

Hierarchical Bayesian Dataset Selection

Xiaona Zhou

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Application

Ismini Lourentzou, Chair

Ran Jin

Christopher Lee Thomas

May 1, 2024

Blacksburg, Virginia

Keywords: Hierarchical Bayesian, Data-Sharing, Reinforcement Learning, Dataset Selection

Copyright 2024, Xiaona Zhou

Hierarchical Bayesian Dataset Selection

Xiaona Zhou

(ABSTRACT)

Despite the profound impact of deep learning across various domains, supervised model training critically depends on access to large, high-quality datasets, which are often challenging to identify. To address this, we introduce **Hierarchical Bayesian Dataset Selection (HBDS)**, the first dataset selection algorithm that utilizes hierarchical Bayesian modeling, designed for collaborative data-sharing ecosystems. The proposed method efficiently decomposes the contributions of dataset groups and individual datasets to local model performance using Bayesian updates with small data samples. Our experiments on two benchmark datasets demonstrate that HBDS not only offers a computationally lightweight solution but also enhances interpretability compared to existing data selection methods, by revealing deep insights into dataset interrelationships through learned posterior distributions. HBDS outperforms traditional non-hierarchical methods by correctly identifying all relevant datasets, achieving optimal accuracy with fewer computational steps, even when initial model accuracy is low. Specifically, HBDS surpasses its non-hierarchical counterpart by 1.8% on **DIGIT-FIVE** and 0.7% on **DOMAINNET**, on average. In settings with limited resources, HBDS achieves a 6.9% higher accuracy than its non-hierarchical counterpart. These results confirm HBDS’s effectiveness in identifying datasets that improve the accuracy and efficiency of deep learning models when collaborative data utilization is essential.

Hierarchical Bayesian Dataset Selection

Xiaona Zhou

(GENERAL AUDIENCE ABSTRACT)

Deep learning technologies have revolutionized many domains and applications, from voice recognition in smartphones to automated recommendations on streaming services. However, the success of these technologies heavily relies on having access to large and high-quality datasets. In many cases, selecting the right datasets can be a daunting challenge. To tackle this, we have developed a new method that can quickly figure out which datasets or groups of datasets contribute most to improving the performance of a model with only a small amount of data needed. Our tests prove that this method is not only effective and light on computation but also helps us understand better how different datasets relate to each other.

Dedication

This work is dedicated to Professor Boyan Kostadinov. Thank you for your support, encouragement, and advice since the beginning of my academic journey.

Acknowledgments

I would like to extend my deepest gratitude to my advisor, Dr. Ismini Lourentzou, for her invaluable guidance and unwavering support throughout the course of this thesis. Her expertise and insightful feedback have been pivotal in shaping this project. I am profoundly grateful for her mentorship and the opportunities I have been afforded under her guidance. Additionally, I would like to thank Dr. Ran Jin and his student Yingyan Zeng for their assistance and guidance on the completion of this project.

Contents

List of Figures	ix
List of Tables	xi
1 Introduction	1
1.1 Motivation	1
1.2 Thesis Contributions	3
2 Review of Literature	5
2.1 Data Selection	5
2.1.1 Active Learning	5
2.1.2 Data Valuation	6
2.1.3 Reinforcement Learning	6
2.1.4 Other Techniques	7
2.2 Hierarchical Bandit	8
2.2.1 Application in Recommendation Systems	8
2.2.2 Broad Applications Across Various Domains	9
2.2.3 Advanced Developments in Adaptive Exploration	9

3	Methodology	11
3.1	Overview	11
3.2	Problem Definition	11
3.3	HBDS Initialization	12
3.4	Posterior Distribution Computation	14
3.5	Dataset Selection Based on Posterior Distributions	15
4	Experiments and Results	17
4.1	Experiments	17
4.1.1	Datasets	17
4.1.2	Baselines	20
4.1.3	Implementation details	21
4.2	Experimental Results	23
4.3	Ablation Studies	26
4.3.1	Comparison Under Limited Exploration	27
4.3.2	Local Model with Low Initial Accuracy	28
4.3.3	Absence of Relevant Datasets	29
4.3.4	Presents of more datasets	29
4.4	Qualitative Analysis	30
5	Conclusions	33

List of Figures

1.1	In scenarios where there are abundant datasets, e.g., in a data-sharing ecosystem, the task of identifying beneficial datasets for model training is challenging. HBDS makes use of the hierarchical structure of the datasets and dataset groups and employs hierarchical Bayesian modeling to efficiently identify beneficial datasets.	2
3.1	Diagram of the HBDS 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. .	12
4.1	Heatmap illustrating the accuracies of local classifiers post-training on different DOMAINNET subsets. The first column displays local accuracies on test sets, while the last column represents the optimal accuracy achievable considering all available relevant same-domain datasets. The middle columns depict accuracies after training on additional relevant subsets from the same domain.	22
4.2	# Steps required on DIGIT-FIVE across baselines	25
4.3	# Steps required on DOMAINNET across baselines	25
4.4	Performance comparison under limited exploration settings	27

4.5	When there exists no relevant dataset available across all groups, HBDS posterior means remain low even after 600 exploration steps. This ensures HBDS can reliably discern irrelevant data.	29
4.6	Scalability comparison when more datasets across five data groups are present. HBDS achieves higher accuracy across all models with the presence of additional datasets.	30
4.7	The posterior means of BB, HBDS and HBDS (mixed).	31
4.8	Dataset and dataset group posterior means as determined by HBDS after training on all relevant available data. (a) Datasets within the same group exhibit consistently high posterior means, indicating their relevance to the local data. (b) Likewise, the highest values are assigned to groups containing the most relevant datasets.	32

List of Tables

4.1	Performance comparison on <code>DIGIT-FIVE</code> of HBDS against baselines (averaged over 5 runs) under perfect and mixed group settings. Best performance is highlighted in blue bold and second-best is bold	24
4.2	Performance comparison on <code>DOMAINNET</code> of HBDS against baselines (averaged over 5 runs) under perfect and mixed group settings. Best performance is highlighted in blue bold and second-best is bold	24
4.3	HBDS improves model performance even when the initial local model exhibits extremely low starting accuracy	28

Chapter 1

Introduction

1.1 Motivation

Deep learning models have achieved remarkable success in various supervised learning tasks due to their ability to process and learn from large volumes of data. However, the effectiveness of these models is contingent upon the availability of vast, high-quality datasets. In many industries, assembling such datasets is not only challenging but also cost-prohibitive, especially for smaller entities or niche markets. These challenges call for innovative approaches to data acquisition and utilization. The concept of a data-sharing ecosystem, where stakeholders share access to their datasets, presents a promising solution to address these challenges. In a data-sharing ecosystem, every participant benefits from the pooled data resources, reducing the individual burden of data collection and processing.

However, with the increasing availability of shared data, there is an emerging need for methods that can automatically select the most relevant datasets for specific tasks. The goal is to identify datasets that are most likely to improve model accuracy without incurring substantial computational costs. Effective dataset selection mechanisms become paramount, as indiscriminate data usage can lead to inefficient training and suboptimal model performance. In addition, in a data-sharing ecosystem participants may usually own multiple datasets that are often similar due to analogous production processes. These datasets naturally form groups, which correspond to the typical organizational structures found in industry

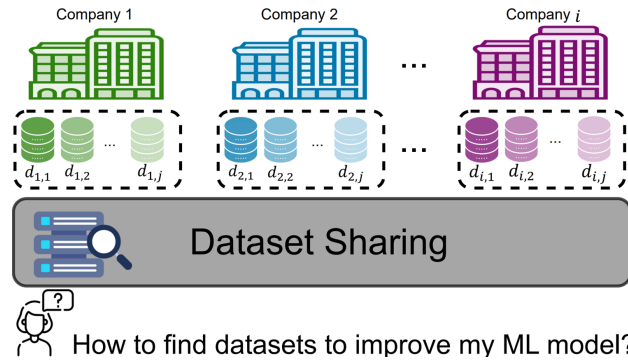


Figure 1.1: In scenarios where there are abundant datasets, e.g., in a data-sharing ecosystem, the task of identifying beneficial datasets for model training is challenging. HBDS makes use of the hierarchical structure of the datasets and dataset groups and employs hierarchical Bayesian modeling to efficiently identify beneficial datasets.

sectors such as manufacturing or digital services, where similar processes yield similar types of data. Within each group, datasets are interconnected and share characteristics, creating a hierarchical structure between the group and its individual datasets.

This hierarchical organization not only mirrors the complexity of real-world data interactions but also opens new directions for more effective data management and utilization strategies. Previous studies have aimed to optimize data utility for model training, employing strategies from active learning to algorithms that select high-value data subsets to improve model performance [3, 7, 8, 35, 42]. However, data selection methods operate independently of the inherent quality of data points within a dataset, as they prioritize the selection of individual points based on predefined criteria rather than comprehensive dataset quality assessment. On the other hand, dataset selection typically requires considering the collective characteristics, distribution, and quality of the entire dataset to ensure its suitability for a specific task or analysis. Relying solely on data point selection may overlook important aspects such as data coherence and representativeness of the dataset as a whole. Moreover, these methods often incur significant computational costs, rendering them impractical when dealing with a large number of data points. Moreover, the benefits derived from each selection are of-

ten minimal, underscoring the need for more scalable and computationally efficient dataset selection approaches.

To address these challenges, we introduce **Hierarchical Bayesian Dataset Selection (HBDS)**, a novel dataset selection technique that groups datasets based on similarities in their collection processes and anticipated usefulness, thus mitigating the computational overhead typically associated with data evaluation methods. HBDS leverages hierarchical Bayesian modeling to estimate the contribution of dataset groups and datasets to the training of a local model. We perform experiments with three baselines on two benchmark datasets under varying settings. Experimental results demonstrate that HBDS outperforms the baselines after just 15 steps, even with non-ideal dataset grouping, i.e., dissimilar datasets are grouped together. When training continues until reaching the natural stopping condition, HBDS consistently surpasses the non-hierarchical counterpart, requiring around 25 fewer training steps. Furthermore, when tested on local models with low initial accuracy, HBDS consistently improves model performance with selected datasets.

1.2 Thesis Contributions

We introduce a framework designed to optimize dataset selection effectively. Our approach utilizes hierarchical Bayesian modeling to infer the potential contributions of dataset groups and individual datasets to the performance improvement of a local model. The priors are updated using a single representative data point from the datasets at each step, ensuring that our framework remains computationally lightweight. The primary computational effort involves generating predictions from the local model, rather than updating the priors. Our experiments validate that considering the hierarchical structure between dataset groups and individual datasets significantly improves selection efficiency. Moreover, the learned

posterior distributions accurately capture the interrelationships among datasets, providing interpretability that existing data selection methods lack.

We summarize the key contributions as follows:

- (1) We introduce HBDS, the first algorithm dedicated to dataset selection utilizing hierarchical Bayesian modeling. HBDS is a hierarchical Bayesian Bandit method that leverages the hierarchical structure of dataset groups and individual datasets, facilitating more efficient selection compared to traditional non-hierarchical approaches.
- (2) Through rigorous testing on two benchmark datasets, we demonstrate that HBDS consistently outperforms traditional non-hierarchical counterparts, achieving optimal or near-optimal accuracy with fewer computational steps.
- (3) Ablation studies demonstrate that even when the initial accuracy of the local model is as low as 10%, HBDS consistently enhances performance, underscoring its robustness and adaptability. In scenarios lacking relevant datasets, the learned posterior means remain low after extensive exploration, effectively signaling the absence of useful data. Additionally, HBDS is scalable to a larger data-sharing ecosystem and maintains good performance. Qualitative analysis shows that the learned posterior distributions closely capture the relationship between datasets, data groups, and the local model.

Chapter 2

Review of Literature

2.1 Data Selection

Research in data selection has focused on improving the efficiency and quality of training data used in machine learning models. While these methods are effective in improving model performance through strategic data selection, they are not suitable for our task. Given the numerous datasets within a data-sharing ecosystem and the vast number of data points, any form of data selection becomes impractical due to the overwhelming computational demands.

2.1.1 Active Learning

Active learning strategies are employed to iteratively select unlabeled training examples for labeling, enhancing model performance with minimal labeling effort. This includes work by Sener and Savarese [42], who use core-set selection to select a subset of points such that a model trained on this subset performs well on the remaining data points. Fraga-Silva et al. [7] employ a variety of criteria to select data for labeling in the area of speech recognition. Gal et al. [8] propose combining Bayesian deep learning with active learning for selecting high-dimensional data. Christen et al. [3] utilize an informativeness measure to efficiently expand the training dataset without requiring extensive manual labeling, thus reducing the overall effort needed while still maintaining or even enhancing model performance. Sarawagi

and Bhamidipaty [39] use active learning to pinpoint and prioritize data inconsistencies, automating the challenging aspects of deduplication, and minimizing the instances required to attain high accuracy. Complementing these efforts, several works [18, 20, 35] employ active learning for subset selection, efficiently constructing high-quality subsets from available data, thus maintaining or enhancing model performance efficiently.

2.1.2 Data Valuation

Data valuation methods aim to quantify the usefulness of data sources and assign value to individual data points, guiding data selection strategies. Yoon et al. [53] develop a data value estimator that employs a meta-learning framework to determine the value of data by jointly learning with the target task predictor model, while Kwon and Zou [24] apply the out-of-bag estimate in bagging models to evaluate data point contributions. Although Shapley values are commonly used for data valuation [4, 12, 23, 28, 34, 40, 46], their use is computationally intensive due to reliance on algorithms that estimate marginal contributions, and they face challenges like randomness from stochastic gradient descent [47]. To counter this, Wang and Jia [47] propose the Banzhaf value, derived from cooperative game theory, offering a more stable alternative and an effective estimation algorithm. Furthermore, Just et al. [17] present a model-agnostic framework using class-wise Wasserstein distance to value data points effectively. Practical applications are exemplified by Kim and Lee [19] in human activity recognition, demonstrating these methodologies' real-world applicability.

2.1.3 Reinforcement Learning

In reinforcement learning for data selection, Gutiérrez et al. [13] utilize Thompson Sampling within a Multi-Armed Bandit (MAB) framework to enhance medical image analysis by

selecting training data from partitioned clusters based on prediction accuracy. However, the effectiveness of this method is limited in contexts where not all data samples improve model performance or when necessary meta-information is unavailable. Additionally, Wang and Zeng [49] apply MAB to select high-value data items from large data streams, integrating various data characteristics into the reward function.

2.1.4 Other Techniques

Several other data selection techniques have been proposed to enhance efficiency and accuracy in various applications, including those handling large, high-dimensional datasets. Bernhardsson [2] utilize randomized tree forests to enhance approximate nearest neighbor searches, allowing for the efficient expansion of training sets with similar data points in a high-dimensional space and reducing both computational resources and query times. Lü et al. [29] introduce a method for redistributing weights within sentence pairs of parallel corpora, effectively prioritizing the most impactful data for machine translation without requiring extra resources. Everaert and Potts [6] propose minimizing the Kullback-Leibler (KL) divergence between a target set and the selected set to construct a subset for training, enhancing the relevance and effectiveness of the training data. Mirzasoleiman et al. [31] focus on identifying representative subsets of training data by maximizing a submodular function, which accelerates the learning process without compromising on model accuracy. Complementing this, Engstrom et al. [5] conceptualize the selection of training data as an optimization problem, aiming to maximize model performance by precisely modeling how individual data points contribute to learning and predicting target tasks.

2.2 Hierarchical Bandit

Hierarchical Bandit algorithms address the challenges of adaptive decision-making in environments where decisions are structured hierarchically and are particularly effective in scenarios where actions at a higher level influence the options available at a lower level.

2.2.1 Application in Recommendation Systems

Hierarchical bandit has been applied in recommendation systems. One earlier work [54] introduce a hierarchical bandit algorithm that accelerates contextual bandit learning in recommender systems by utilizing a coarse-to-fine feature space, resulting in significantly enhanced efficiency and outperforming traditional methods. Another work [48] present a novel hierarchical multi-armed bandit algorithm for automating IT service recommendations, structuring different automation in a hierarchy that aligns with the complexity of IT problems. A more recent work [51] apply hierarchical adaptive contextual bandits to make recommendations under resource constraints, featuring an upper layer for resource allocation and a lower layer focused on personalized recommendation. Zuo et al. [56] propose hierarchical bandit algorithms for conversational recommendations, leveraging the hierarchical relationship between key-term questions and item recommendations to enhance recommendation efficiency. In contrast, Santana et al. [38] propose to dynamically select the most appropriate system based on the current context by employing a meta-bandit structure that oversees a set of pre-trained recommender systems. These examples underline the effectiveness of hierarchical bandit algorithms in various decision-making tasks, suggesting their potential utility in dataset selection for enhancing model training.

2.2.2 Broad Applications Across Various Domains

Neyshabouri et al. [32] develop a hierarchical bandit framework for sequential learning that efficiently partitions the context space and combines mappings to bandit arms, achieving near-optimal arm selection in adversarial environments. By employing binary trees and other structures, their algorithm outshines existing methods in diverse experimental scenarios with both real and synthetic data. Zhao et al. [55] develop a hierarchical adversarial bandit framework to enhance relay selection in underwater acoustic networks, helping the network process and transfer data more efficiently and reducing communication costs. Ngo et al. [33] propose a strategy that leverages hierarchical edge computing to improve anomaly detection in IoT devices. This approach balances the trade-off between model complexity and response time, achieving improved accuracy and reduced delays, as validated through simulations on real IoT testbeds. Closer to our work, Wigmore et al. [50] employ Hierarchical Thompson Sampling for multi-band radio channel selection, demonstrating how leveraging the hierarchical structure between bands and channels significantly improves algorithmic performance over traditional methods. The hierarchical structure observed in frequency bands and channels bears a resemblance to the grouping of dataset groups and datasets in our context, providing a compelling parallel to our aim to explore whether such a hierarchical framework could enhance the dataset selection task.

2.2.3 Advanced Developments in Adaptive Exploration

To enhance the adaptability of bandit algorithms in dynamically complex environments, Kveton et al. [22] introduce Meta-Thompson Sampling (MetaTS), an innovative variant that meta-learns from interactions with bandit instances from an unknown prior. This approach enhances the exploration process in bandit algorithms by using past experience to inform bet-

ter decision-making in environments characterized by uncertainty. Hong et al. [14] propose HierTS, a hierarchical Bayesian model with a novel Thompson sampling algorithm designed to optimize decision-making across multiple hierarchical data structures. They provide an efficient implementation for Gaussian hierarchies, supported by theoretical analysis.

Chapter 3

Methodology

3.1 Overview

HBDS employs Bayesian modeling to assess the potential contributions of both dataset groups and individual datasets to a local model’s performance. Initially, HBDS represents dataset groups as the first layer of arms and individual datasets as the second layer, each associated with Gaussian distributions. HBDS then samples from these distributions to select the arms with the highest values. Upon receiving a reward, HBDS updates its knowledge using closed-form Bayesian posterior calculations. Figure 3.1 presents an overview of the method.

3.2 Problem Definition

Consider n data groups $\mathbf{g} = \{g_1, g_2, \dots, g_n\} = \{g_i\}_{i \in [n]}$, where each data group contains one or more datasets. The set of datasets belonging to data group g_i is denoted as $d_i = \{d_{i,j}\}_{j \in [m]}$, where i denotes the group index and j denotes the j -th dataset. Each dataset contains an arbitrary number of data points. The set of all datasets across all data groups is represented as $\mathbf{d} = \{d_i\}_{i \in [n]}$. The goal is to identify datasets that significantly enhance the performance of a local model for a given group.

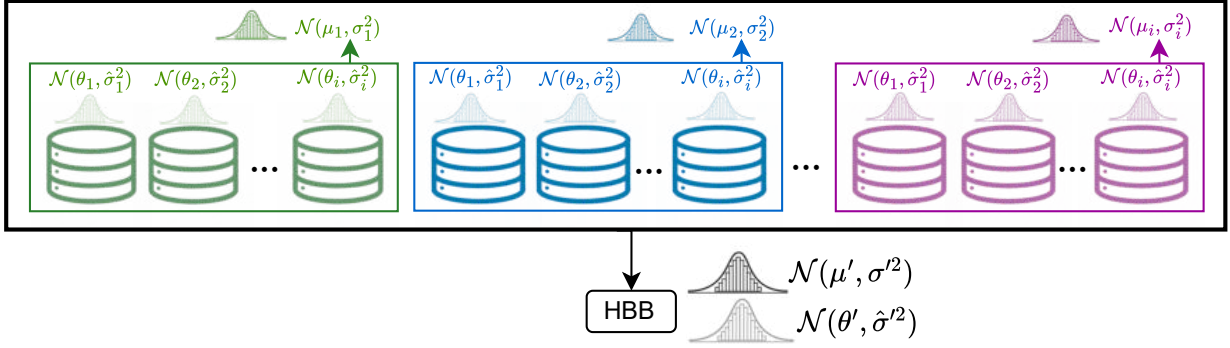


Figure 3.1: Diagram of the HBDS dataset selection method. Each dataset and its corresponding group are modeled using Gaussian distributions $\mathcal{N}(\theta_i, \hat{\sigma}_i^2)$ and $\mathcal{N}(\mu_i, \sigma_i^2)$ for datasets and dataset groups, respectively. The selection process involves choosing 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.

3.3 HBDS Initialization

HBDS addresses the dataset selection problem with a Hierarchical Bayesian Bandit (HBB) model. Bayesian modeling is a powerful statistical framework for making inferences about unknown parameters in a dataset while accounting for uncertainty [10, 11]. Each data group g_i is characterized by θ_i , and each dataset $d_{i,j}$ by $\theta_{i,j}$, with corresponding reward distributions $r_{i,j}(t)$. We assume normal distributions for data group and dataset priors, as well as reward distributions, with unknown means and known variances. Note that, given $\theta_{i,j}$, the reward $r_{i,j}(t)$ is conditionally independent of the data group parameter θ_i . The prior model is described by:

$$\begin{aligned}
 \theta_i &\sim, \mathcal{N}(\mu_i, \sigma_i^2), \forall i \in [n] \\
 \theta_{i,j} | \theta_i &\sim, \mathcal{N}(\theta_i, \hat{\sigma}_i^2), \forall j \in [m] \\
 r_{i,j}(t) | \theta_{i,j} &\sim \mathcal{N}(\theta_{i,j}, \sigma_r^2), \forall D(t) = d_{i,j},
 \end{aligned} \tag{3.1}$$

Algorithm 1 HBDS Dataset Selection

Initialize $P(\theta_i|r_i)$ group distributions, $P(\theta_{i,j}|r_{i,j})$ dataset distributions, $\mathcal{N}(\theta_{i,j}, \sigma_r^2)$ reward distributions

for $t = 1, \dots, T$ **do**

for $i = 1, \dots, n$ **do**

 Sample $\hat{\theta}_i(t) \sim P(\theta_i|r_i)$

end for

$i = \operatorname{argmax}[\theta_i(t) \forall i \in [n]]$ $\triangleright g_i$ is chosen

for $j = 1, \dots, m$ **do**

 Sample $\hat{\theta}_{i,j}(t) \sim P(\theta_{i,j}|r_{i,j})$

end for

$j = \operatorname{argmax}[\theta_{i,j}(t) \forall j \in [m]]$ $\triangleright d_{i,j}$ is chosen

 receive reward $r_{i,j}(t) = \mathbb{1}\{\hat{y} = y\}$

 update $P(\theta_i|r_i)$, $P(\theta_{i,j}|r_{i,j})$ \triangleright Eq. (3.2) and Eq. (3.3)

end for

where μ_i represents the mean of the prior distribution for data group g_i , σ_i^2 is the variance, $\hat{\sigma}_i^2$ is the variance of the dataset prior distribution $\theta_{i,j}$, and σ_r^2 is the variance of the reward distribution. 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, HBDS converges towards accurate estimations of the true distributions for both θ_i and $\theta_{i,j}$ after a number of iterations. The HBDS algorithm is described in Algorithm 1.

HBDS starts by assigning all dataset groups \mathbf{g} with the same group and dataset prior distributions $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{N}(\theta_0, \hat{\sigma}_0^2)$. At every 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 , HBDS 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 $D(t) = d_{i,j}$.

3.4 Posterior Distribution Computation

HBDS receives a reward from the chosen dataset and updates the distribution associated with the chosen dataset group and dataset using Eqs. (3.3) and (3.5). The posterior distribution of θ_i after observing reward values $r_i = \{r_{i,j}\}, j \in [m]$, where $r_{i,j} = \{r_{i,j}(t), \forall D(t) = d_{i,j}\}$, represents all rewards received by dataset $d_{i,j}$ up to current time step, is as follows:

$$\int_{\theta_{i,j}} \left(\prod_{j=1}^m \mathcal{N}(r_{i,j}; \theta_{i,j}, \sigma_r^2) \right) \mathcal{N}(\theta_{i,j}; \theta_i, \hat{\sigma}_i^2) d\theta_{i,j} \mathcal{N}(\theta_i; \mu_i, \sigma_i^2). \quad (3.2)$$

From Eq.(3.2), we can derive the posterior distribution of θ_i as

$$P(\theta_i | r_i) = \mathcal{N} \left(\lambda_i^2 \left(\frac{\mu_i}{\sigma_i^2} + \frac{\bar{s}_i}{\hat{\sigma}_i^2 + \frac{\sigma_r^2}{n_i}} \right), \lambda_i^2 \right) \quad (3.3)$$

where $\lambda_i^2 = \left(\frac{1}{\sigma_i^2} + \frac{1}{\hat{\sigma}_i^2 + \frac{\sigma_r^2}{n_i}} \right)^{-1}$ and $\bar{s}_i = \frac{\sum_{j=1}^m r_{i,j}}{n_i}$ is the mean of r_i . Here, n_i is the number of times dataset group g_i has been selected. The derivation of Equation (3.3) can be found in [10, 22]. The posterior distribution becomes the new prior for the next computation. Notice that the posterior mean is the weighted sum of the prior mean μ_i and the average reward \bar{s}_i . The weight of \bar{s}_i changes from $\frac{1}{\hat{\sigma}_i^2 + \sigma_r^2}$, when $n_i = 1$, to $\frac{1}{\hat{\sigma}_i^2}$, when $n_i \rightarrow \infty$. $\frac{1}{\hat{\sigma}_i^2}$ is the minimum amount of uncertainty that cannot be reduced by more selection.

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 current time step t , is:

$$\begin{aligned} P(\theta_{i,j} | r_{i,j}) &\propto P(\theta_{i,j}) p(r_{i,j} | \theta_{i,j}) \\ &= P(\theta_{i,j}) \prod_t p(r_{i,j}(t) | \theta_{i,j}), \forall D(t) = d_{i,j}. \end{aligned} \quad (3.4)$$

The posterior distribution of $\theta_{i,j}$ can be derived as

$$P(\theta_{i,j}|r_{i,j}) = \mathcal{N} \left(\lambda_{i,j}^2 \left(\frac{\theta_i}{\hat{\sigma}_i^2} + \frac{\bar{s}_{i,j}}{\frac{\sigma_r^2}{n_{i,j}}} \right), \lambda_{i,j}^2 \right) \quad (3.5)$$

where $\lambda_{i,j}^2 = \left(\frac{1}{\hat{\sigma}_i^2} + \frac{n_{i,j}}{\sigma_r^2} \right)^{-1}$, and $\bar{s}_{i,j} = \frac{r_{i,j}}{n_{i,j}}$ is the mean of $r_{i,j}$. Here, $n_{i,j}$ is the number of times dataset $d_{i,j}$ has been selected. The derivation of Equation 3.5 can be found in [11]. Different from the dataset group posterior, the dataset posterior only depends on the rewards received by the dataset. The mean is a weighted average of the conditional dataset group parameter θ_i and the averaged reward $\bar{s}_{i,j}$. 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 $n_{i,j} \rightarrow \infty$, the dataset posterior variance goes to zero, and the dataset posterior mean approaches the averaged reward $\bar{s}_{i,j}$.

3.5 Dataset Selection Based on Posterior Distributions

We formalize dataset selection using posterior means with a two-step process: first selecting a dataset group, then selecting a dataset within that group. A dataset or dataset group is selected if its posterior mean exceeds a percentile-based threshold within the context of all evaluated datasets or dataset groups, as described by, i.e.,

$$\text{Select if } \mu > F^{-1}(x),$$

where μ represents a posterior mean, and F^{-1} is the inverse of the cumulative distribution function (CDF) for 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 crite-

tion, 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.

Chapter 4

Experiments and Results

4.1 Experiments

4.1.1 Datasets

To validate the effectiveness of the proposed method, we leveraged two publicly available datasets **DIGIT-FIVE** and **DOMAINNET** [36]. Both datasets have been widely used to evaluate domain adaptation models [16, 21, 27, 30, 41, 43, 44, 52] and provide groupings based on different domain types for the same task.

DIGIT-FIVE Dataset. The **DIGIT-FIVE** dataset is a collection of five handwritten digit images and is commonly used for training machine learning digit recognition models. The dataset contains images of handwritten digits (0-9) from multiple writers, with variability in writing styles, stroke thickness, and other characteristics. Each image is associated with a label indicating the digit it represents. The dataset can be divided into five different groups:

- **MNIST [25]:** Comprises clean, grayscale images of handwritten digits in a uniform style, ideal for basic image processing.
- **MNIST-M [9]:** Augments **MNIST** digits onto backgrounds extracted from **BSDS500**, introducing complex color backgrounds and varying image textures.
- **USPS [15]:** Features smaller (16x16 pixels) grayscale images of digits from scanned

mail, characterized by variations in scale and stroke thickness.

- **SVHN [37]** : Contains real-world, full-color images of house numbers captured in natural scenes, with a range of fonts, overlapping numbers, and diverse lighting conditions.
- **SYN [9]**: Includes synthetic images of digits manipulated with different font styles and digital effects such as blur and noise, simulating a variety of digital environments.

For our experiments, we utilize preprocessed data as provided by Schrod et al. [41]. Digit images from **MNIST**, **MNIST-M**, and **USPS** are processed to dimensions of $(3, 28, 28)$, while those from **SYN** and **SVHN** resized to $(3, 32, 32)$. Initially, a feature extractor was trained on these images to obtain feature vectors of shape $(16, 7, 7)$, and the feature vectors were used for the experiments, not the actual image data. For each of the five digit datasets, we randomly sample data points to divide them into three mutually exclusive groups. We refer to each subgroups from **MNIST** subset as $\{mn0, mn1, mn2\}$, subgroups from **MNIST-M** as $\{mm0, mm1, mm2\}$, subgroups from **USPS** as $\{us0, us1, us2\}$, subgroups from **SVHN** as $\{sv0, sv1, sv2\}$, and subgroups from **SYN** as $\{sy0, sy1, sy2\}$.

DOMAINNET Dataset. The **DOMAINNET** [36] dataset is a large-scale unsupervised domain adaptation dataset that covers a wide range of object categories across six different domains, including real, clipart, painting, sketch, infograph, and quickdraw. Each domain represents a distinct artistic or graphical style. A subset of the **DOMAINNET** dataset is used to validate the proposed method. For our experiments, we select images from 15 classes across four domains: **CLIPART** (a collection of clip art images), **QUICKDRAW** (drawings from the global players of the game “Quick Draw”), **REAL** (photos and real-world images), and **SKETCH** (sketches of specific objects). Data pre-processing is conducted similarly to that of the **DIGIT-FIVE** dataset. For each of the domain subsets, we randomly sample data points to divide them into three mutually exclusive groups. We refer to each subgroups from the **CLIPART** subset

as $\{cl0, cl1, cl2\}$, QUICKDRAW subgroups as $\{qu0, qu1, qu2\}$, subgroups from REAL as $\{re0, re1, re2\}$, and subgroups from SKETCH as $\{sk0, sk1, sk2\}$.

The baseline HEG and the proposed HBDS consider the hierarchical structure of dataset groups. In practice, datasets can be either formed naturally, e.g., due to data generated by different stakeholders, or can be formed due to other kinds of metadata or machine operational status [26]. In our experiments, we consider two distinct settings of varying difficulty:

- **Perfect Groups:** Subsets derived from the same dataset form groups that represent different domains or data distributions. To simulate experiments when data from the same source share similar characteristics, we divide DIGIT-FIVE into subgroups based on domain, i.e., forming five subgroups, each with three distinct data subsets: $\{mn0, mn1, mn2\}$, $\{mm0, mm1, mm2\}$, $\{us0, us1, us2\}$, $\{sv0, sv1, sv2\}$, and $\{sy0, sy1, sy2\}$. Similarly, DOMAINNET forms five subgroups based on domain, each with three distinct data subsets, i.e., $\{cl0, cl1, cl2\}$, $\{qu0, qu1, qu2\}$, $\{re0, re1, re2\}$, $\{sk0, sk1, sk2\}$.
- **Mixed Grouping:** In other practical settings, cross-domain data collection is possible, e.g., an institution may possess datasets that originate from multiple domains that aim to solve distinct but similar tasks. To assess the robustness of HBDS under this challenging scenario, we experiment with mixed groupings where subsets from different datasets are combined. The DIGIT-FIVE dataset is divided into the following subgroups: $\{mn1, mn2, mm0\}$, $\{mm1, mm2, us0\}$, $\{us1, us2, sv0\}$, $\{sv1, sv2, sy0\}$, $\{sy1, sy2, mn0\}$. As for the DOMAINNET dataset, the mixed groupings are $\{cl1, cl2, qu0\}$, $\{qu1, qu2, re0\}$, $\{re1, re2, sk0\}$, $\{sk1, sk2, cl0\}$.

4.1.2 Baselines

To evaluate the proposed framework, we perform comparisons to demonstrate: (1) the advantage of the hierarchical structure in selecting datasets compared to non-hierarchical methods, and (2) the effectiveness of capturing dependencies between datasets. We compare the proposed HBDS with three baselines:

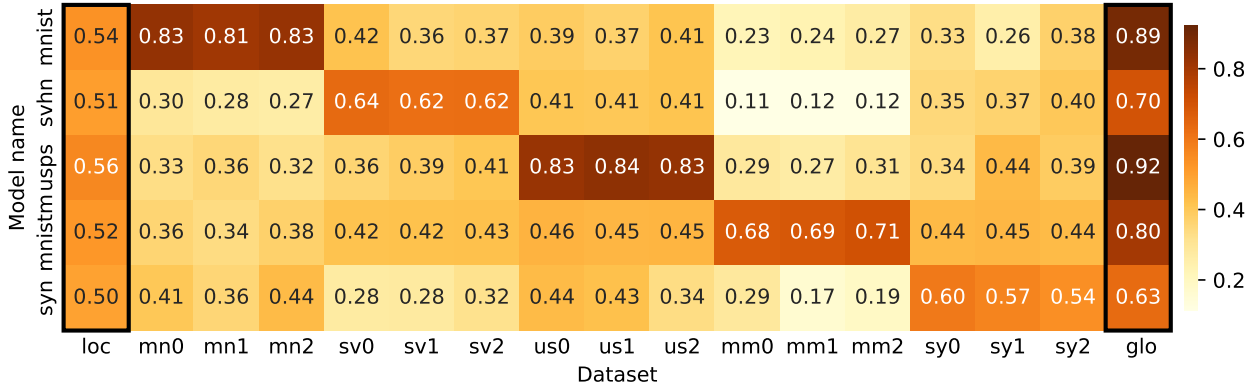
- **Epsilon-Greedy (EG)** [45]: Selects the best-performing dataset most of the time ($1-\epsilon$) but randomly selects any other dataset with probability ϵ , ensuring both exploration and exploitation. EG treats each dataset independently without considering any hierarchical group structure.
- **Bayesian Bandit (BB)**: Operating similarly to Thompson Sampling [1]. BB is closest to HBDS but without considering the hierarchical grouping of datasets.
- **Hierarchical Epsilon-Greedy (HEG)**: Adapts the EG strategy to a hierarchical context, using a two-stage selection process where the group is chosen based on the EG principle followed by dataset selection within the chosen group through another EG mechanism.
- **Local**: Model trained only on the local datasets from a specific data group, with no additional datasets used for training.
- **Global**: Model train on the local datasets from a specific data group, as well as the additional datasets originating from the same domain, irrespective of dataset groups, e.g., a model trained on all MNIST data subsets $\{\text{mn0}, \text{mn1}, \text{mn2}\}$.

4.1.3 Implementation details

For the **DIGIT-FIVE** dataset, the local classifiers consist of a single CNN layer. Each model is trained on a training set and evaluated on its test set, and the resulting accuracy is referred to as the *loc* (local accuracy). Figure 4.1 presents the ground truth accuracy heatmap, where the first column displays the local accuracy for each digit classifier on the **MNIST** ($\{\text{mn0}, \text{mn1}, \text{mn2}\}$), **MNIST-M** ($\{\text{mm0}, \text{mm1}, \text{mm2}\}$), **USPS** ($\{\text{us0}, \text{us1}, \text{us2}\}$), **SVHN** ($\{\text{sv0}, \text{sv1}, \text{sv2}\}$) and **SYN** ($\{\text{sy0}, \text{sy1}, \text{sy2}\}$) subgroups while the last column reveals the global accuracies achieved after each classifier is trained on the relevant subsets sampled from its corresponding dataset. For example, the global accuracy of 0.90 for **MNIST** is achieved by training the local model on $\{\text{mn0}, \text{mn1}, \text{mn2}\}$. Training on other datasets yields lower accuracies than the local accuracy, suggesting a degradation in performance. Therefore, the optimal performance for **MNIST** is attained by training on the datasets *mn0*, *mn1*, and *mn2*. The middle columns depict the accuracies of the local classifiers after additional training on each individual subset.

For the **DOMAINNET**[36] dataset, the local classifier consists of three fully connected layers. Figure 4.1 (b) shows that the **CLIPART** model exhibits the lowest local accuracy at 0.42, while the sketch model achieves the highest local accuracy at 0.67. In this benchmark, although the local model still gains the most improvement when trained on external sets from the same domain, datasets from other domains also improve model accuracy. For instance, in the case of the **REAL** model, **CLIPART** datasets $\{\text{cl0}, \text{cl1}, \text{cl2}\}$ contribute to enhancing local model performance. These characteristics render the **DOMAINNET** experiments closer to realistic data-sharing settings.

For EG and HEG, we use $\epsilon = 0.1$ as the exploration-exploitation trade-off parameter. For HBDS and BB, we set μ_0 and θ_i to zero, and σ_0^2 and $\hat{\sigma}_0^2$ to 2, as prior distributions for all



(a) DIGIT-FIVE



(b) DOMAINNET

Figure 4.1: Heatmap illustrating the accuracies of local classifiers post-training on different **DOMAINNET** subsets. The first column displays local accuracies on test sets, while the last column represents the optimal accuracy achievable considering all available relevant same-domain datasets. The middle columns depict accuracies after training on additional relevant subsets from the same domain.

dataset groups and datasets. We set the pre-defined percentile posterior mean threshold to 80 and 60 for the perfect and mixed groups, respectively. At every time step, HBDS decides on a dataset to select, retrieves a sample, and the local model predicts the sample’s label. The accuracy of this prediction determines the reward, i.e.,

$$r_{i,j}(t) = \begin{cases} 1 & \text{if } \hat{y} = y, \\ 0 & \text{otherwise,} \end{cases}$$

where \hat{y} represents the predicted label and y the actual label of the sample. This reward, either 1 for a correct prediction or 0 for an incorrect one, serves as the sole feedback for the algorithm to update its prior beliefs. The algorithm systematically refines these beliefs in response to the observed reward outcomes. For accurate and efficient dataset selection, we employ K-means clustering to identify representative data points, selecting five points nearest to the centroids in each cluster to encapsulate the dataset’s characteristics. Specifically, for **DIGIT-FIVE**, which comprises 10 distinct classes, we configure the clustering algorithm to generate 10 clusters to ensure that the variability inherent in each class is captured effectively. The model’s priors are updated exclusively using 5 near-centroid points from each cluster. Similarly, for the **DOMAINNET** data, we generate 15 clusters corresponding to the 15 classes in the dataset and use 5 near-centroid points from each cluster. An empirically determined natural stopping condition is employed, whereby the data selection stops when all representative points from a particular dataset are selected, indicating that the selection model has identified a specific dataset as likely to significantly enhance model performance. The total number of steps required to explore all representative points from all 15 **DIGIT-FIVE** data subsets is 750 (corresponding to the 15 data subsets, each with 10 clusters and 5 near-centroid points for each cluster). Similarly, the total number of steps required to explore all representative points from **DOMAINNET** is 1125. However, our experiments verify that the proposed empirical stopping criterion requires significantly fewer number of steps.

4.2 Experimental Results

In Table 4.1, we summarize results for **DIGIT-FIVE** dataset. We report mean and standard deviation over 5 experimental runs. To showcase the effectiveness of the proposed HBDS, we compare with global and local models, non-hierarchical baselines EG and BB, and the

Table 4.1: Performance comparison on **DIGIT-FIVE** of HBDS against baselines (averaged over 5 runs) under perfect and mixed group settings. Best performance is highlighted in **blue bold** and second-best is **bold**.

Method	Hierachical	MNIST	SVHN	USPS	MNIST-M	SYN	AVG
Local	\times	0.527 \pm .065	0.509 \pm .034	0.522 \pm .032	0.494 \pm .024	0.509 \pm .051	0.512 \pm .041
Global	\times	0.890 \pm .011	0.701 \pm .014	0.924 \pm .007	0.798 \pm .016	0.634 \pm .014	0.789 \pm .112
EG	\times	0.848 \pm .021	0.508 \pm .000	0.622 \pm .140	0.556 \pm .077	0.498 \pm .000	0.606 \pm .146
BB	\times	0.884\pm.022	0.703\pm.029	0.910\pm.017	0.790\pm.029	0.585\pm.064	0.774\pm.126
HEG	\checkmark	0.873 \pm .017	0.543 \pm .079	0.560 \pm .000	0.628 \pm .147	0.498 \pm .000	0.620 \pm .160
HEG (mixed)	\checkmark	0.869 \pm .009	0.536 \pm .064	0.628 \pm .153	0.522 \pm .000	0.512 \pm .031	0.613 \pm .051
HBDS	\checkmark	0.893\pm.011	0.707\pm.010	0.915\pm.013	0.792\pm.021	0.653\pm.026	0.792\pm.016
HBDS (mixed)	\checkmark	0.880 \pm .035	0.698 \pm .027	0.904 \pm .030	0.770 \pm .028	0.576 \pm .072	0.766 \pm .130

Table 4.2: Performance comparison on **DOMAINNET** of HBDS against baselines (averaged over 5 runs) under perfect and mixed group settings. Best performance is highlighted in **blue bold** and second-best is **bold**.

Method	Hierachical	CLIPART	QUICKDRAW	REAL	SKETCH	AVG
Local	\times	0.400 \pm .024	0.640 \pm .021	0.610 \pm .011	0.675 \pm .011	0.581 \pm .017
Global	\times	0.784 \pm .006	0.868 \pm .013	0.876 \pm .015	0.737 \pm .009	0.816 \pm .011
EG	\times	0.739 \pm .015	0.742 \pm .123	0.803 \pm .115	0.670 \pm .000	0.739 \pm .063
BB	\times	0.767 \pm .009	0.874\pm.011	0.885\pm.009	0.711\pm.008	0.809\pm.009
HEG	\checkmark	0.772 \pm .030	0.784 \pm .120	0.880 \pm .009	0.677 \pm .015	0.778 \pm .044
HEG (mixed)	\checkmark	0.697 \pm .117	0.776 \pm .113	0.852 \pm .055	0.689 \pm .019	0.754 \pm .076
HBDS	\checkmark	0.785\pm.001	0.878\pm.004	0.886\pm.009	0.716\pm.021	0.816\pm.009
HBDS (mixed)	\checkmark	0.769\pm.029	0.866 \pm .008	0.884 \pm .010	0.69 \pm .021	0.802 \pm .017

hierarchical HEG baseline, each after continuing training the local model on the datasets selected by each method. The best and second-best performance is highlighted in bold.

HBDS achieves or approaches global accuracy for all five models under perfect grouping. The Bayesian Bandit (BB) baseline is also close to the global accuracy, but its performance on SYN and its average accuracy are lower than HBDS by 6.8% and 1.8% respectively. Despite appearing nearly as effective as the hierarchical HBDS, BB requires more training steps. In particular, as shown in Figure 4.2, hierarchical methods reach their stopping condition roughly 25 steps earlier than their non-hierarchical counterparts, showcasing a more efficient learning process. Epsilon Greedy (EG) and Hierarchical Epsilon Greedy (HEG) take fewer steps, but their performance is sub-optimal. Specifically, EG fails to identify any of the

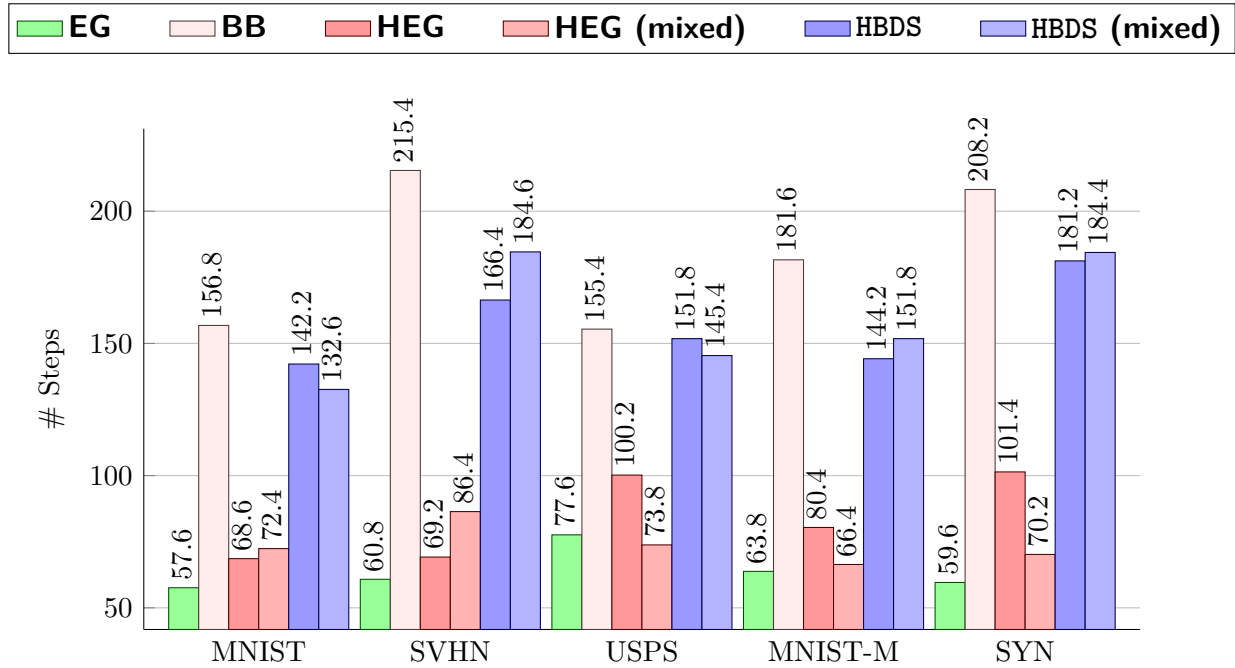


Figure 4.2: # Steps required on DIGIT-FIVE across baselines

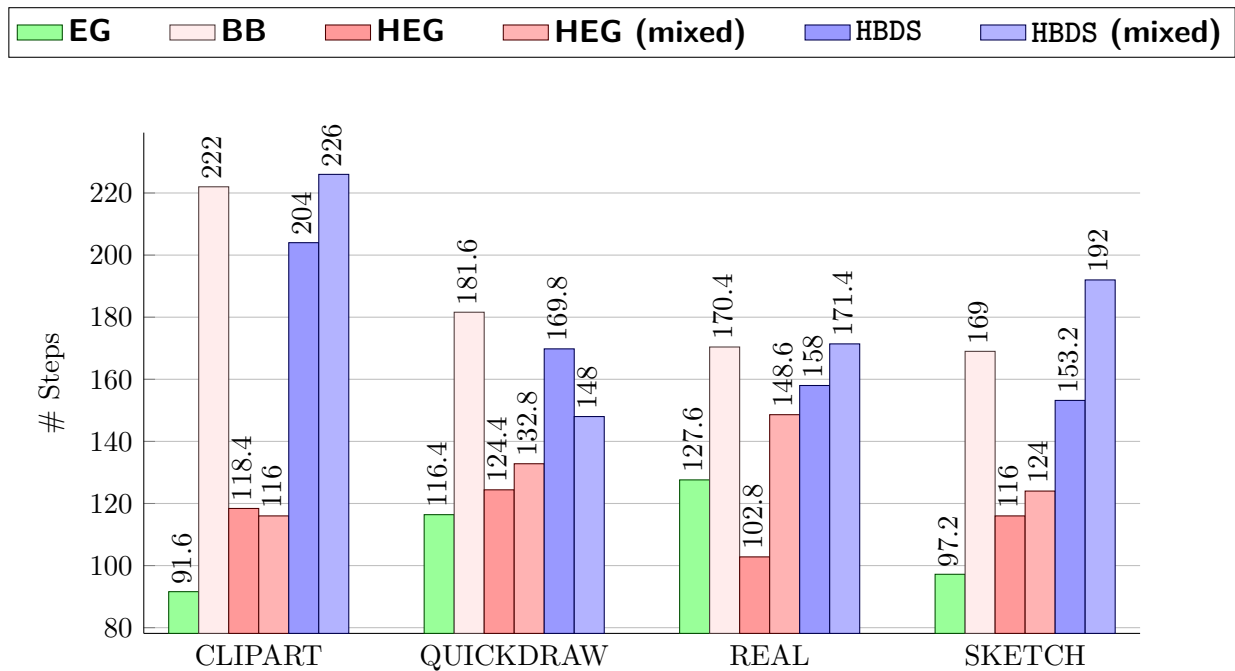


Figure 4.3: # Steps required on DOMAINNET across baselines

relevant datasets for SVHN and SYN, resulting in 0% improvement over the local model. Similarly, HEG fails on MNIST-M and SYN. The average performance of EG and HEG are

lower than HBDS by 18.2% and 16.8% respectively.

Under mixed grouping, HBDS achieves or approaches global accuracy for all models except for SYN, and reaches the stopping condition roughly 23 steps sooner than the BB method, while the average accuracy is only 0.8% lower than BB. In comparison, the HEG under mixed grouping fails to identify any relevant datasets for MNIST-M and achieves marginal performance improvements for SVHN and SYN. In general, the mixed group setting deteriorates the selection accuracy but only slightly, e.g., about 2% decrease in accuracy on average.

In Table 4.2 and Figure 4.3, we observe similar conclusions in experiments conducted on the DomainNet dataset. The BB and HBDS both achieve or approach global accuracy for all four models. The average accuracy for HBDS under perfect grouping is 0.7% higher than the BB method and takes about 15 fewer steps. On the contrary, the EG and HEG methods fail to improve model performance on SKETCH, and the EG and HEG average accuracy is lower than the global average by 7.7% and 3.8%, respectively. Comparing under mixed grouping setting, HBDS has 6.3% higher accuracy than the EG method on average. Additionally, HBDS achieves 4.8% relative performance gains over the HEG. While the difference is not as pronounced as with the DIGIT-FIVE dataset, this is likely due to the fact that the data used in the DOMAINNET experiments use the same pre-trained ResNet 18 as a feature extractor after training on the dataset. This uniformity makes the datasets more similar, thus complicating the identification of a dataset.

4.3 Ablation Studies

To gain a comprehensive understanding of HBDS’s capabilities, we consider the following questions for ablations on the DIGIT-FIVE dataset: **(1) Limited exploration:** How does HBDS perform in comparison to baselines under low-resource exploration settings? **(2) Subpar**

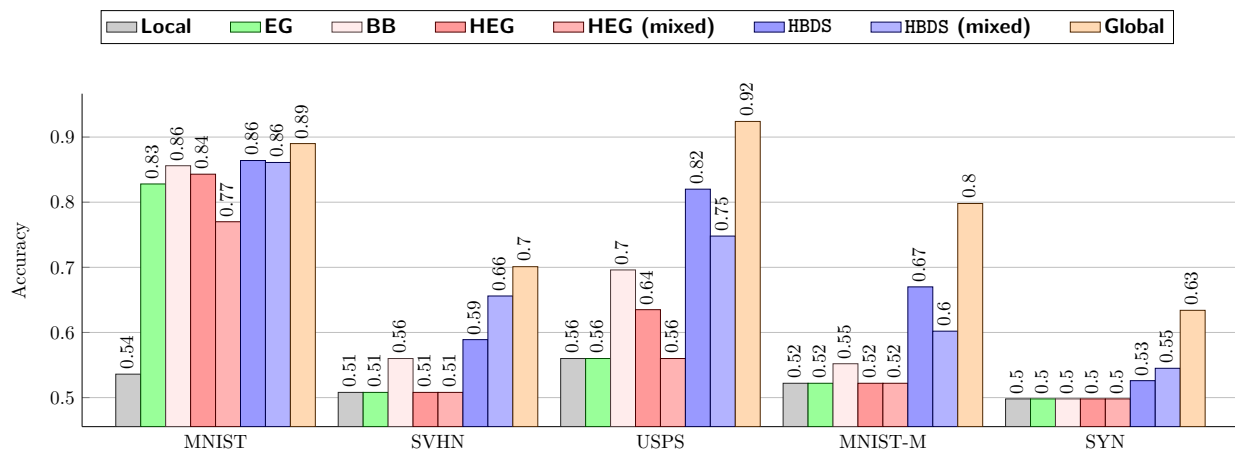


Figure 4.4: Performace comparison under limited exploration settings

local model performance: What insights can be gleaned from analyzing HBDS when the initial local model accuracy is low? **(3) Absence of relevant datasets:** What happens when no relevant datasets are available across all dataset groups? **(4) Scalability:** Is HBDS similarly effective when there are more datasets present in each group?

4.3.1 Comparison Under Limited Exploration

In this low-resource scenario, each method is allowed to explore each dataset only once on average, equating to 15 steps considering the total of 15 datasets in **DIGIT-FIVE**. Figure 4.4 depicts the local model accuracy improvements after training on datasets selected by each method. While the performance of all methods on the **MNIST** closely aligns, possibly due to the dataset’s simplicity, both **HBDS** and **HBDS (mixed)** outperform the baselines. For the other four datasets, **HBDS** and **HBDS (mixed)** also consistently surpass the baselines. The **EG** method fails to identify any useful datasets after just 15 training steps. Similarly, the **HEG** method only succeeds in enhancing performance on **USPS**. Meanwhile, the selections made by **HBDS** and **HBDS (mixed)** lead to accuracy improvements for all five data subsets, underscoring its effectiveness in low-resource conditions. Quantitatively, **HBDS** outperforms

Table 4.3: HBDS improves model performance even when the initial local model exhibits extremely low starting accuracy

% Train Data	Name	Initial Local Acc.	HBDS Acc.
10%	MNIST	0.176	0.315
	SVHN	0.128	0.242
	USPS	0.096	0.135
	MNIST-M	0.206	0.551
	SYN	0.266	0.376
20%	MNIST	0.236	0.896
	SVHN	0.212	0.215
	USPS	0.128	0.286
	MNIST-M	0.288	0.576
	SYN	0.214	0.249
50%	MNIST	0.366	0.896
	SVHN	0.356	0.667
	USPS	0.312	0.914
	MNIST-M	0.442	0.793
	SYN	0.274	0.410

its non-hierarchical counterpart, BB, by 6.9% on average.

4.3.2 Local Model with Low Initial Accuracy

We further investigate whether HBDS can effectively enhance performance when the initial local model accuracy is extremely low. To assess this, we train the initial local classifiers using only 10%, 20%, and 50% of the available training data. Results presented in Table 4.3 demonstrate consistent gains in accuracy across all tested conditions, even when the initial local accuracy is as low as 0.096%. These results indicate that HBDS is robust against significant variations in initial local model performance (prior to dataset selection).

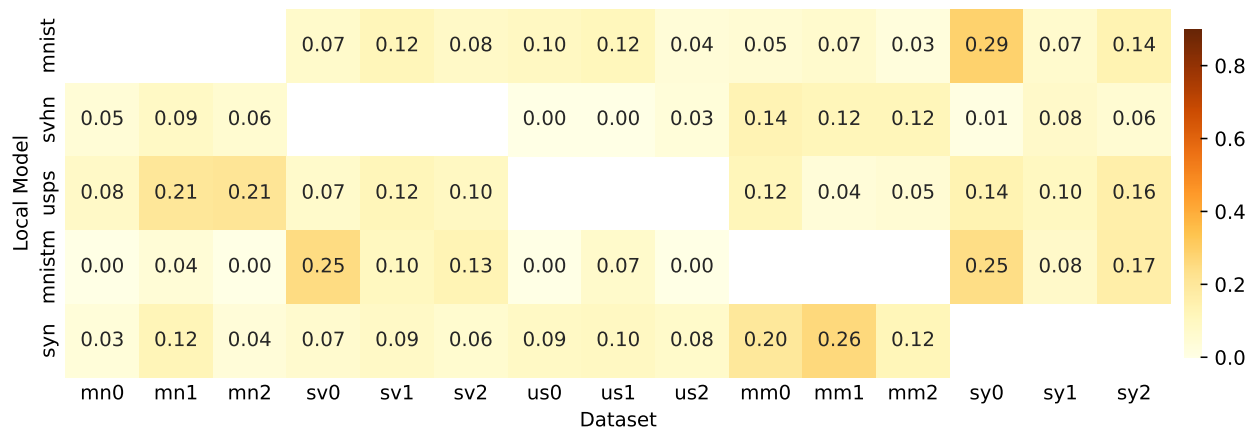


Figure 4.5: When there exists no relevant dataset available across all groups, HBDS posterior means remain low even after 600 exploration steps. This ensures HBDS can reliably discern irrelevant data.

4.3.3 Absence of Relevant Datasets

In addition, we evaluate the performance of HBDS in scenarios where no relevant dataset is available. The analysis in Figure 4.5 reveals that the posterior means remain low, not exceeding 0.3, even after extensive exploration of the dataset pool. This suggests that HBDS can effectively identify the lack of beneficial datasets, avoiding misguided training efforts.

4.3.4 Presents of more datasets

We also run experiments to test the scalability of HBDS within a larger data-sharing ecosystem. We construct more datasets from each DIGIT-FIVE dataset group. The MNIST, SVHN, USPS, MNISTM, and SYN groups have 10, 12, 11, 9, 9 datasets, respectively. We apply HBDS to newly constructed datasets and observe that it is able to identify all useful datasets and achieve higher accuracy on all five models (See Figure 4.6). The computation time and resource usage remain the same for each time step. In addition, The scaling of the number of steps required is sublinear, e.g., SVHN group contains 4 times as many datasets but only takes 2.6 times more steps.

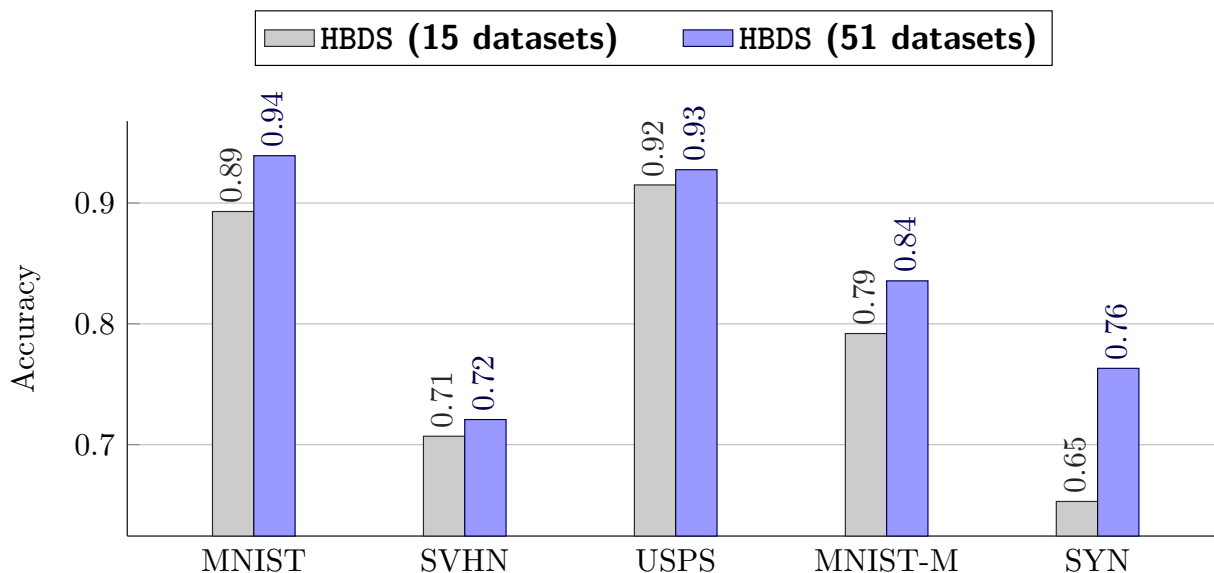
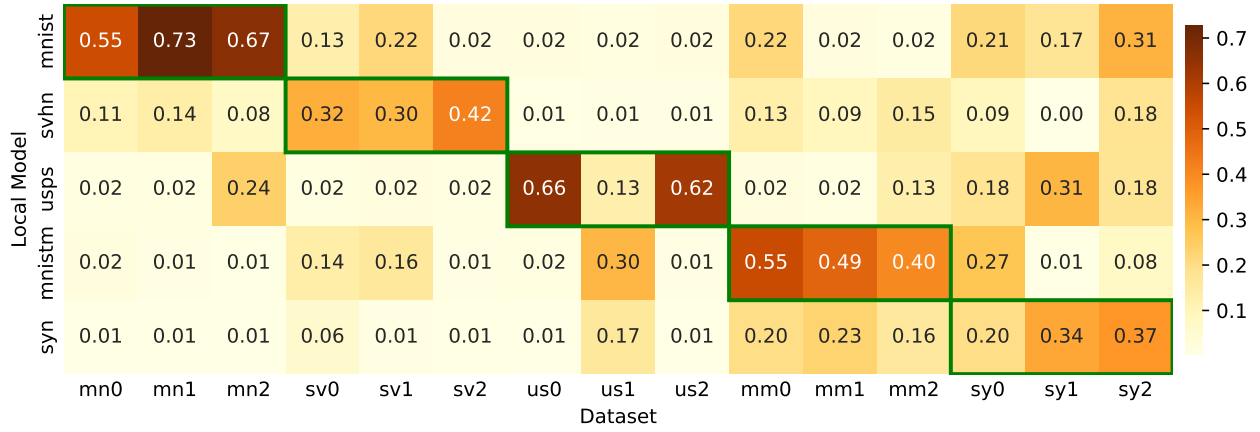


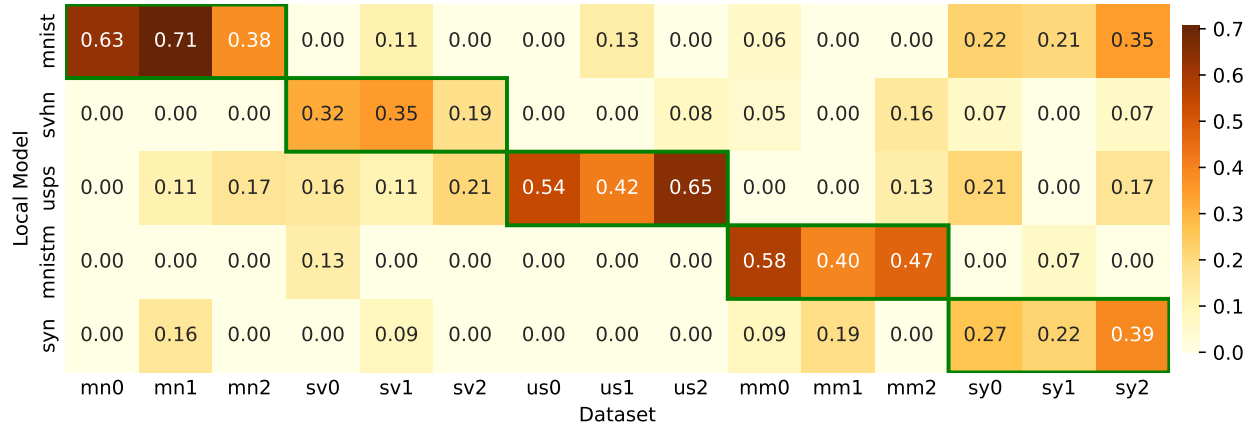
Figure 4.6: Scalability comparison when more datasets across five data groups are present. HBDS achieves higher accuracy across all models with the presence of additional datasets.

4.4 Qualitative Analysis

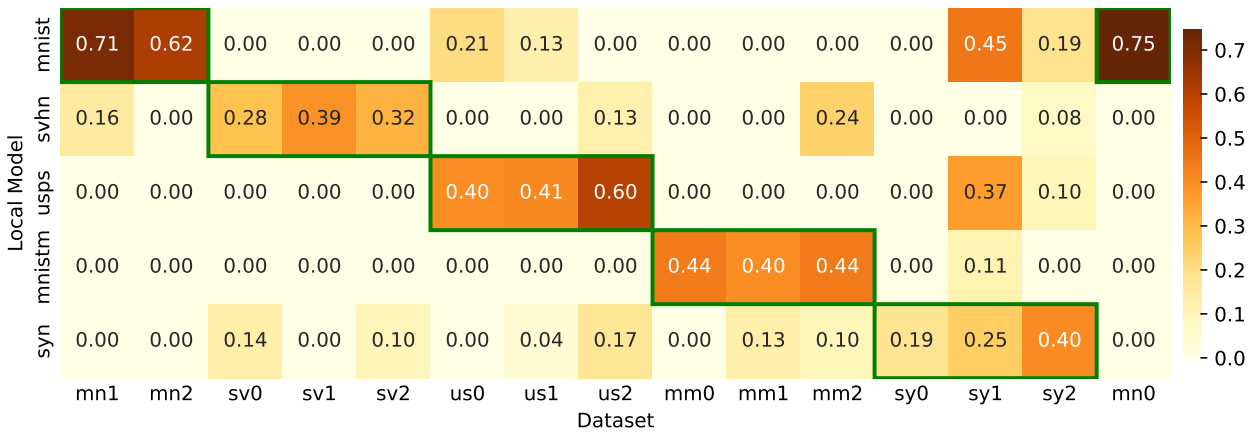
We qualitatively analyze the learned posterior distributions of each method. Figure 4.7 presents the posterior means for BB, HBDS, and HBDS (mixed). The heatmap analysis reveals that HBDS and HBDS (mixed) effectively capture the relationship between the local model and the relevant datasets, using fewer computational steps than baselines. In contrast, the BB method demonstrates less precision in identifying the most relevant datasets for USPS and SYN. We further analyze the posterior distributions of the datasets and dataset groups for HBDS after training on all relevant data subsets available. Figure 4.8 reveals a clearer distinction between beneficial datasets (Fig. 4.8a) and relevant dataset groups (Fig. 4.8b).



(a) BB Posterior Means



(b) HBDS Posterior Means



(c) HBDS (mixed) Posterior Means

Figure 4.7: The posterior means of BB, HBDS and HBDS (mixed).

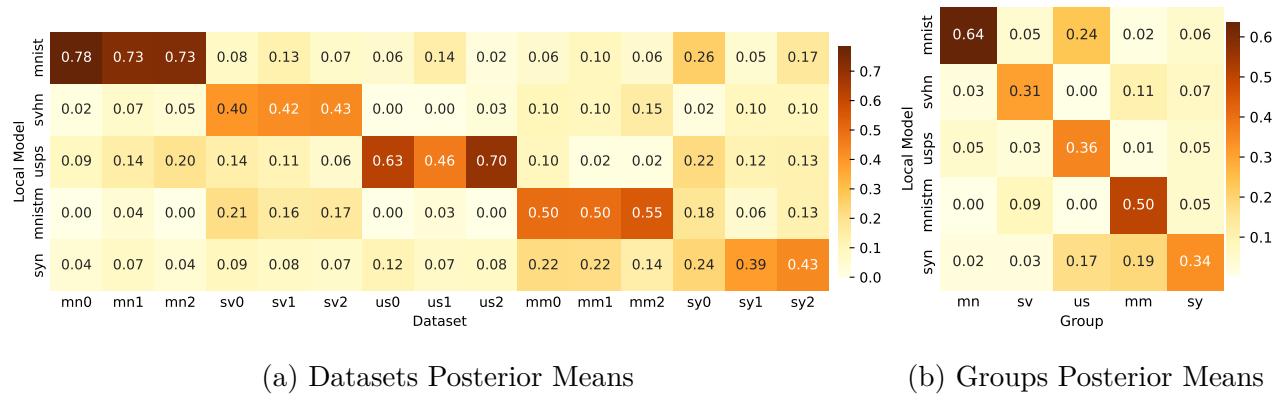


Figure 4.8: Dataset and dataset group posterior means as determined by HBDS after training on all relevant available data. (a) Datasets within the same group exhibit consistently high posterior means, indicating their relevance to the local data. (b) Likewise, the highest values are assigned to groups containing the most relevant datasets.

Chapter 5

Conclusions

This work introduces HBDS, the first dataset selection method designed to identify relevant datasets for model training by jointly assessing the potential contributions of dataset groups and individual datasets. Experimental results validate the robustness of HBDS, particularly under non-ideal dataset groupings typical in real-world scenarios. Experiments on two benchmark datasets demonstrate that HBDS not only selects datasets that improve model performance under different resource settings but does so more efficiently than non-hierarchical approaches. In low-resource environments, HBDS significantly outperformed the baselines by effectively using limited dataset explorations to identify datasets that improve local model performance. Qualitative analysis shows that posterior means accurately reflect relevant dataset groups and datasets.

In the future, we aim to further validate our method with manufacturing datasets to demonstrate its practical capabilities, as well as datasets from various domains, e.g., time series, text, and graph datasets. Additionally, we plan to refine our approach for obtaining representative data points to better estimate posterior distributions. While we exclusively used the Gaussian distribution to estimate potential distributions, other distributions such as the Bernoulli distribution could be equally effective. Further experiments are needed to verify this adaptability. Currently, we have employed the same model structure throughout our experiments; however, this is not a requirement. The local model can take any form, as the selection mechanism is independent of the model structure. For instance, in the classi-

fication tasks of **DIGIT-FIVE**, we could employ KNN or SVMs instead of a single CNN layer model. Moreover, we intend to enhance our model by integrating state-of-the-art data selection techniques after datasets are selected to address the variability in data quality within identified datasets. Future research will also explore scenarios where adversarial users upload fake datasets to undermine the method. For example, if there is data leakage from the local dataset, an adversarial user may construct a fake dataset based on the local dataset. Another crucial aspect of the data-sharing ecosystem is privacy. If standard procedures are used to obtain proxy data, such as the feature extraction from digit images we employed, and the local model is trained on this proxy data, the proposed method should be able to identify proxy datasets that enhance its performance. However, the real challenge emerges when the local model is trained on raw data while all the shared data are proxy data, or when proxy data is produced differently by various data owners. One way to ensure improved local model performance while protecting data privacy is to automate the entire process, thereby preventing any access to raw data.

Bibliography

- [1] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.
- [2] Erik Bernhardsson. Annoy (approximate nearest neighbors oh yeah). <https://github.com/spotify/annoy>, 2017. Apache-2.0 license.
- [3] Victor Christen, Peter Christen, and Erhard Rahm. Informativeness-based active learning for entity resolution. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 125–141. Springer, 2020.
- [4] Christie Courtnage and Evgueni Smirnov. Shapley-value data valuation for semi-supervised learning. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*, pages 94–108. Springer, 2021.
- [5] Logan Engstrom, Axel Feldmann, and Aleksander Madry. Dsdm: Model-aware dataset selection with datamodels. *arXiv preprint arXiv:2401.12926*, 2024.
- [6] Dante Everaert and Christopher Potts. Gio: Gradient information optimization for training dataset selection. In *The Twelfth International Conference on Learning Representations*, 2023.
- [7] Thiago Fraga-Silva, Jean-Luc Gauvain, Lori Lamel, Antoine Laurent, Viet-Bac Le, and Abdel Messaoudi. Active learning based data selection for limited resource stt and kws.

- In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [8] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.
- [9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.
- [10] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY, 2007.
- [11] Andrew Gelman, John Carlin, Hal Stern, Donald Rubin, David Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC Press, 2013.
- [12] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.
- [13] Benjamín Gutiérrez, Loïc Peter, Tassilo Klein, and Christian Wachinger. A multi-armed bandit to smartly select a training set from big medical data. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 38–45. Springer, 2017.
- [14] Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7724–7741. PMLR, 2022.

- [15] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.
- [16] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Re-energizing domain discriminator with sample relabeling for adversarial domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9174–9183, 2021.
- [17] Hoang Anh Just, Feiyang Kang, Tianhao Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. Lava: Data valuation without pre-specified learning algorithms. In *The Eleventh International Conference on Learning Representations*. OpenReview, 2023.
- [18] Vishal Kaushal, Anurag Sahoo, Khoshrav Doctor, Narasimha Raju, Suyash Shetty, Pankaj Singh, Rishabh Iyer, and Ganesh Ramakrishnan. Learning from less data: Diversified subset selection and active learning in image classification tasks. *arXiv preprint arXiv:1805.11191*, 2018.
- [19] Yeon-Wook Kim and Sangmin Lee. Data valuation algorithm for inertial measurement unit-based human activity recognition. *Sensors*, 23(1):184, 2022.
- [20] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.
- [21] Tatsuya Komatsu, Tomoko Matsui, and Junbin Gao. Multi-source domain adaptation with sinkhorn barycenter. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 1371–1375. IEEE, 2021.
- [22] Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. Meta-thompson sampling. In *International Conference on Machine Learning*, pages 5884–5893. PMLR, 2021.

- [23] Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 8780–8802. PMLR, 2022.
- [24] Yongchan Kwon and James Zou. Data-oob: out-of-bag estimate as a simple and efficient data value. In *International Conference on Machine Learning*, pages 18135–18152. PMLR, 2023.
- [25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [26] Yifu Li, Xinwei Deng, Shan Ba, William R Myers, William A Brenneman, Steve J Lange, Ron Zink, and Ran Jin. Cluster-based data filtering for manufacturing big data systems. *Journal of Quality Technology*, 54(3):290–302, 2022.
- [27] Yunsheng Li, Lu Yuan, Yinpeng Chen, Pei Wang, and Nuno Vasconcelos. Dynamic transfer for multi-source domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10998–11007, 2021.
- [28] Zhihong Liu, Hoang Anh Just, Xiangyu Chang, Xi Chen, and Ruoxi Jia. 2d-shapley: a framework for fragmented data valuation. In *International Conference on Machine Learning*, pages 21730–21755. PMLR, 2023.
- [29] Yajuan Lü, Jin Huang, and Qun Liu. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350, 2007.
- [30] Shuang Luo, Didi Zhu, Zexi Li, and Chao Wu. Ensemble federated adversarial training with non-iid data. *arXiv preprint arXiv:2110.14814*, 2021.

- [31] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020.
- [32] Mohammadreza Mohaghegh Neyshabouri, Kaan Gokcesu, Hakan Gokcesu, Huseyin Ozkan, and Suleyman Serdar Kozat. Asymptotically optimal contextual bandit algorithm using hierarchical structures. *IEEE transactions on neural networks and learning systems*, 30(3):923–937, 2018.
- [33] Mao V Ngo, Tie Luo, and Tony QS Quek. Adaptive anomaly detection for internet of things in hierarchical edge computing: A contextual-bandit approach. *ACM Transactions on Internet of Things*, 3(1):1–23, 2021.
- [34] Konstantin D Pandl, Fabian Feiland, Scott Thiebes, and Ali Sunyaev. Trustworthy machine learning for health care: scalable data valuation with the shapley value. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 47–57, 2021.
- [35] Sujoy Paul, Jawadul H Bappy, and Amit K Roy-Chowdhury. Non-uniform subset selection for active learning in structured data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6846–6855, 2017.
- [36] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.
- [37] P Roy, S Ghosh, S Bhattacharya, and U Pal. Effects of degradations on deep neural network architectures. arxiv 2018. *arXiv preprint arXiv:1807.10108*, 1807.
- [38] Marlesson RO Santana, Luckeciano C Melo, Fernando HF Camargo, Bruno Brandão, Anderson Soares, Renan M Oliveira, and Sandor Caetano. Contextual meta-bandit

- for recommender systems selection. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 444–449, 2020.
- [39] Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–278, 2002.
- [40] Stephanie Schoch, Haifeng Xu, and Yangfeng Ji. Cs-shapley: class-wise shapley values for data valuation in classification. *Advances in Neural Information Processing Systems*, 35:34574–34585, 2022.
- [41] Stefan Schrod, Jonas Lippl, Andreas Schäfer, and Michael Altenbuchinger. Fact: Federated adversarial cross training. *arXiv preprint arXiv:2306.00607*, 2023.
- [42] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.
- [43] Christian Simon, Masoud Faraki, Yi-Hsuan Tsai, Xiang Yu, Samuel Schuler, Yumin Suh, Mehrtash Harandi, and Manmohan Chandraker. On generalizing beyond domains in cross-domain continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9265–9274, 2022.
- [44] Ankit Singh. Clda: Contrastive learning for semi-supervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:5089–5101, 2021.
- [45] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [46] Siyi Tang, Amirata Ghorbani, Rikiya Yamashita, Sameer Rehman, Jared A Dunnmon, James Zou, and Daniel L Rubin. Data valuation for medical imaging using shapley

- value and application to a large-scale chest x-ray dataset. *Scientific reports*, 11(1):8366, 2021.
- [47] Jiachen T Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6388–6421. PMLR, 2023.
- [48] Qing Wang, Tao Li, SS Iyengar, Larisa Shwartz, and Genady Ya Grabarnik. Online it ticket automation recommendation using hierarchical multi-armed bandit algorithms. In *Proceedings of the 2018 SIAM international conference on data mining*, pages 657–665. SIAM, 2018.
- [49] Shun Wang and Guosun Zeng. A novel approach to select high-reward data items in big data stream based on multiarmed bandit. *IEEE Transactions on Computational Social Systems*, 9(4):1144–1153, 2021.
- [50] Jerrod Wigmore, Brooke Shrader, and Eytan Modiano. Hierarchical thompson sampling for multi-band radio channel selection. In *2023 IFIP Networking Conference (IFIP Networking)*, pages 1–9. IEEE, 2023.
- [51] Mengyue Yang, Qingyang Li, Zhiwei Qin, and Jieping Ye. Hierarchical adaptive contextual bandits for resource constraint based recommendation. In *Proceedings of the web conference 2020*, pages 292–302, 2020.
- [52] Chun-Han Yao, Boqing Gong, Hang Qi, Yin Cui, Yukun Zhu, and Ming-Hsuan Yang. Federated multi-target domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1424–1433, 2022.
- [53] Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement

- learning. In *International Conference on Machine Learning*, pages 10842–10851. PMLR, 2020.
- [54] Yisong Yue, Sue Ann Hong, and Carlos Guestrin. Hierarchical exploration for accelerating contextual bandits. In *Proceedings of the 29th International Conference on Machine Learning*, pages 979–986, 2012.
- [55] Haihong Zhao, Xinbin Li, Song Han, Lei Yan, and Junzhi Yu. Adaptive relay selection strategy in underwater acoustic cooperative networks: A hierarchical adversarial bandit learning approach. *IEEE Transactions on Mobile Computing*, 22(4):1938–1949, 2021.
- [56] Jinhang Zuo, Songwen Hu, Tong Yu, Shuai Li, Handong Zhao, and Carlee Joe-Wong. Hierarchical conversational preference elicitation with bandit feedback. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2827–2836, 2022.