ESTIMATING DATASET LEARNING UTILITY BEFORE TRAINING: A MANIFOLD VOLUME PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Assessing the quality of a candidate dataset before model training is increasingly important for data acquisition and emerging data marketplaces. However, dataset size can be easily inflated by duplicate or highly redundant samples, while existing intrinsic quality metrics do not directly reflect downstream learning utility and model-aware valuation methods often require training on the candidate data itself. In this work, we study label-free and fine-tuning-free dataset-level quality assessment, aiming to estimate the future learning utility of an entire candidate dataset before training on it. We propose a simple geometric criterion based on the manifold volume of frozen pretrained representations. Our key hypothesis is that informative and non-redundant data span richer and more diverse feature directions, whereas redundant data concentrate in a more restricted subspace. Theoretically, we show that manifold volume is invariant to trivial sample replication, becomes dimensionally constrained under severe redundancy, and exhibits diminishing returns along already well-covered feature directions. We evaluate the proposed criterion through controlled redundancy experiments on nine datasets spanning three modalities, four architecture families, and 13 dataset-backbone pairs. Manifold volume consistently exhibits a strong positive linear association with downstream learning utility, achieving the highest average Pearson correlation of 0.816.

1 INTRODUCTION

High-quality annotated data is critical for modern machine learning systems, especially in specialized domains such as medical imaging, industrial fault diagnosis, rare-event autonomous driving, and scientific discovery, where reliable annotation requires substantial expert effort. As model development increasingly relies on external data acquisition, curation, fine-tuning, and post-training, determining whether a candidate dataset is truly valuable has become increasingly important for both model developers and emerging data marketplaces. However, dataset size alone is insufficient to characterize its actual value. A dataset can be substantially enlarged by duplicating existing examples, introducing near-duplicate samples, or padding it with highly redundant data, without providing a proportional amount of new information for model learning. Recent studies also show that the marginal utility of training data can diminish rapidly under repeated exposure Goyal et al. (2024). Therefore, a meaningful notion of dataset quality should go beyond how much data is present and instead reflect how much useful learning signal the data can provide to a model.

Existing dataset quality assessment methods primarily characterize data through intrinsic properties, such as noise, redundancy, representativeness, relevance, and diversity Gong et al. (2023). Prior studies have shown that dataset bias and insufficient representativeness can substantially affect cross-dataset generalization Torralba & Efros (2011), while annotation errors constitute another important source of dataset quality degradation Northcutt et al. (2021). Redundancy is also a critical factor: large-scale training corpora may contain substantial exact or near-duplicate content, and removing such redundancy can preserve or even improve learning efficiency Lee et al. (2022). Subsequent studies have moved beyond nominal dataset size by measuring effective diversity or complementary information in lexical, semantic, or representation spaces using statistical, spectral, or similarity-based criteria Friedman & Dieng (2023); Amiri et al. (2023). These approaches provide efficient indicators of whether a dataset is noisy, redundant, diverse, representative, or complementary to existing data. However, they mainly characterize properties of the data itself, rather than the actual capability improvement that the candidate data can bring to a model after training.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To more directly characterize learning utility, model-aware data valuation methods estimate the contribution of training data to model predictions, losses, or generalization. Early studies mainly focus on instance-level valuation through marginal performance contribution, gradients, training trajectories, or counterfactual model behavior Ghorbani & Zou (2019); Koh & Liang (2017); Pruthi et al. (2020); Ilyas et al. (2022); Park et al. (2023). Utility-driven data selection further prioritizes examples according to whether they remain learnable or useful for subsequent model improvement Mindermann et al. (2022); Xia et al. (2024). More recent work extends this idea from individual examples to data subsets, domains, and entire datasets, estimating their learning potential through proxy models, data-mixture optimization, scaling laws, or gradient-based dataset comparison Xie et al. (2023); Fan et al. (2024); Jiang et al. (2025); Zeng et al. (2026). Although these approaches more directly connect data value with model performance, their utility estimates are typically obtained through additional model training, gradient computation, proxy optimization, or online learning dynamics. Training-free valuation has also been explored at the instance level Ki et al. (2023), but efficiently assessing the utility of an entire candidate dataset before training remains underexplored. This limitation is particularly problematic in data marketplaces, where the candidate dataset, especially its valuable annotations, must remain undisclosed before pricing and acquisition. Consequently, any valuation method that requires training on the candidate data is impractical.

To address this gap, we study label-free and fine-tuning-free dataset quality assessment to estimate the future learning utility of an entire candidate dataset. Given a candidate dataset and an available pretrained model, we hypothesize that its effective information can be characterized by the geometry of its pretrained representations: informative and non-redundant samples span richer and more diverse feature directions, whereas redundant samples concentrate along fewer effective directions. Accordingly, we represent the candidate dataset as a point-cloud manifold and quantify its joint geometric span using a regularized log-determinant manifold volume computed from the centered feature covariance. From a spectral perspective, this quantity jointly reflects the number of effective feature directions occupied by the dataset and the variation along those directions. We further show theoretically that the proposed volume is invariant to trivial sample replication, becomes dimensionally constrained under severe redundancy, and exhibits diminishing returns along already well-covered feature directions. These properties provide a theoretical basis to characterize effective information beyond nominal dataset size. Thus, manifold volume provides a label-free and fine-tuning-free dataset-level utility estimate using only frozen-model feature extraction.

We evaluate this hypothesis through controlled redundancy experiments across image, text, and audio classification. For each dataset, we progressively replace samples within each class with copies of a randomly selected seed while keeping the total dataset size, label distribution, and class balance unchanged, thereby reducing effective information without changing nominal data quantity. Across nine datasets, three modalities, four architecture families, and 13 dataset-backbone pairs spanning ResNet, ViT, BERT, and AST, manifold volume generally decreases as redundancy increases, accompanied by corresponding degradation in downstream fine-tuning performance. More importantly, manifold volume shows a consistently strong positive linear association with downstream accuracy, achieving the highest average Pearson correlation of 0.816 among the compared metrics. These results suggest that the geometry of frozen pretrained representations provides a simple yet informative signal of a dataset's future learning utility before fine-tuning. Our main contributions can be summarized as follows:

- We formulate a new dataset-level quality assessment problem that aims to estimate the future learning utility of an entire candidate dataset before task-specific training, under a label-free and fine-tuning-free setting suitable for practical data acquisition.
- We propose manifold volume in the pretrained representation space as a model-aware dataset-level quality score. Our theoretical analysis establishes its invariance to trivial sample replication, dimensional constraint under severe redundancy, and diminishing returns along well-covered feature directions.
- We conduct a systematic evaluation on 13 dataset-backbone pairs across nine datasets, three modalities, four architecture families, using controlled redundancy benchmarks. Manifold volume shows a strong positive linear association with downstream learning utility, achieving the highest average Pearson correlation of 0.816 among the compared metrics.

2 PROBLEM FORMULATION

We consider a candidate training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where x_i denotes an input sample and y_i is its annotation for downstream training. Although annotations may be available when the dataset is eventually used for model training, our quality-assessment procedure does not access y_i. We therefore denote the observable input collection by $\mathcal{X}_\mathcal{D} = \{x_i\}_{i=1}^n$. Given a pretrained model $f_{\boldsymbol{\theta}_0}$, our goal is to assign a scalar quality score to the entire candidate dataset, $Q(\mathcal{D}; f_{\boldsymbol{\theta}_0})$, such that the score reflects how useful $\mathcal{D}$ will be when it is subsequently used to train or fine-tune the model.

Ideally, the quality of a candidate dataset should be determined by the capability improvement it can induce after model training. Let $\mathcal{A}(f_{\boldsymbol{\theta}_0}, \mathcal{D})$ denote a fixed downstream training procedure initialized from $f_{\boldsymbol{\theta}_0}$, and let $f_{\boldsymbol{\theta}_\mathcal{D}} = \mathcal{A}(f_{\boldsymbol{\theta}_0}, \mathcal{D})$ be the resulting model. Given a held-out evaluation set $\mathcal{D}_{\text{test}}$ and a performance metric $\mathcal{M}$, we define the downstream learning utility of $\mathcal{D}$ as

$$U(\mathcal{D}; f_{\boldsymbol{\theta}_0}) = \mathcal{M}\big(f_{\boldsymbol{\theta}_\mathcal{D}}, \mathcal{D}_{\text{test}}\big).$$

This utility provides a direct measure of how well the model performs after learning on $\mathcal{D}$. However, obtaining $U(\mathcal{D}; f_{\boldsymbol{\theta}_0})$ requires training with repeated optimization, validation, and hyperparameter selection. Therefore, it defeats the purpose of pre-training quality assessment.

We seek a valuation function

$$Q(\mathcal{D}; f_{\boldsymbol{\theta}_0}) = g(\mathcal{X}_\mathcal{D}, f_{\boldsymbol{\theta}_0})$$

that satisfies two constraints. Label-free: computing Q only requires the inputs $\mathcal{X}_\mathcal{D}$, without accessing their annotations. Fine-tuning-free: computing Q does not involve updating model parameters, training a proxy model, or observing optimization trajectories on $\mathcal{D}$.

The desired score should preserve the ordering induced by downstream utility. Given two candidate datasets $\mathcal{D}_a$ and $\mathcal{D}_b$ evaluated under the same pretrained model and downstream protocol, we expect

$$Q(\mathcal{D}_a; f_{\boldsymbol{\theta}_0}) > Q(\mathcal{D}_b; f_{\boldsymbol{\theta}_0}) \quad \Longleftrightarrow \quad U(\mathcal{D}_a; f_{\boldsymbol{\theta}_0}) > U(\mathcal{D}_b; f_{\boldsymbol{\theta}_0}).$$

Accordingly, we treat downstream fine-tuning performance as an evaluation signal rather than an input to the valuation procedure. A successful dataset-quality metric should therefore distinguish informative datasets from low-utility ones, while faithfully predicting which datasets will lead to better downstream models once training is eventually performed.

3 METHOD

To estimate the dataset-level quality $Q(\mathcal{D}; f_{\theta_0})$, our central hypothesis is that the representation space of a pretrained model provides a model-aware view of the effective information contained in a dataset. Informative and non-redundant samples tend to span a broader region of the representation space, whereas duplicated or highly similar samples concentrate in a more restricted subspace. We characterize the candidate dataset as a point-cloud manifold in the pretrained feature space and measure its geometric span through manifold volume. The resulting quality score depends only on the candidate samples' representations and can be computed through a single forward pass of the frozen model. Therefore, the entire valuation procedure requires neither labels, gradient computation, proxy-model training, nor downstream fine-tuning.

3.1 FEATURE MANIFOLD CONSTRUCTION

We use the pretrained model $f_{\boldsymbol{\theta}_0}$ as a frozen feature extractor and denote its representation mapping by $\phi_{\boldsymbol{\theta}_0}$. For each input $x_i \in \mathcal{X}_\mathcal{D}$, we extract a m-dimensional representation $\boldsymbol{z}_i = \phi_{\boldsymbol{\theta}_0}(x_i) \in \mathbb{R}^m$. We compute the empirical feature mean as

$$\boldsymbol{\mu}_Z = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i, \tag{1}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_1^\top, \boldsymbol{z}_2^\top, \ldots, \boldsymbol{z}_n^\top]^\top \in \mathbb{R}^{n \times m}$ and obtain the centered feature matrix

$$\bar{\mathbf{Z}} = \mathbf{Z} - \mathbf{1}_n \boldsymbol{\mu}_Z^\top, \tag{2}$$

where $\mathbf{1}_n$ denotes an n-dimensional all-ones vector. The empirical covariance is then

$$\boldsymbol{C}_{\mathcal{D}} = \frac{1}{n} \bar{\mathbf{Z}}^\top \bar{\mathbf{Z}} \in \mathbb{R}^{m \times m}. \tag{3}$$

Centering removes the global offset of the representation distribution, allowing $\boldsymbol{C}_{\mathcal{D}}$ to characterize how broadly the candidate samples vary around their mean in the pretrained feature space.

3.2 Manifold Volume for Dataset Valuation

We quantify the geometric span of the candidate representations using the volume of their feature manifold. The volume spanned by a set of feature vectors is related to the product of the singular values of the centered feature matrix, or equivalently to the determinant of its covariance matrix. However, deep representations often lie near a low-dimensional manifold embedded in a high-dimensional feature space. Consequently, $\boldsymbol{C}_{\mathcal{D}}$ can be rank deficient, making its direct determinant zero even when substantial variation exists along its intrinsic dimensions.

To obtain a generalized volume that remains well-defined in such cases, we regularize the covariance matrix with an identity matrix and define

$$\mathrm{Vol}(\boldsymbol{Z}) = \frac{1}{2} \log_2 \det \left(\boldsymbol{I}_m + \frac{1}{n} \bar{\mathbf{Z}}^\top \bar{\mathbf{Z}} \right), \tag{4}$$

where $\boldsymbol{I}_m$ is the m-dimensional identity matrix. The identity term makes the matrix positive definite, while the logarithmic transformation avoids the numerical instability caused by directly multiplying a large number of eigenvalues. We instantiate the valuation function introduced in Section 2 as

$$Q(\mathcal{D}; f_{\boldsymbol{\theta}_0}) = \mathrm{Vol}(\boldsymbol{Z}) \tag{5}$$

A larger $Q(\mathcal{D}; f_{\boldsymbol{\theta}_0})$ indicates that the candidate samples collectively span a richer region of the pretrained representation space, whereas a smaller score indicates a more concentrated or degenerate feature distribution. Importantly, the score is determined by the joint geometry of the entire dataset rather than by the nominal number of samples alone. Therefore, adding redundant samples without expanding the effective representation space provides little increase in the estimated dataset quality.

3.3 Theoretical Analysis

We analyze several important properties of manifold volume for dataset-level quality assessment.

Spectral interpretation: Let $\lambda_1, \ldots, \lambda_m \geq 0$ denote the eigenvalues of $\boldsymbol{C}_{\mathcal{D}}$. Since $\boldsymbol{C}_{\mathcal{D}} = \boldsymbol{U} \,\mathrm{diag}(\lambda_1, \ldots, \lambda_m) \boldsymbol{U}^\top$, Eq.(5) can be equivalently written as

$$Q(\mathcal{D}; f_{\boldsymbol{\theta}_0}) = \frac{1}{2} \sum_{j=1}^{m} \log_2(1 + \lambda_j). \tag{6}$$

Eq.(6) shows that the proposed score jointly captures the number of effective feature directions occupied by the dataset and the variation along these directions. When the candidate representations spread across multiple independent directions, more eigenvalues become non-negligible and the resulting manifold volume is larger. In contrast, increasing redundancy concentrates the representation distribution into fewer effective directions, resulting in a more degenerate covariance spectrum.

Replication invariance: **Proposition 1.** Let $\mathcal{D}^{(k)}$ be obtained by repeating every sample in $\mathcal{D}$ exactly k times. Then

$$Q(\mathcal{D}^{(k)}; f_{\boldsymbol{\theta}_0}) = Q(\mathcal{D}; f_{\boldsymbol{\theta}_0}). \tag{7}$$

Proof. Repeating all samples k times leaves the empirical feature mean unchanged. Let $\bar{\mathbf{Z}}^{(k)}$ denote the centered feature matrix of the replicated dataset. Since every row of $\bar{\mathbf{Z}}$ appears k times,

$$(\bar{\mathbf{Z}}^{(k)})^\top \bar{\mathbf{Z}}^{(k)} = k\bar{\mathbf{Z}}^\top \bar{\mathbf{Z}}. \tag{8}$$

Since $\mathcal{D}^{(k)}$ contains kn samples,

$$\frac{1}{kn}(\bar{\mathbf{Z}}^{(k)})^\top \bar{\mathbf{Z}}^{(k)} = \frac{1}{n}\bar{\mathbf{Z}}^\top \bar{\mathbf{Z}} \implies Q(\mathcal{D}^{(k)}; f_{\boldsymbol{\theta}_0}) = Q(\mathcal{D}; f_{\boldsymbol{\theta}_0}). \tag{9}$$

Proposition 1 shows that increasing the nominal dataset size through uniform sample replication cannot artificially increase the proposed quality score. Thus, the metric responds to changes in effective representation geometry rather than to the number of repeated samples.

Redundancy constrains the effective feature dimension: Proposition 2. Suppose the candidate dataset contains only q distinct feature representations. Then

$$\operatorname{rank}(\boldsymbol{C}_{\mathcal{D}}) \le \min(m, q-1). \tag{10}$$

Proof. After mean centering, the q distinct representations span at most $q-1$ independent directions. We have

$$\operatorname{rank}(\boldsymbol{C}_{\mathcal{D}}) = \operatorname{rank}(\frac{1}{n}\bar{\mathbf{Z}}^\top \bar{\mathbf{Z}}) = \operatorname{rank}(\bar{\mathbf{Z}}) \le \min(m, q-1). \tag{11}$$

Therefore, if only q distinct representations remain, at most $q-1$ eigenvalues in Eq.(6) can be non-zero. Severe duplication thus constrains the effective dimensionality of the feature manifold even if the nominal number of samples remains unchanged.

Diminishing returns along covered directions. From Eq.(6), the sensitivity of the quality score to the j-th eigenvalue is

$$\frac{\partial Q}{\partial \lambda_j} = \frac{1}{2\ln 2\,(1+\lambda_j)}. \tag{12}$$

For two feature directions satisfying $\lambda_a < \lambda_b$, we have

$$\frac{\partial Q}{\partial \lambda_a} > \frac{\partial Q}{\partial \lambda_b}. \tag{13}$$

Hence, increasing variation along a weakly represented direction contributes more than adding the same amount of variation along an already dominant direction. This diminishing-return property makes manifold volume sensitive to the coverage of complementary feature directions while limiting the benefit of repeatedly reinforcing already well-covered directions.

The above analysis does not imply that manifold volume uniquely determines downstream performance for arbitrary neural networks or optimization procedures. Instead, it establishes several properties that make manifold volume a meaningful proxy for dataset-level learning utility: it measures the effective spread of pretrained representations, is invariant to trivial replication, is constrained by the number of distinct feature patterns, and exhibits diminishing returns along already covered representation directions. Accordingly, our objective is not to predict the exact downstream performance from the manifold volume. As formulated in Section 2, we instead expect the proposed score to preserve the relative quality ordering of candidate datasets under a pretrained model.

Table 1: Datasets and backbones used in our experiments.

Modality	Dataset	Backbone	Arch.	Modality	Dataset	Backbone	Arch.
Image	CIFAR-10	ResNet-18	CNN	Text	AG News	BERT-base	BERT
	CIFAR-100	ResNet-18	CNN		THUCNews	BERT-base	BERT
	ImageNet-1K	ResNet-50	CNN		TNews	BERT-base	BERT
		ViT-B/16, ViT-L/14	ViT	Audio	ESC-50	AST	AST
	Places365	ResNet-50	CNN		GSC	AST	AST
		ViT-B/16, ViT-L/14	ViT				

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and backbones. We evaluate the proposed dataset quality score on nine datasets spanning three modalities. The image benchmarks include CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), ImageNet-1K (Deng et al., 2009), and Places365 (Zhou et al., 2018); the text benchmarks include AG News (Zhang et al., 2015), THUCNews (Tsinghua NLP Lab, 2016), and TNews (Xu et al., 2020); and the audio benchmarks include ESC-50 (Piczak, 2015) and Google Speech Commands (GSC) (Warden, 2018). We consider six backbones from four architecture families, including ResNet-18 and ResNet-50 (He et al., 2016), ViT-B/16 and ViT-L/14 (Dosovitskiy et al., 2021), BERT-base (Devlin et al., 2019), and AST (Gong et al., 2021). The resulting 13 dataset-backbone pairs are summarized in Table 1. We follow the standard train/test splits. For the ResNet experiments, the backbone is pretrained on one half of the training set, while candidate datasets are constructed from the remaining half. For ViT-B/16, ViT-L/14, BERT-base, and AST, we use publicly available pretrained checkpoints and construct candidate datasets from the full training split.

Controlled redundancy construction. To obtain candidate datasets with progressively reduced effective information while keeping their nominal size unchanged, we inject controlled intra-class redundancy. Specifically, for each redundancy ratio $r \in \{0, 20, 40, 60, 80, 100\}\%$, we randomly select one sample within each class and replace $r\%$ of the samples in that class with copies of the selected sample. Note that class labels are only used to construct these controlled benchmark variants and never accessed when computing the proposed quality score.

Dataset quality score. For each candidate dataset, we freeze the corresponding pretrained backbone and extract representations for each sample in the candidate dataset. For ResNet, we use the feature output before the classification head. For BERT, AST, and ViT, we use the final `[CLS]` representation. Then, we compute the manifold volume (Vol) defined in Eq. (4). No additional feature normalization, projection, or task-specific transformation is applied.

Downstream learning utility. To evaluate whether manifold volume predicts future learning utility, we fine-tune each pretrained backbone on every redundancy-corrupted candidate dataset and measure its test accuracy on the corresponding held-out evaluation set (Acc). Fine-tuning is performed only to obtain the downstream utility. All models for a given dataset-backbone pair share the same initialization and are equipped with a randomly initialized classification head. We follow the standard optimization settings of the corresponding backbone for downstream fine-tuning.

Comparison metrics. We compare with three existing dataset quality metrics: Vendi Score (Vendi) (Friedman & Dieng, 2023), which quantifies sample diversity through the spectral entropy of a similarity matrix, and the diversity (Div) and relevance (Rel) (Amiri et al., 2023), which characterize statistical differences between candidate and reference data using second-order statistics.

Evaluation protocol. Our primary goal is to determine whether the proposed score preserves the ordering induced by downstream learning utility. We therefore use Spearman's rank correlation coefficient ρ_s as the primary evaluation metric between each metric and downstream test accuracy across the six redundancy levels of each dataset-backbone pair. We additionally report Pearson's correlation coefficient ρ_r as a secondary measure of linear association.

4.2 Experimental Results

We first consider a practical data-acquisition scenario in which a model developer already owns an initial model f_{θ_0} trained on its internal data and seeks to acquire an additional candidate dataset $\mathcal{D}$ to further improve this model through fine-tuning. In this setting, the initial model serves two roles: it provides the frozen representation space for assessing the quality of $\mathcal{D}$, and is subsequently fine-tuned on $\mathcal{D}$ to obtain the upgraded downstream model. Using the controlled redundancy construction, we examine whether manifold volume captures the effective information available to the initial model and whether it can predict the downstream learning utility of the candidate dataset. The results are presented in Table 2, from which we draw the following observations.

Impact of effective information. Table 2 shows that increasing redundancy generally reduces Vol, Vendi, and Rel, while Acc also decreases. Although the dataset size and class balance remain unchanged, replacing distinct samples with repeated instances reduces the amount of non-redundant information available for learning. For example, as the redundancy ratio increases from 0% to 100%, the accuracy on CIFAR-100 with ResNet-18 drops from 67.15% to 53.75%, accompanied by a decrease in Vol from 57.59 to 45.74. Similar reductions are observed across the other dataset-backbone pairs. These results indicate that dataset size alone does not determine downstream utility: the effective information contained in distinct samples also matters.

Correlation with downstream accuracy. Figure 1 illustrates the normalized trends of the four metrics and downstream accuracy for three representative dataset-backbone pairs. Vol decreases alongside Acc as redundancy increases, including in cases where accuracy changes nonlinearly with the redundancy ratio. The correlation results in Table 2 further quantify this relationship. Across all 13 dataset-backbone pairs, Vol achieves a Spearman correlation of at least 0.943, reaching 1.000 in eight pairs. Its Pearson correlation ranges from 0.585 to 0.974, indicating a consistently positive association with downstream accuracy across the evaluated redundancy levels. For the three pairs shown in Figure 1, the Pearson correlations of Vol with Acc are 0.974, 0.760, and 0.939, respectively. Vendi and Rel also exhibit strong positive rank correlations, whereas Div exhibits strong negative rank correlations. Thus, the controlled redundancy experiment reveals a shared monotonic response among several metrics, while their Pearson correlations distinguish how closely their numerical changes track changes in downstream accuracy.

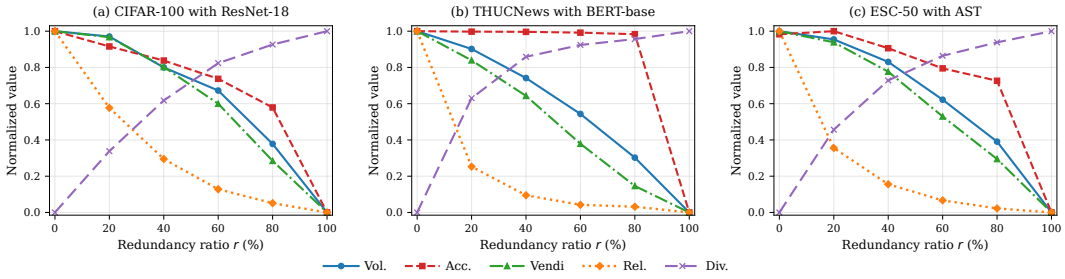

Figure 1: Normalized metric trends under increasing redundancy for three dataset-backbone pairs: CIFAR-100/ResNet-18, THUCNews/BERT-base, and ESC-50/AST.

Overall performance. Averaged across the 13 dataset-backbone pairs, Vol, Vendi, and Rel achieve Spearman correlations of 0.978, 0.982, and 0.987, respectively, while Div achieves -0.987. These results indicate that all four metrics exhibit strong monotonic associations with downstream accuracy under controlled redundancy, although the direction of association differs for Div. In terms of Pearson correlation, Vol achieves the highest macro-average value of 0.816, compared with 0.801 for Vendi, 0.500 for Rel, and -0.575 for Div. Moreover, Vol exhibits the largest positive Pearson correlation in seven dataset-backbone pairs and the second-largest in the remaining six. Together, these results demonstrate that Vol maintains a consistently high positive linear correlation with downstream accuracy across the evaluated dataset-backbone pairs.

Table 2: Vendi, Div and Rel, Vol, Acc(%), and their correlations under different redundancy ratios r. The overall performance is computed by averaging the correlations across all dataset-backbone pairs. The best and sub-optimal values are highlighted in bold and underlined, respectively.

Dataset	Backbone	Metric	Redundancy ratio r (%)						Spearman	Pearson
			0	20	40	60	80	100	ρ_s	ρ_r
CIFAR-10	ResNet-18	Vendi	8.55	8.23	7.59	7.32	6.59	5.75	1.000	**0.861**
		Rel	1.00	0.12	0.06	0.03	0.02	0.02	1.000	0.362
		Div	0.00	0.76	0.84	0.91	0.93	0.93	-1.000	-0.407
		Vol	9.96	9.88	9.68	8.69	8.33	8.07	1.000	0.737
		Acc	91.22	90.83	90.49	90.17	89.34	82.59	–	–
CIFAR-100	ResNet-18	Vendi	33.16	32.74	30.65	28.07	24.06	20.44	1.000	0.949
		Rel	1.00	0.67	0.45	0.32	0.26	0.22	1.000	0.728
		Div	0.00	0.23	0.42	0.56	0.63	0.68	-1.000	-0.773
		Vol	57.59	57.24	55.22	53.71	50.22	45.74	1.000	**0.974**
		Acc	67.15	66.02	64.98	63.63	61.51	53.75	–	–
ImageNet-1K	ResNet-50	Vendi	39.16	38.84	38.67	37.90	36.50	35.01	1.000	**0.890**
		Rel	1.00	0.64	0.41	0.31	0.24	0.20	1.000	0.549
		Div	0.00	0.20	0.40	0.51	0.59	0.65	-1.000	-0.630
		Vol	176.81	176.25	175.38	173.60	169.02	167.67	1.000	0.791
		Acc	73.27	72.43	71.90	71.14	69.51	51.92	–	–
ImageNet-1K	ViT-B/16	Vendi	25.05	24.99	24.86	24.67	23.77	24.06	0.943	0.536
		Rel	1.00	0.94	0.84	0.75	0.67	0.59	1.000	**0.718**
		Div	0.00	0.02	0.08	0.14	0.21	0.27	-1.000	-0.752
		Vol	31.78	31.75	31.73	31.72	31.45	31.51	0.943	0.585
		Acc	79.56	78.42	77.34	76.31	73.62	19.34	–	–
ImageNet-1K	ViT-L/14	Vendi	32.23	32.21	31.77	31.12	30.17	29.47	1.000	0.846
		Rel	1.00	0.66	0.54	0.46	0.40	0.35	1.000	0.597
		Div	0.00	0.04	0.11	0.20	0.28	0.35	-1.000	-0.792
		Vol	89.58	89.45	89.04	88.49	87.40	86.65	1.000	**0.857**
		Acc	83.64	82.66	82.13	81.14	79.16	63.11	–	–
Places365	ResNet-50	Vendi	23.57	23.51	22.81	22.34	20.62	19.55	1.000	**0.848**
		Rel	1.00	0.42	0.25	0.17	0.13	0.11	1.000	0.461
		Div	0.00	0.40	0.57	0.69	0.75	0.79	-1.000	-0.543
		Vol	91.24	90.25	90.65	87.84	85.38	83.03	0.943	0.843
		Acc	54.92	54.31	53.92	53.28	51.85	38.00	–	–
Places365	ViT-B/16	Vendi	27.33	27.30	26.55	25.46	24.98	22.43	1.000	**0.925**
		Rel	1.00	0.87	0.71	0.59	0.48	0.42	1.000	0.698
		Div	0.00	0.06	0.17	0.28	0.39	0.46	-1.000	-0.752
		Vol	31.87	31.86	31.66	31.34	31.28	30.70	1.000	0.902
		Acc	57.41	54.46	52.94	52.43	51.65	28.75	–	–
Places365	ViT-L/14	Vendi	35.04	34.84	33.87	31.73	30.39	26.23	1.000	**0.903**
		Rel	1.00	0.57	0.43	0.34	0.28	0.24	1.000	0.597
		Div	0.00	0.10	0.25	0.36	0.47	0.54	-1.000	-0.721
		Vol	87.42	87.05	86.14	84.18	82.37	78.79	1.000	0.889
		Acc	58.95	54.75	54.44	54.24	53.26	33.72	–	–
AG News	BERT-base	Vendi	7.09	6.53	5.39	4.01	3.12	2.34	1.000	0.660
		Rel	1.00	0.18	0.11	0.08	0.05	0.03	1.000	0.302
		Div	0.00	0.73	0.84	0.87	0.92	0.95	-1.000	-0.346
		Vol	32.56	30.23	25.60	19.86	13.81	6.52	1.000	**0.764**
		Acc	94.25	94.08	93.55	93.14	92.16	58.64	–	–
THUCNews	BERT-base	Vendi	10.94	9.85	8.52	6.73	5.15	4.16	1.000	0.635
		Rel	1.00	0.29	0.14	0.09	0.08	0.05	1.000	0.309
		Div	0.00	0.58	0.79	0.85	0.88	0.92	-1.000	-0.362
		Vol	73.77	68.60	60.14	49.71	37.01	21.07	1.000	**0.760**
		Acc	97.17	97.06	97.01	96.82	96.45	53.89	–	–
TNews	BERT-base	Vendi	8.33	7.78	7.03	5.82	4.52	3.59	0.943	0.710
		Rel	1.00	0.31	0.14	0.12	0.10	0.07	0.943	0.330
		Div	0.00	0.56	0.78	0.83	0.86	0.90	-0.943	-0.380
		Vol	67.04	63.37	56.27	47.07	35.00	21.78	0.943	**0.777**
		Acc	66.56	66.00	66.16	65.32	64.58	29.38	–	–
ESC-50	AST	Vendi	56.66	54.67	49.40	41.37	33.74	24.16	0.943	0.898
		Rel	1.00	0.42	0.24	0.16	0.12	0.10	0.943	0.525
		Div	0.00	0.37	0.59	0.70	0.76	0.81	-0.943	-0.625
		Vol	218.36	212.91	198.02	173.02	145.26	98.37	0.943	**0.939**
		Acc	94.00	94.50	91.75	88.50	86.50	65.25	–	–
GSC	AST	Vendi	5.61	5.39	5.04	4.60	4.11	3.49	0.943	0.753
		Rel	1.00	0.34	0.19	0.13	0.10	0.08	0.943	0.326
		Div	0.00	0.49	0.69	0.79	0.84	0.87	-0.943	-0.393
		Vol	110.10	105.31	96.13	83.86	68.37	47.32	0.943	**0.788**
		Acc	96.07	96.22	95.55	95.27	94.08	7.41	–	–
Overall Performance		Vendi	–	–	–	–	–	–	0.982	0.801
		Rel	–	–	–	–	–	–	0.987	0.500
		Div	–	–	–	–	–	–	-0.987	-0.575
		Vol	–	–	–	–	–	–	0.978	**0.816**

Table 3: Cross-model dataset quality assessment and downstream learning utility.

Dataset	Models (Val. → Train.)	Metric	Redundancy ratio r (%)					
			0	20	40	60	80	100
CIFAR-100	ResNet-18 → ResNet-50	Vol	57.59	57.24	55.22	53.71	50.22	45.74
		Initial Acc	61.52	61.52	61.52	61.52	61.52	61.52
		Internal Acc	62.10	62.10	62.10	62.10	62.10	62.10
		Joint Acc	72.63	70.49	70.74	68.32	64.74	59.40
		Gain	10.53	8.39	8.64	6.22	2.64	-2.70

4.3 Generalization to Cross-Model Utility Prediction

We further consider a more challenging data-acquisition scenario in which a model developer seeks to acquire additional data not to fine-tune the existing model, but to train a new model with a larger capacity. Specifically, the developer owns an internal dataset $\mathcal{D}_0$ and an initial model $f_{\boldsymbol{\theta}_0}$ trained on $\mathcal{D}_0$. The initial model is used to evaluate the quality of a candidate dataset $\mathcal{D}$ before acquisition, while the acquired data will subsequently be combined with $\mathcal{D}_0$ to train a larger model $F_{\boldsymbol{\phi}}$. This setting allows us to investigate whether manifold volume captures effective information that remains useful beyond the specific model used for valuation.

We instantiate this setting on CIFAR-100, using ResNet-18 as $f_{\boldsymbol{\theta}_0}$ and ResNet-50 as $F_{\boldsymbol{\phi}}$. For each candidate dataset with redundancy ratio $r \in \{0, 20, 40, 60, 80, 100\}\%$, we compute Vol from the frozen ResNet-18 representations. We then train ResNet-50 either on the fixed internal dataset $\mathcal{D}_0$ alone or on $\mathcal{D}_0 \cup \mathcal{D}$, yielding *Internal Acc* and *Joint Acc*, respectively. We additionally report the test accuracy of the initial ResNet-18 model as *Initial Acc*. The learning utility contributed by the candidate dataset is measured by *Gain*, i.e., the difference between *Joint Acc* and *Internal Acc*, where positive and negative values indicate performance improvement and degradation, respectively.

As shown in Table 3, we can see that (1) Vol estimated by the frozen ResNet-18 exhibits a trend consistent with the downstream Gain obtained by ResNet-50. As Vol decreases from 57.59 to 45.74, Gain generally decreases from 10.53% to -2.70%. This consistency suggests that Vol estimated by the initial model can reflect the learning utility of candidate data for a different downstream model. (2) Comparing the two models trained on the same internal dataset further illustrates the benefit of increased model capacity. Under the same training data, the larger ResNet-50 achieves a higher accuracy (62.10%) than the ResNet-18 initial model (61.52%). (3) Compared with training ResNet-50 on the internal dataset alone, incorporating effective candidate data can further improve model performance. At $r = 0\%$, Joint Acc reaches 72.63%, which is 10.53% higher than the Internal Acc of 62.10%. As redundancy increases, this additional benefit gradually decreases, with Gain dropping to 2.64% at $r = 80\%$. When $r = 100\%$, Joint Acc decreases to 59.40%, even lower than Internal Acc. These results indicate that effective data growth can benefit the larger model, whereas excessively redundant data may provide little additional utility or even degrade performance.

5 Conclusion

In this work, we have studied label-free and fine-tuning-free dataset-level quality assessment to estimate the future learning utility of an entire candidate dataset before training on it. We have proposed manifold volume as a simple model-aware criterion that characterizes the effective information of a dataset through the geometry of frozen pretrained representations. Our theoretical analysis has shown that manifold volume is invariant to trivial sample replication, becomes dimensionally constrained under severe redundancy, and exhibits diminishing returns along already well-covered feature directions, providing a geometric basis for distinguishing effective information from nominal dataset size. Through controlled redundancy experiments on 13 dataset-backbone pairs, we have shown that manifold volume consistently exhibits a strong positive linear association with downstream learning utility, achieving the highest average Pearson correlation of 0.816 among the compared metrics. We have further observed that the quality estimated by a smaller valuation model is consistent with the learning gain obtained by a larger downstream model, providing initial evidence for cross-model utility assessment. In the future, we will examine this principle under more diverse candidate datasets, model pairs, and data acquisition settings.

AI USE STATEMENT

The authors used AI tools for language polishing and LaTeX formatting assistance. The authors are responsible for the method, experimental design, reported numbers, interpretations, and final claims.

REFERENCES

Mohammad Mohammadi Amiri, Frédéric Berdoz, and Ramesh Raskar. Fundamentals of task-agnostic data valuation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. Doge: Domain reweighting with generalization estimation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, and Lingzhong Meng. A survey on dataset quality in machine learning. *Information and Software Technology*, 162:107268, 2023. doi: 10.1016/j.infsof.2023.107268. URL https://doi.org/10.1016/j.infsof.2023.107268.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Proceedings of Interspeech*, pp. 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

Sachin Goyal, Pratyush Maini, Zachary C. Lipton, Aditi Raghunathan, and J. Zico Kolter. Scaling laws for data filtering—data curation cannot be compute agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Yiding Jiang, Allan Zhou, Zhili Feng, Sadhika Malladi, and J. Zico Kolter. Adaptive data optimization: Dynamic sample selection with scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025.

Nohyun Ki, Hoyong Choi, and Hye Won Chung. Data valuation without training of a model. In *International Conference on Learning Representations (ICLR)*, 2023.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022. doi: 10.18653/v1/2022.acl-long.577.

Sören Mindermann, Jan M. Brauner, Muhammed T. Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Karol J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1521–1528, 2011. doi: 10.1109/CVPR.2011.5995347.

Tsinghua NLP Lab. THUCTC: An efficient chinese text classifier and the THUCNews dataset. `https://github.com/thunlp/THUCTC`, 2016.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. Clue: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 4762–4772, 2020.

Siqi Zeng, Christopher Jung, Rui Li, Zhe Kang, Ming Li, Nima Noorshams, Zhigang Wang, Fuchun Peng, Han Zhao, and Xue Feng. Convex dataset valuation for post-training. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=oAAL0wZSMF`.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.