; Zhu, D.; Li, Z.; and Wu, C. 2021. Ensemble federated adversarial training with non-iid data. *arXiv preprint arXiv:2110.14814*.

Mohammed, S.; Budach, L.; Feuerpfeil, M.; Ihde, N.; Nathansen, A.; Noack, N.; Patzlaff, H.; Naumann, F.; and Harmouch, H. 2025. The effects of data quality on machine learning performance. *Information Systems*.

Munos, R.; et al. 2014. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*.

Pandl, K. D.; Feiland, F.; Thiebes, S.; and Sunyaev, A. 2021. Trustworthy Machine Learning for Health Care: Scalable Data Valuation with the Shapley Value. In *Conference on Health, Inference, and Learning*.

Paul, S.; Bappy, J. H.; and Roy-Chowdhury, A. K. 2017. Non-Uniform Subset Selection for Active Learning in Structured Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *IEEE/CVF International Conference on Computer Vision*.

Roy, P.; Ghosh, S.; Bhattacharya, S.; and Pal, U. 1807. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*.

Santana, M. R.; Melo, L. C.; Camargo, F. H.; Brandão, B.; Soares, A.; Oliveira, R. M.; and Caetano, S. 2020. Contextual Meta-Bandit for Recommender Systems Selection. In *ACM Conference on Recommender Systems*.

Schoch, S.; Xu, H.; and Ji, Y. 2022. CS-Shapley: class-wise Shapley values for data valuation in classification. *Advances in Neural Information Processing Systems (NeurIPS)*.

Schrod, S.; Lippl, J.; Schäfer, A.; and Altenbuchinger, M. 2023. FACT: Federated Adversarial Cross Training. *arXiv preprint arXiv:2306.00607*.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.

Simon, C.; Faraki, M.; Tsai, Y.-H.; Yu, X.; Schulter, S.; Suh, Y.; Harandi, M.; and Chandraker, M. 2022. On generalizing beyond domains in cross-domain continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Singh, A. 2021. Clda: Contrastive learning for semi-supervised domain adaptation. *Advances in Neural Information Processing Systems (NeurIPS)*.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tang, S.; Ghorbani, A.; Yamashita, R.; Rehman, S.; Dunnmon, J. A.; Zou, J.; and Rubin, D. L. 2021. Data Valuation for Medical Imaging Using Shapley Value and Application to a Large-Scale Chest X-Ray Dataset. *Scientific Reports*.

Wang, J. T.; and Jia, R. 2023. Data Banzhaf: A Robust Data Valuation Framework for Machine Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Wang, L.; Wang, X.; Ji, Q.; Wang, L.; and Jin, R. 2023. Mutual Active Learning for Engineering Regulated Statistical Digital Twin Models. *IEEE Transactions on Industrial Informatics*.

Wang, Q.; Li, T.; Iyengar, S.; Shwartz, L.; and Grabarnik, G. Y. 2018. Online IT Ticket Automation Recommendation Using Hierarchical Multi-Armed Bandit Algorithms. In *SIAM International Conference on Data Mining*.

Xie, Y.; Ding, M.; Tomizuka, M.; and Zhan, W. 2023a. Towards free data selection with general-purpose models. *Advances in Neural Information Processing Systems (NeurIPS)*.

Xie, Y.; Lu, H.; Yan, J.; Yang, X.; Tomizuka, M.; and Zhan, W. 2023b. Active finetuning: Exploiting annotation budget in the pretraining-finetuning paradigm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yang, M.; Li, Q.; Qin, Z.; and Ye, J. 2020. Hierarchical Adaptive Contextual Bandits for Resource Constraint Based Recommendation. In *The Web Conference*.

Yao, C.-H.; Gong, B.; Qi, H.; Cui, Y.; Zhu, Y.; and Yang, M.-H. 2022. Federated multi-target domain adaptation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Yoon, J.; Arik, S.; and Pfister, T. 2020. Data Valuation Using Reinforcement Learning. In *International Conference on Machine Learning*.

Yue, Y.; Hong, S. A.; and Guestrin, C. 2012. Hierarchical Exploration for Accelerating Contextual Bandits. In *International Conference on Machine Learning*.

Zeng, Y.; Chen, X.; and Jin, R. 2023. Ensemble Active Learning by Contextual Bandits for AI Incubation in Manufacturing. *ACM Transactions on Intelligent Systems and Technology*.

Zhang, W.; Deng, L.; Zhang, L.; and Wu, D. 2022. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zuo, J.; Hu, S.; Yu, T.; Li, S.; Zhao, H.; and Joe-Wong, C. 2022. Hierarchical conversational preference elicitation with bandit feedback. In *ACM International Conference on Information & Knowledge Management*.