A Medical Data-Effective Learning Benchmark for Highly Efficient Pre-training of Foundation Models

Anonymous Authors

ABSTRACT

Foundation models, pre-trained on massive datasets, have achieved unprecedented generalizability. However, is it truly necessary to involve such vast amounts of data in pre-training, consuming extensive computational resources? This paper introduces data-effective learning, aiming to use data in the most impactful way to pre-train foundation models. This involves strategies that focus on data quality rather than quantity, ensuring the data used for training has high informational value. Data-effective learning plays a profound role in accelerating foundation model training, reducing computational costs, and saving data storage, which is very important as the volume of medical data in recent years has grown beyond many people's expectations. However, due to the lack of standards and comprehensive benchmark, research on medical data-effective learning is poorly studied. To address this gap, our paper introduces a comprehensive benchmark specifically for evaluating data-effective learning in the medical field. This benchmark includes a dataset with millions of data samples from 31 medical centers (DataDEL), a baseline method for comparison (MedDEL), and a new evaluation metric (NormDEL) to objectively measure data-effective learning performance. Our extensive experimental results show the baseline MedDEL can achieve performance comparable to the original large dataset with only 5% of the data. Establishing such an open data-effective learning benchmark is crucial for the medical foundation model research community because it facilitates efficient data use, promotes collaborative breakthroughs, and fosters the development of cost-effective, scalable, and impactful healthcare solutions.

KEYWORDS

Medical benchmark, Data-Effective learning, Endoscopic Image Processing

1 INTRODUCTION

The effectiveness of foundation models depends on the abundance of pre-training data, a notion seemingly supported by consensus: more pre-training data leads to enhanced model performance. However, is this assumption truly accurate? To explore this issue, we introduce the concept of data-effective learning, which plays a significant role in the field of medical data, accelerating foundation model training and reducing storage burden in the era of big data

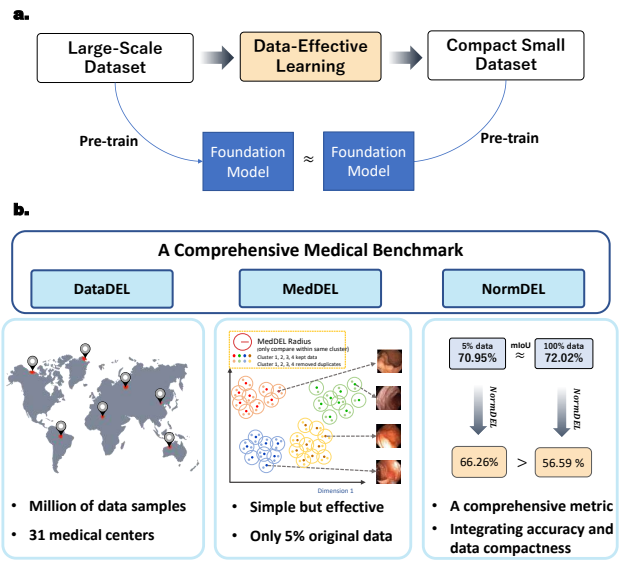


Figure 1: Data-Effective Learning (DEL) enables more efficient pre-training of foundational models. (a) Data-effective learning aims to obtain a compact small dataset from a large-scale pre-training dataset, but the two datasets have similar effects on Foundation model pre-training. (b) Demonstration of our comprehensive benchmark for data-effective learning. The benchmark includes a dataset of millions of data samples from 31 medical centers (DataDEL), a baseline method for comparison (MedDEL), and a new evaluation metric (NormDEL).

(Figure 1(a)). The global endoscopy surgery market is undergoing significant development, with an estimated daily addition of 22,546,800,000 video frames [1]. However, a substantial presence of disruptive and invalid data [44] significantly hampers training efficiency and occupies a considerable amount of storage space [59]. Therefore, achieving data-effective in endoscopy datasets holds the following special advantages:

- **Storage Savings**: Assuming the use of traditional high-definition endoscopes (1080p) [25], the daily uncompressed endoscopy examination videos would require 12,756,493 TB (about 13.06 exabytes) of storage space. Condensing endoscopy datasets can result in substantial storage savings.
- **Enhanced Model Efficiency**: [8] released the largest digestive system image dataset (Hyper-Kvasir) to date and showed that upon analyzing data sources, over 90% of video frames consist of disruptive and invalid data. Core critical data comprises only 2% of the entire dataset [29, 30]. Efficient utilization of core critical data can significantly improve

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACM MM, 2024, Melbourne, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Overview of the proposed DataDEL in our benchmark. We integrate 31 medical centers, including data on the scale of millions, to build a high-quality, large-scale, multi-disease comprehensive dataset, aiming to provide researchers with a better data platform.

Division	Dataset	Task	Medical centers	Videos	Frames	Disease
Model Pre-training	Gastrovision [28]	Detection and classification	2	None	8,000	GI
	Hyper-Kvasir [8]	Classification	1	374	110,079	GI
	Kvasir-Capsule [51]	Classification	1	117	47,238	GI
Model downstream testing	CVC-12k (CVC-ClinicDB) [6]	Segmentation	1	None	612	polyp
	CVC-300 [54]	Segmentation	Not mentioned	None	60	polyp
	CVC-ColonDB [7]	Segmentation	Not mentioned	None	380	polyp
	EAD2019 [5]	Endoscopic artifacts	6	None	2,991	cancers
	EDD2020 [3]	Segmentation, detection, and localization	4	None	380	polyp
	ETIS [50]	Segmentation	Not mentioned	None	196	polyp
	ImageCLEFmed [26]	Polyp segmentation	1	None	66,662	polyp
	Kvasir-Instrument [27]	Segmentation, detection, and localization	1	None	590	polyp
	Kvasir-SEG [30]	Segmentation, detection, and localization	4	None	1,000	polyp
	Kvasir-Sessile [29]	Segmentation, detection, and localization	4	None	196	polyp
	PolypGen2021 [4]	Polyp segmentation	6	None	3,762	polyp

the efficiency of model training, in order to achieve rapid convergence.

- **Computational Resource Savings**: With the use of a single RTX 3090 graphics card and the VGG16 model [32] (approximately 138 million parameters) for 32-bit floating-point (FP32) calculations, processing 325.6 images per second is achievable [34]. In this configuration, training on the daily added video frames would require 19,200 hours. Utilizing only core critical data could save nearly 18,816 hours without compromising precision.

The issue of data-effective learning is not exclusive to endoscopy datasets; it is also evident in other types of medical datasets. With the rapid expansion of future medical data, efficiently handling medical datasets is the next crucial research problem in data-driven learning methods[46].

In recent years, there has been relevant research on data-effective learning in natural image datasets such as adversarial samples [19], and data biases [55], aiming to enhance the robustness and generalization capabilities of deep learning models. These works in natural datasets have shown preliminary effectiveness, indicating significant potential advantages such as data storage, computational resource savings, and efficient model training. However, despite the exponential growth of medical datasets, there is currently a lack of such exploration in the medical field [57]. This is attributed to the absence of relevant benchmarks, encompassing unified datasets, baseline methods, and comprehensive evaluation metrics, providing the academic community with a basis for developing and comparing advanced methods.

In addition to the absence of benchmark, achieving data-effective in the field of endoscopy remains a huge challenge [49] and is worth exploring in depth. Traditional data-effective techniques may struggle to work effectively with datasets that have high-dimensional features or sparse data. In endoscopy datasets, the challenge lies more in identifying subtle differences in images, which can be crucial for medical diagnosis. Therefore, accurately determining the similarity of images in endoscopy datasets is a top priority in research. Maintaining data quality and integrity is a key issue in data-effective learning [17], as it directly affects the usability and accuracy of the dataset.

Our research contributions can be summarized as follows:

- We introduce the concept of data-effective learning and provide a corresponding medical benchmark (Figure 1(b)) to guide data-effective algorithm research in the medical field. Furthermore, we integrate an open-source dataset called DataDEL, sourced from a million-level dataset spanning 31 medical centers.
- In our benchmark, we introduce a baseline method called MedDEL for data-effective learning, which can outperform the use of 100% of the data in downstream tasks with 5% of the pretraining data in extreme cases.
- We develop a new metric called NormDEL, to assess the performance of data-effective in datasets, which considers the relationship between the proportion of the dataset retained and the performance of downstream tasks.

2 RELATED WORK

In the natural image domain, benchmarks for assessing data-effective learning typically involve various techniques such as structural similarity index [16], convolutional neural networks [38], and local feature descriptors [35]. These methods [43, 56] are evaluated on widely used image datasets like ImageNet [14, 48] and CIFAR [33], or specialized datasets. PXDedup [58] explores the issue of data-effective learning in JPEG images, pointing out that traditional binary stream-based techniques do not work well for compressed JPEG images. [9] introduces a high-precision image data effective method that identifies and eliminates duplicate images through feature extraction and high-dimensional indexing [31]. [10] uses wavelet decomposition [42] to extract feature vectors from images and calculate the Manhattan distance to determine image similarity, thus achieving the detection and removal of duplicate images. However, when handling datasets with high-dimensional features [47] or sparse data [58], traditional data-effective techniques might not work effectively. MFDedup [60] surpasses existing technologies in

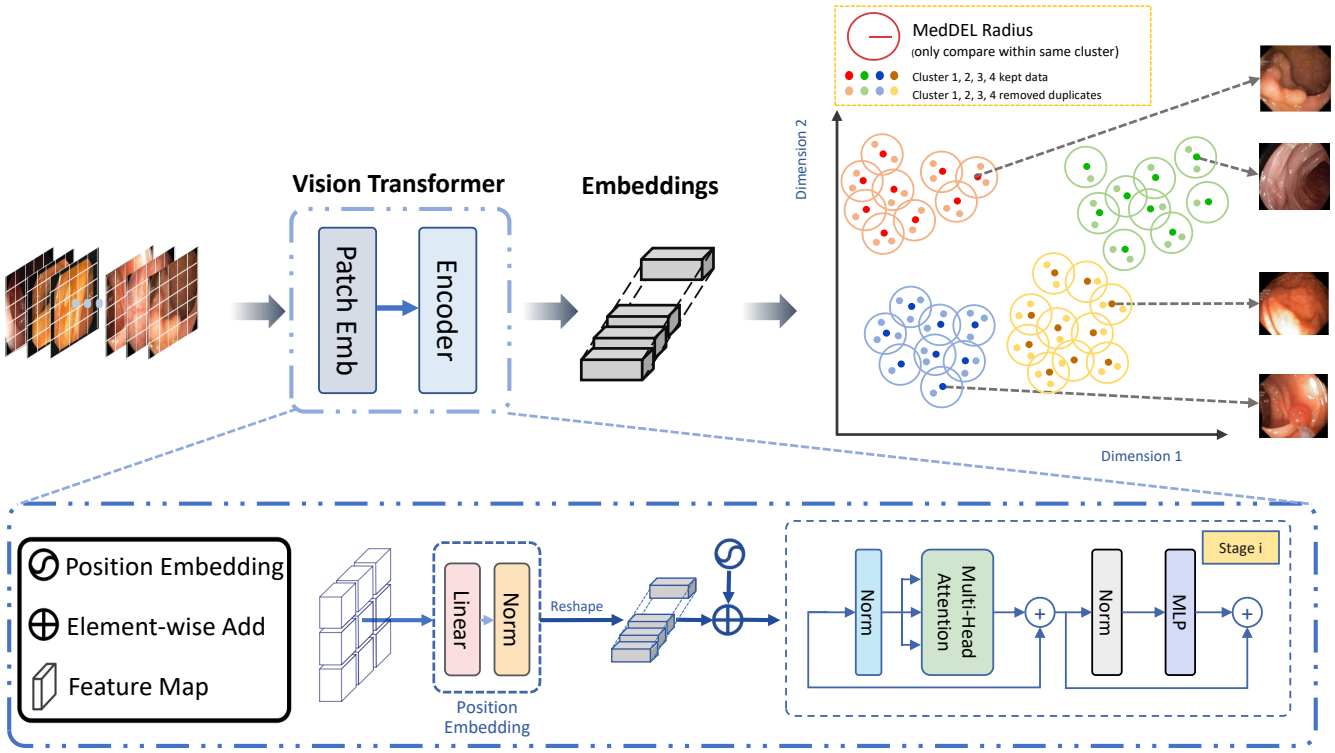


Figure 2: Pipeline of the baseline method (MedDEL) for data-effective learning in our benchmark. It illustrates effective removal of disruptive and invalid data from the dataset, aiming to save storage space and computational resources while enhancing model efficiency.

improving data reduction rates and recovery throughput, while also reducing the cost of garbage collection. Storage and memory capacity [13] can also become limiting factors. [23] discusses how to accurately estimate the data reduction ratio achieved through data effective and compression techniques for specific datasets. Hash methods [45] indeed have a wide application in data effective processing. The CE-Dedup [37] framework combines hash-based image data-effective techniques with deep learning image classification tasks. By adjusting the deduplication threshold, effectively balances the trade-off between data reduction rate and model accuracy [20]. [39] proposes a deep supervised hash algorithm for image retrieval by providing the model with pairs of images labeled as similar/dissimilar.

However, in the medical field, there currently exists no such benchmark for the research of data-effective learning algorithms.

3 DATA-EFFECTIVE LEARNING MEDICAL BENCHMARK

In this section, we introduce the definition of benchmark tasks. Data-effective learning refers to the practice of using a limited amount of data for the pre-training phase [18]. Therefore, this benchmark encourages researchers to develop advanced data-effective learning methods to generate a compact version of the originally collected large-scale dataset, thereby obtaining a compact small-scale new dataset for pre-training foundational models.

In the era of big data, large models often require massive pre-training data [11]. However, our perspective challenges this conventional thinking, and we have undertaken the following work to explore this issue in depth.

Our benchmark consists of three parts, a dataset with millions of data samples (DataDEL), a baseline method for comparison (MedDEL) and a new evaluation metric (NormDEL).

3.1 The benchmark dataset: DataDEL

Given the urgent demand for a comprehensive benchmark in the medical field, we are facing the pressing task of integrating diverse large datasets. Currently, prevailing challenges within existing medical datasets encompass issues like the uniformity of data sources, the standardization of task execution, the limited scope of covered disease types, imbalances in dataset categories, and the uniformity of modalities [36].

These limitations hinder the comprehensive development of medical research. Therefore, there is an urgent need for us to construct a comprehensive medical dataset. Our efforts focus on integrating datasets from over 31 different centers and 23 different countries, spanning multiple modalities such as images and videos. Furthermore, our dataset surpasses the million-scale mark, becoming a large-scale collection that encompasses multiple tasks, modalities, sources, and diseases. This dataset can provide rich and diverse

support for subsequent research, offering significant convenience for healthcare professionals and research institutions.

3.2 The benchmark baseline: MedDEL

We introduce a benchmark baseline method (MedDEL) in the endoscopic medical field, which is based on the principles of the SemDeDup method [2]. Endoscopic data holds significant importance in medical diagnosis and treatment. However, due to various factors such as organ morphological changes [52], lighting conditions [12], and noise interference [22], the analysis of this data becomes complex and challenging.

Algorithm 1 Pseudo Code for MedDEL

Input: Image sequence $\{I_1, I_2, \ldots, I_n\}$
Parameter: Thresholds ϵ and η

```
Let t = 0
while t < max_iterations do
    for cluster_i ∈ clusters do
        for j = 1 to size(cluster_i) do
            if dis(F_{x,j}, centers_i) > ε then
                Delete I_j
            else
                for k = j to Size(cluster_i) do
                    if cos(F_j, F_k) > η then
                        if dis(F_j, C_i) > dis(F_k, C_i) then
                            Delete I_k
                        else
                            Delete I_j
                        end if
                    end if
                end for
            end if
        end for
    end for
    Increment iteration counter: t ← t + 1
end while
```

Specifically, the MedDEL method takes an image I as an ith example, the encoder of Vision Transformer (ViT) [15] extracts deep features from the image, ultimately producing a 768-dimensional feature output. This output serves as our information embedding. The high-dimensional representation not only captures the semantic information of the image more effectively but also addresses semantic repetitiveness issues that are challenging to resolve in low-dimensional spaces. Considering the encoding process of image I through the ViT model, we can represent it with the following formula:

$$F_x = ViT(I) \tag{1}$$

Where x represents the final layer of the ViT encoder.

After obtaining the feature embeddings for each image in the dataset, we can explore the similarity between images in a high-dimensional space. Let the embedding corresponding to the ith image be denoted as $F(x, i)$, representing all image features as $\{F(x, 1), F(x, 2), \ldots, F(x, n)\}$. For typical data-effective methods, we have to calculate the similarity between any two pairs of embeddings, resulting in an overall time complexity of $O(n^2)$. Taking our proposed upstream dataset as an example, which includes 88,282 images, the total number of calculations reaches 744 million.

However, by introducing the K-means algorithm [41], the overall computational complexity is reduced from $O(n^2)$ to $O(N^2/k)$, where k is the number of clusters. After applying the K-means algorithm, we obtain k clusters. Assuming that the feature $F(x, i)$ belongs to the k cluster, we can derive the following k sequences:

$$C_k = \{F(x, i) \mid i \in N^*, F(x, i) \in cluster_k\} \tag{2}$$

In each cluster, data points that are far from the cluster centroid are considered disruptive or invalid data. To filter out interference and invalid data, we introduce a predefined threshold ϵ. Specifically, in each cluster C_j we define the distance from image i to the cluster centroid as d_{ij}. By using the following inequality, we classify image i as potentially affected by disruptive or invalid data.

Specifically, we introduce a similarity threshold η, which is used to define semantic duplicate information. In category j, we perform similarity calculations for each pair of embeddings using the following formula:

$$S(p, q) = cos(F(x, p), F(x, q)) \tag{3}$$

Through this similarity calculation, we retain a set of features that are closer to the cluster center, achieving the goal of filtering effective and core-important data. The pseudocode for a simplified algorithm regarding data effectiveness is presented in Algorithm 1.

Table 2: Experimental setting for the parameter of η in Algorithm 1 that controls the data remaining ratio. We set several different ratios to fully assess the performance of MedDEL under various proportions of remaining data. Additionally, we also calculate the number of training epochs required for different ratios of data at the same computational power for fair comparison.

η	Remaining data	ratio	Epochs
1.0	88,282	100%	200
0.9	43,484	50%	400
0.85	28,352	33%	600
0.8	16,789	20%	1,000
0.75	9,055	10%	2,000
0.7	4,461	5%	4,000

3.3 The benchmark evaluation metric: NormDEL

Currently, there is no comprehensive objective metric for evaluating the performance of data-effective learning algorithms, where comprehensive refers to integrating both the test accuracy of downstream tasks and the compactness of pre-training data. Hence, We propose a new data-effective metric, the Normalized Data-Effective Learning Index (NormDEL). Our objective is that within the same task, if a model can achieve the same level of performance in downstream tasks using fewer pre-training data, its data-effective performance should be considered superior. Taking the commonly used

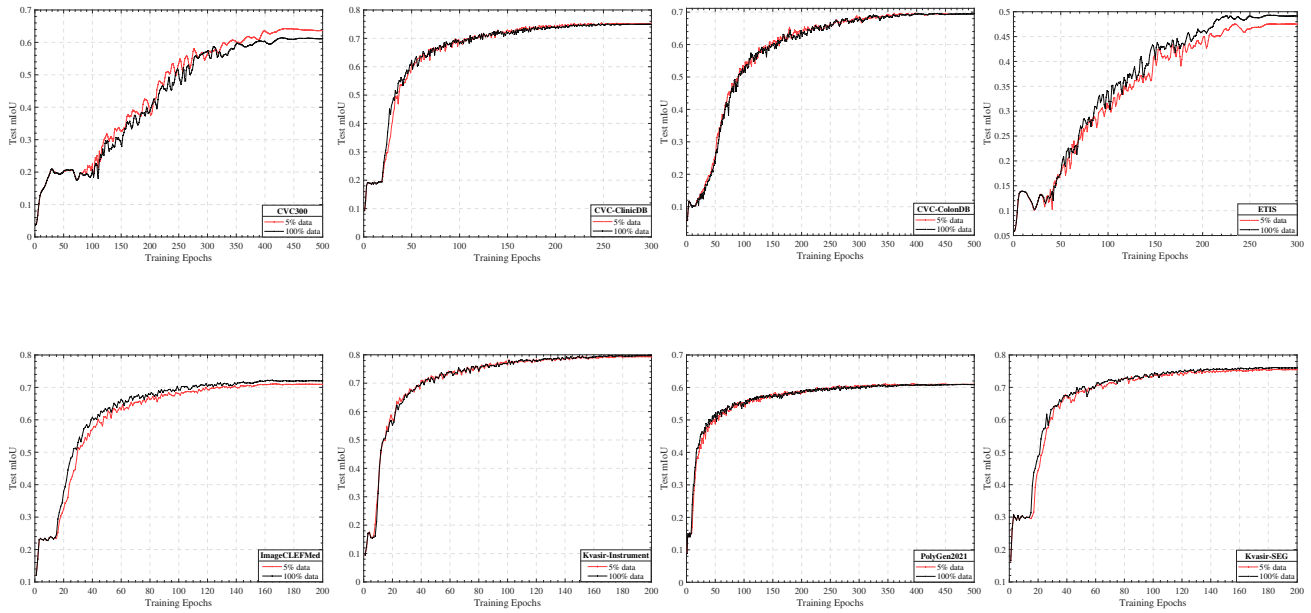


Figure 3: Demonstration of the feasibility of data-effective learning. We compared the performance differences between using only 5% of the pre-training data (in red) and using 100% of the data (in black) in 8 datasets. The results indicate that using only 5% of the pre-training data can achieve results comparable to using 100% of the pre-training data, which fully demonstrates the validity of the MedDEL method.

Table 3: Demonstration of the rationality of NormDEL. The table shows the performance of different proportions of pre-trained data on mIoU (↑) as well as NormDEL (↑) in eight different datasets. The results indicate that the performance of mIoU is almost comparable in different scales, suggesting that it primarily measures performance without considering the scale of the pre-training data used. In contrast, NormDEL incorporates the factor of data scale, and can evaluate data compactness even with only 5% of the data used. This demonstrates the rationality of the NormDEL method, as it not only assesses performance but also effectively utilizes the data scale.

Dataset	5% pretraining data		10% pretraining data		20% pretraining data		30% pretraining data		50% pretraining data		100% pretraining data	
	mIoU	NormDEL	mIoU	NormDEL	mIoU	NormDEL	mIoU	NormDEL	mIoU	NormDEL	mIoU	NormDEL
Kvasir-Instrument	79.38	**68.03**	79.52	67.25	80.22	65.85	80.38	64.06	**80.48**	61.97	79.70	57.28
Kvasir-SEG	75.45	**67.21**	75.74	66.49	76.37	65.14	76.03	63.33	**76.77**	61.43	76.02	56.95
ImageCLEFmed	70.95	**66.26**	71.80	65.69	71.44	64.22	**72.58**	62.76	71.23	60.64	72.02	56.59
ETIS	47.50	**61.11**	49.44	61.00	50.20	60.13	**50.28**	58.94	49.62	57.47	49.13	54.51
PolypGen2021	60.93	**64.10**	61.61	63.59	61.47	62.32	**62.28**	61.01	62.20	59.32	60.89	55.58
CVC-300	**63.67**	**64.69**	56.69	62.55	61.40	62.31	62.85	61.11	61.97	59.29	61.16	55.60
CVC-ClinicDB	75.24	**67.16**	74.58	66.26	75.27	64.94	**75.50**	63.25	75.49	61.25	74.95	56.85
CVC-ColonDB	69.52	**65.96**	68.58	65.03	69.90	63.93	**71.60**	62.59	69.72	60.42	69.48	56.36

segmentation task in endoscopy as an example, where the general metric for measuring performance is often mIoU (mean Intersection over Union), we can formulate the following equation:

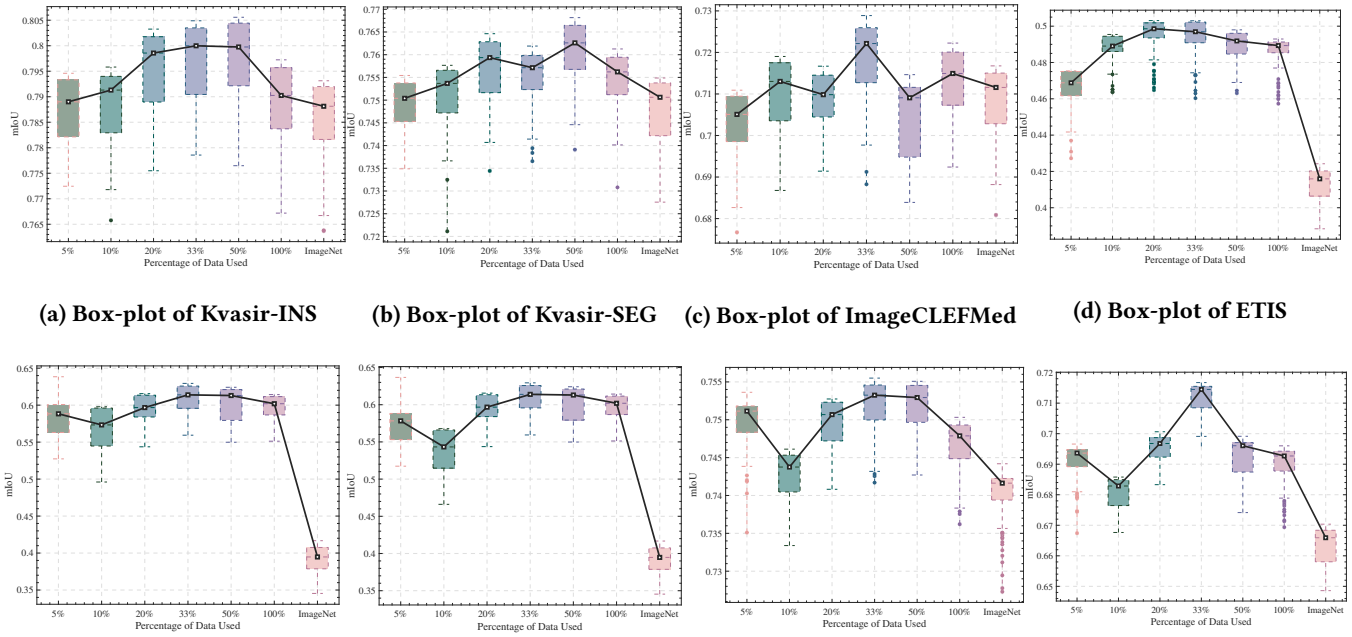
$$DEL = mIoU \cdot e^{-\alpha \cdot R} \tag{4}$$

Where α serves as a positive weight parameter used to adjust the influence of mIoU and the retention ratio in DEL, and R represents the proportion of the retained dataset.

Subsequently, we aim to normalize DEL to a range between 0 and 1 for a clearer definition of data-effective performance. Thus, we obtain the definition of NormDEL:

$$NormDEL = \frac{1}{1 + e^{-DEL}} \tag{5}$$

Through NormDEL, we can comprehensively evaluate the adaptability of the data-effective model with limited training data, emphasizing the critical factors of model data-effective in large-scale datasets.



(a) Box-plot of Kvasir-INS (b) Box-plot of Kvasir-SEG (c) Box-plot of ImageCLEFMed (d) Box-plot of ETIS

(e) Box-plot of PolyGen2021 (f) Box-plot of CVC-300 (g) Box-plot of CVC-ClinicDB (h) Box-plot of CVC-ColonDB

Figure 4: Box plot of the baseline method (MedDEL) for testing mIoU under retaining different data ratios. Under the same computational resources, we plot the impact of varying proportions of pre-training data on downstream task performance, using the data proportions from Table 2 and models pre-trained on ImageNet for reference. The results demonstrate that the performance of downstream tasks initially improves with an increase in pre-training data but eventually decreases. This is because the initially added data is effective, however, as more data is introduced beyond a certain point, it includes redundancy or erroneous information, which adversely affects the model's performance.

4 EXPERIMENT

To assess the capability of MedDEL, we train the Foundation Model on three pre-training datasets as presented in Table 1. Subsequently, we conduct finetuning on downstream task datasets to evaluate performance.

4.1 Settings

To achieve the pretraining of the Foundation Model, enabling it to provide a better initial value for the majority of endoscopy tasks, we use the unsupervised Masked Auto Encoder (MAE-ViT-Base) [24].

During the training of all models, we selectively filter the pre-training dataset capacity based on different thresholds of ϵ and η. Specifically, we set ϵ to 0.9 and η from 0.7 to 0.9 with a step size of 0.05. Consequently, we obtain the dataset quantities as presented in Table 2. The batch size is set to 16, and the training is conducted using AdamW [40] and a cosine learning rate [21]. The model maintains a peak learning rate of 1.5×10^{-4} throughout the training process. In the validation of downstream tasks, we implement a simple Multi-Layer Perceptron (MLP) [53] to achieve segmentation heads.

4.2 Effective Data Utilization with MedDEL

Figure 3 demonstrates that, under equivalent computational resources, the MedDEL method which operates on a reduced dataset (utilizing only 5% of the data, indicated by the red curve), generally performs comparably to the full dataset (using 100% of the data, indicated by the black curve) across various downstream tasks. Moreover, in certain specific scenarios, due to its smaller dataset size and higher data quality, the MedDEL method exhibits superior performance when trained on just 5% of the data. This advantage is particularly pronounced, as seen in the case of CVC-300, highlighting its effectiveness in efficiently leveraging a small amount of high-quality data. This underscores the MedDEL method's characteristic focus and refinement during the learning process, enabling it to more effectively capture crucial task features, resulting in improved performance in low-data scenarios. This result emphasizes the significance of MedDEL method in enhancing data effectiveness, particularly in the era of big data when high-performance models are needed, showcasing its capability for effective data utilization.

4.3 Storage Saving Analysis

Table 3 presents the impact of different amounts of pre-training data on mIoU and NormDEL across 8 downstream datasets. It is worth noting, in the majority of tasks, the mIoU achieved with only 5% of the data is comparable to that achieved with 100% of the data. To clearly assess the influence of pre-training data set proportion on performance, we computed NormDEL for each proportion. The results indicate that in some tasks, although the mIoU attained with

Table 4: Illustrating the impact of different pre-training time (epochs) on performance metrics (mIoU) using 20% and 33% of pretraining data. The results indicate that the optimal performance is often not achieved with the standard pretraining duration (1,000 epochs for 20% and 600 epochs for 33%) but rather occurs with early pretraining time, demonstrating the effectiveness of MedDEL in reducing computational resource consumption.

mIoU / Pretraining Time / Dataset	20% pretraining data								33% pretraining data					
	200	300	400	500	700	800	900	1,000	240	360	420	480	540	600
Kvasir-Instrument	80.21	79.73	79.98	80.01	80.06	80.11	**80.22**	80.22	80.25	**80.50**	80.32	80.39	80.41	80.38
Kvasir-SEG	75.80	75.69	76.13	75.90	76.25	76.33	**76.39**	76.37	75.73	75.94	75.84	**76.07**	75.78	76.03
ImageCLEFmed	71.44	71.60	71.44	**72.91**	71.51	71.52	71.84	71.44	**72.59**	72.32	72.34	72.54	72.19	72.58
ETIS	49.07	49.70	50.88	51.11	50.96	**51.13**	51.01	50.20	49.54	51.23	**51.30**	50.28	50.79	50.28
PolypGen2021	**61.64**	61.32	61.54	61.00	61.47	61.15	61.30	61.47	61.15	61.62	62.10	62.28	**62.29**	62.28
CVC-300	58.66	57.52	56.81	60.59	**61.51**	60.78	60.00	61.40	62.09	62.34	**64.40**	61.40	63.34	62.85
CVC-ClinicDB	74.66	**75.56**	75.48	75.29	75.11	74.83	75.41	75.27	75.44	75.11	75.24	75.17	75.28	**75.50**
CVC-ColonDB	69.83	70.17	**70.68**	69.65	70.62	70.10	69.89	69.90	**72.09**	71.62	70.89	71.43	71.95	71.60

Table 5: MedDEL vs. random selection method. For both MedDEL and random method, the influence of various proportions of pretraining data on the average mIoU over the last 50 epochs was investigated across 8 different datasets. MedDEL outperforms the random method across almost all proportions of the pretraining dataset. Moreover, as the size of the pretraining dataset decreases, the average mIoU over the last 50 epochs does not exhibit a significant decrease. This underscores the effectiveness of utilizing MedDEL for dataset structural optimization.

mIoU / Dataset	5% pretraining data		20% pretraining data		33% pretraining data		50% pretraining data	
	MedDEL	Random	MedDEL	Random	MedDEL	Random	MedDEL	Random
Kvasir-Instrument	**79.50**	79.24	**80.11**	79.67	**80.29**	79.24	**80.35**	79.23
Kvasir-SEG	**75.67**	75.33	**76.23**	76.04	**75.96**	75.75	**76.60**	76.13
ImageCLEFmed	70.90	**71.20**	**71.69**	71.11	**71.42**	71.06	**72.56**	71.29
ETIS	**47.32**	46.22	**49.29**	48.71	50.11	**50.51**	**50.10**	49.93
PolypGen2021	**60.92**	60.70	**61.51**	61.47	**62.24**	60.46	**62.19**	60.91
CVC-300	**63.81**	60.17	**61.40**	60.53	**62.84**	60.27	**62.60**	62.19
CVC-ClinicDB	**75.17**	74.57	**74.52**	73.12	**75.18**	71.64	**75.42**	74.75
CVC-ColonDB	**69.46**	68.62	**69.85**	69.01	**71.54**	69.50	**69.71**	69.48

5% of the data may not be the highest when considering dataset size comprehensively, its NormDEL reaches the highest.

Furthermore, To analyze and evaluate the accuracy and robustness of the baseline method under the same time performance, Figure 4 illustrates a box plot of different proportions of pre-trained data in various downstream datasets, based on the mIoU probabilities. Each box represents the predictive performance of the model during the several final epochs, employing the median, approximate quartiles, and the lowest and highest probabilities to intuitively display the level, spread, and symmetry of the mIoU distribution.

Figure 4 demonstrates a unique trend in the performance of models (measured by the median mIoU) across most downstream tasks, in relation to the increase in the volume of pre-training data. Initially, performance improves with an increase in data volume, reaching a peak, and then starts to decline as data volume continues to grow. Specifically, when a smaller volume of data is used (as in the case of the 5% data volume), despite the limited quantity, the average information content per image is higher, enabling the model to achieve relatively decent results. However, as the data volume increases, redundancy in information also grows. At a certain point, a balance is reached between information content and redundancy, culminating in peak model performance. Beyond this point, further increases in data volume lead to a decline in performance. This trend is reflected in the length of the boxes in the plot, where the box length first increases with the data volume, reaches a peak, and then starts to shorten.

However, in the results from Figure 4, the trends are not as stable, and there is fluctuation observed in some of the downstream tasks. This also reflects the limitations of the MedDEL method. These fluctuations may arise from differences in data characteristics, task complexities, or other factors. It is precisely this diversity that complicates the optimization of data effectiveness, requiring in-depth research and fine-tuning.

Nevertheless, these findings still underscore the effectiveness of MedDEL in reducing storage burden and achieving good results with smaller proportions of data. We encourage more researchers to engage with our study and propose more refined methods to make the most of limited data resources.

4.4 Computing Power Saving Analysis

Section 4.3 demonstrates that MedDEL effectively enhances storage performance and achieves results comparable to large-scale datasets

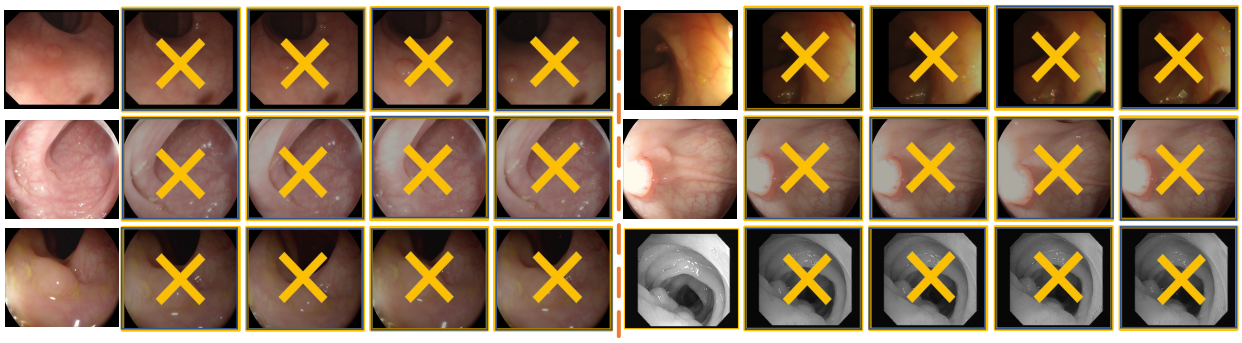


Figure 5: Demonstration of images deleted by MedDEL. This figure shows MedDEL deleting semantically similar images, which appear to have no significant differences between them from a perceptual perspective,

by using fewer data. Building upon this, we further explore the optimization of MedDEL's time performance. Table 4 presents models trained with 20% and 33% of pretraining data, using the epoch numbers from Table 2 as a reference. We progressively decrease the pretraining time in 10% increments, aiming to explore MedDEL's performance metrics (mIoU values) across various downstream datasets at different pretraining times (epochs).

Remarkably, we observe that optimal performance is not consistently achieved with models trained using 100% of the pretraining time across nearly all datasets. Instead, we note the presence of a complex uncertainty and fluctuation, where the optimal result may be distributed among different pretraining time points in diverse datasets and with varying proportions of pretraining data. This underscores the effectiveness of MedDEL in conserving computational resources. However, it also suggests that optimizing pretraining time may require different strategies for various tasks and data contexts. Our study has not yet uncovered the specific relationships between these strategies, indicating the need for more in-depth research in the future.

4.5 Ablation Study

To validate the effectiveness of our model, we conduct ablation experiments where we compare the MedDEL method and the random method across different downstream task datasets. The metric utilized in Table 5 is the average mIoU over the last 50 epochs, chosen to mitigate errors compared to relying solely on the mIoU from the last epoch. The results are promising, consistently showcasing our approach outperforming random selection.

Specifically, on the CVC-300, CVC-ColonDB, Kvasir-SEG, PolypGen as well as Kvasir-Instrument datasets, MedDEL's performance significantly surpasses that of random selection method with all different proportions of pretraining data. This is particularly evident in the more favorable mIoU (mean intersection over union) curves observed throughout the entire training process.

Furthermore, the superior performance of MedDEL is underscored by its ability to effectively leverage pretraining data across different datasets. By consistently surpassing the random selection method across various datasets, MedDEL demonstrates its robustness and generalizability in different settings. This indicates the effectiveness of our approach in capturing and leveraging the underlying semantic similarities among images, thereby enhancing model performance across diverse downstream tasks.

4.6 Visualization of MedDEL(DEL)

The MedDEL algorithm, as a benchmark method in the field of endoscopic, is based on the Vision Transformer (ViT) model [15] for extracting image features, generating a 768-dimensional feature output. By introducing K-means clustering [41], it successfully reduces computational complexity and further effectively addresses irrelevant data in the endoscopic dataset by setting similarity thresholds and filtering conditions.

Figure 5 visualizes the results of the MedDEL (DEL) algorithm, where upon observation, it is evident that MedDEL eliminates a significant portion of redundant data with highly similar features from the dataset. Considering their similarity, we suppose removing such redundant data is a reasonable strategy to optimize the dataset structure, enhancing dataset quality. This optimization improves the model's understanding of endoscopic images, making it more targeted and interpretable, and providing robust support for medical image analysis [15].

5 CONCLUSION

In the era of large language models, we are the first to propose whether larger pre-training data necessarily leads to improved model performance. In order to investigate this issue, we propose the first medical data-effective Benchmark, which involves a million-level dataset from 31 medical centers (DataDEL), a data-effective approach (MedDEL), and a comprehensive metric for evaluating pretraining data volume (NormDEL). The establishment of an open benchmark for data-effective learning is crucial for the medical artificial intelligence research community, which is an encouraging foundation for subsequent research endeavors.

REFERENCES

[1] 2023. Endoscopy Procedures Estimates Market Volume, Share & Trends Analysis Report. Report ID: GVR-4-68039-915-0, Number of Pages: 118, Format: Electronic (PDF), Historical Range: 2016 - 2021, Industry: Healthcare. Segment Forecasts, 2023 - 2030.

[2] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540* (2023).

[3] Sharib Ali, Noha Ghatwary, Barbara Braden, Dominique Lamarque, Adam Bailey, Stefano Realdon, Renato Cannizzaro, Jens Rittscher, Christian Daul, and James East. 2020. Endoscopy disease detection challenge 2020. *arXiv preprint arXiv:2003.03376* (2020).

[4] Sharib Ali, Debesh Jha, Noha Ghatwary, Stefano Realdon, Renato Cannizzaro, Osama E Salem, Dominique Lamarque, Christian Daul, Michael A Riegler, Kim V Anonsen, et al. 2021. PolypGen: A multi-center polyp detection and segmentation dataset for generalisability assessment. *arXiv preprint arXiv:2106.04463* (2021).

[5] Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James East, Georges Wagnieres, Victor Loschenov, Enrico Grisan, et al. 2019. Endoscopy artifact detection (EAD 2019) challenge dataset. *arXiv preprint arXiv:1905.03209* (2019).

[6] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics* 43 (2015), 99–111.

[7] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. 2012. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition* 45, 9 (2012), 3166–3182.

[8] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data* 7, 1 (2020), 283.

[9] Ming Chen, Shupeng Wang, and Liang Tian. 2013. A High-precision Duplicate Image Deduplication Approach. *J. Comput.* 8, 11 (2013), 2768–2775.

[10] Ming Chen, Yang Wang, Xiaoxiang Zou, Shupeng Wang, and Guangjun Wu. 2012. A duplicate image deduplication approach via Haar wavelet technology. In *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, Vol. 2. IEEE, 624–628.

[11] Xue-Wen Chen and Xiaotong Lin. 2014. Big data deep learning: challenges and perspectives. *IEEE access* 2 (2014), 514–525.

[12] Neil T Clancy, Rui Li, Kevin Rogers, Paul Driscoll, Peter Excel, Ron Yandle, George Hanna, Nigel Copner, and Daniel S Elson. 2012. Development and evaluation of a light-emitting diode endoscopic light source. In *Advanced Biomedical and Clinical Diagnostic Systems X*, Vol. 8214. SPIE, 105–111.

[13] Nelson Cowan. 2001. Metatheory of storage capacity limits. *Behavioral and brain sciences* 24, 1 (2001), 154–176.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[16] Richard Dosselmann and Xue Dong Yang. 2011. A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing* 5 (2011), 81–91.

[17] Wayne W Eckerson. 2002. Data quality and the bottom line. *TDWI Report, The Data Warehouse Institute* (2002), 1–32.

[18] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. 2010. Why does unsupervised pre-training help deep learning?. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 201–208.

[19] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. 2017. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410* (2017).

[20] Pedro Furtado and Henrique Madeira. 1999. Analysis of accuracy of data reduction techniques. In *DataWarehousing and Knowledge Discovery: First International Conference, DaWaK'99 Florence, Italy, August 30–September 1, 1999 Proceedings 1*. Springer, 377–388.

[21] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243* (2018).

[22] Oleksandra Gulenko, Hyunmo Yang, KiSik Kim, Jin Young Youm, Minjae Kim, Yunho Kim, Woonggyu Jung, and Joon-Mo Yang. 2022. Deep-Learning-Based Algorithm for the Removal of Electromagnetic Interference Noise in Photoacoustic Endoscopic Image Processing. *Sensors* 22, 10 (2022), 3961.

[23] Danny Harnik, Oded Margalit, Dalit Naor, Dmitry Sotnikov, and Gil Vernik. 2012. Estimation of deduplication ratios in large data sets. In *2012 IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST)*. IEEE, 1–11.

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. (2021).

[25] Zhongyu He, Peng Wang, Yuelong Liang, Zuoming Fu, and Xuesong Ye. 2021. Clinically available optical imaging technologies in endoscopic lesion detection: current status and future perspective. *Journal of Healthcare Engineering* 2021 (2021), 1–27.

[26] William Hersh, Henning M"uller, and Jayashree Kalpathy-Cramer. 2009. The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging* 22 (2009), 648–655.

[27] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A Riegler, Thomas de Lange, Peter T Schmidt, Håvard D Johansen, et al. 2021. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27*. Springer, 218–229.

[28] Debesh Jha, Vanshali Sharma, Neethi Dasu, Nikhil Kumar Tomar, Steven Hicks, MK Bhuyan, Pradip K Das, Michael A Riegler, Pål Halvorsen, Ulas Bagci, et al. 2023. GastroVision: A Multi-class Endoscopy Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Workshop on Machine Learning for Multimodal Healthcare Data*. Springer, 125–140.

[29] Debesh Jha, Pia H Smedsrud, Dag Johansen, Thomas de Lange, Håvard D Johansen, Pål Halvorsen, and Michael A Riegler. 2021. A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. *IEEE journal of biomedical and health informatics* 25, 6 (2021), 2029–2040.

[30] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*. Springer, 451–462.

[31] Luis O Jimenez-Rodriguez, Emmanuel Arzuaga-Cruz, and Miguel Vélez-Reyes. 2007. Unsupervised linear feature-extraction methods and their effects in the classification of high-dimensional data. *IEEE Transactions on geoscience and remote sensing* 45, 2 (2007), 469–483.

[32] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1646–1654.

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

[34] Rafal Kwasny, Daniel Friar, and Giuseppe Papallo. 2020. Benchmarking Deep Learning Workloads with TensorFlow on the NVIDIA GeForce RTX 3090.

[35] Chengcai Leng, Hai Zhang, Bo Li, Guorong Cai, Zhao Pei, and Li He. 2018. Local feature descriptor for image matching: A survey. *IEEE Access* 7 (2018), 6424–6434.

[36] Mengfang Li, Yuanyuan Jiang, Yanzhou Zhang, and Haisheng Zhu. 2023. Medical image analysis using deep learning algorithms. *Frontiers in Public Health* 11 (2023), 1273253.

[37] Xuan Li, Liqiong Chang, and Xue Liu. 2021. CE-Dedup: Cost-effective convolutional neural nets training based on image deduplication. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 11–18.

[38] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems* 33, 12 (2021), 6999–7019.

[39] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2016. Deep supervised hashing for fast image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2064–2072.

[40] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. (2017).

[41] J. Macqueen. 1967. Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability, 5th* 1 (1967).

[42] Stephane G Mallat. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence* 11, 7 (1989), 674–693.

[43] Dutch T Meyer and William J Bolosky. 2012. A study of practical deduplication. *ACM Transactions on Storage (ToS)* 7, 4 (2012), 1–20.

[44] Hussain Nyeem, Wageeh Boles, and Colin Boyd. 2013. A review of medical image watermarking requirements for teleradiology. *Journal of digital imaging* 26 (2013), 326–343.

[45] JK Periasamy and B Latha. 2021. Efficient hash function–based duplication detection algorithm for data Deduplication deduction and reduction. *Concurrency and Computation: Practice and Experience* 33, 3 (2021), e5213.

[46] M Mostafizur Rahman and Darryl N Davis. 2013. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing* 3, 2 (2013), 224.

[47] Papia Ray, S Surender Reddy, and Tuhina Banerjee. 2021. Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review* 54, 5 (2021), 3473–3515.

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[49] Prakash Chandra Sharma, Sulabh Bansal, Rohit Raja, Phyu Myo Thwe, Moe Moe Htay, and Su Su Hlaing. 2021. Concepts, strategies, and challenges of data deduplication. In *Data Deduplication Approaches*. Elsevier, 37–55.

[50] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery* 9 (2014), 283–293.

[51] Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. 2021. Kvasir-Capsule, a video capsule endoscopy dataset. *Scientific Data* 8, 1 (2021), 142.

[52] CHARLES E Smith and ANTONIO Nanci. 2003. Overview of morphological changes in enamel organ cells associated with major events in amelogenesis. *International Journal of Developmental Biology* 39, 1 (2003), 153–161.

[53] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, and Alexey Dosovitskiy. 2021. MLP-Mixer: An all-MLP Architecture for Vision. (2021).

[54] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, Aaron Courville, et al. 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering* 2017 (2017).

[55] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5310–5319.

[56] Wen Xia, Hong Jiang, Dan Feng, Fred Douglis, Philip Shilane, Yu Hua, Min Fu, Yucheng Zhang, and Yukun Zhou. 2016. A comprehensive study of the past, present, and future of data deduplication. *Proc. IEEE* 104, 9 (2016), 1681–1710.

[57] Ling Xiao, Beiji Zou, Chengzhang Zhu, and Fanbo Nie. 2023. ESDedup: An efficient and secure deduplication scheme based on data similarity and blockchain for cloud-assisted medical storage systems. *The Journal of Supercomputing* 79, 3 (2023), 2932–2960.

[58] Hengxiang Xie, Yuhui Deng, Hao Feng, and Lei Si. 2021. Pxdedup: Deduplicating massive visually identical jpeg image data. *Big Data Research* 23 (2021), 100171.

[59] Martin J Yaffe. 2019. Emergence of "big data" and its potential and current limitations in medical imaging. In *Seminars in Nuclear Medicine*, Vol. 49. Elsevier, 94–104.

[60] Xiangyu Zou, Jingsong Yuan, Philip Shilane, Wen Xia, Haijun Zhang, and Xuan Wang. 2022. From Hyper-Dimensional Structures to Linear Structures: Maintaining Deduplicated Data's Locality. 18, 3 (2022).