Proceedings of 13th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2023 Month: April Volume: 13

Dataset Condensation of Gastrointestinal Tract Endoscopic Images

Shanti Tamang [a], Kiran Chandra Dahal [b]

[a, b] *Institute of Engineering, Tribhuvan University, Nepal*
[a] shanti76msiise16@tcioe.edu.np, [b] dahalkc@ioe.edu.np

Abstract

Gastrointestinal diseases, which affect the gastrointestinal tract, are a common health-related problem. The common procedure for their diagnosis is endoscopy, which results in endoscopic images. Although sharing these images is beneficial, there are difficulties in their storage and transmission. Dataset condensation has recently grown in favor as a data-efficient learning technique for reducing large datasets into concise useful samples for deep neural network training which also retains the performance of the original dataset. Therefore, a dataset containing endoscopic images of the gastrointestinal tract is condensed to obtain a smaller synthetic dataset. The gastrointestinal tract disorders were classified using a convolutional neural network and the distribution matching method is used to condense the images in three settings of 1, 5 and 10 images per class. The classification performance of the original dataset and the condensed dataset is also compared and analyzed.

Keywords

CNN, Dataset Condensation, Dataset Distillation, Distribution Matching, Gastrointestinal Endoscopy

1. Introduction

There has been significant advances in medical technology, particularly in the field of Computer Aided Diagnosis (CAD) that utilizes machine learning techniques to help physicians with disease diagnosis. Deep learning techniques have shown significant results in exploiting large-scale medical data, but the requirement for large datasets with millions of samples poses challenges in terms of storage and computational complexity [1]. Small medical facilities may not have enough data for training deep convolutional neural networks (DCNN), thus cross-flow of data among medical facilities is necessary. However, sharing sensitive medical data raises privacy concerns, and there are also huge storage and transmission costs to consider.

The reduction of large datasets has been attempted using various methods such as instance selection, core-set construction, and dataset pruning [2]. Recently, dataset distillation, also referred to as dataset condensation, has gained popularity. This involves creating a smaller dataset from a larger one, which differs from previous methods in that the distilled dataset gathers knowledge from the entire original dataset, resulting in comparable performance to the larger dataset. This technique is seen as a solution to the challenges posed by the storage and transmission costs of large medical datasets. Synthesized data can be utilized for model training while maintaining privacy and preventing retrieval through MIA (Membership Inference Attack) or visual comparison analysis [3].

Gastrointestinal (GI) tract disease is a common health problem caused by a combination of lifestyle, genetics, and environmental factors. According to research from Cancer Statistics, the United States had the highest occurrence of stomach cancer in 2018 [4]. In Nepal, a large number of people suffer from GI diseases[5]. Endoscopy is the standard diagnostic method for GI disease, providing gastroenterologists with important endoscopic images and videos to make accurate diagnoses and determine the appropriate course of treatment. It is crucial to diagnose stomach cancer early as it is one of the main causes of cancer-related deaths globally. However, collecting a large number of gastrointestinal images can be challenging for small medical facilities.

The condensation of a dataset of GI endoscopic images has been performed, resulting in a condensed dataset that can be used for further computer-aided diagnosis. In a subsequent step, this condensed dataset is used to classify various GI diseases using CNN. Additionally, a comparison of its performance with the original, larger dataset has been conducted.

2. Related Works

2.1 Dataset Condensation

The current state of the art machine learning models require large datasets with millions of samples, which can be difficult to store, pre-process, and train on. To address these constraints, a recent approach called training set synthesis has emerged, aiming to generate a small synthetic dataset that can be used to train deep neural networks for specific tasks. This approach was introduced in the Dataset Distillation (DD) [2], which utilizes gradient optimization to construct synthetic images from a set of original training images that are most helpful for empirical risk minimization with respect to model parameters. However, a major drawback of this approach is the time-consuming optimization process which includes unrolling the recursive computation graph and changing network weights for subsequent steps for each outer iteration.

Further, Dataset Condensation with Gradient Matching (DC) [6] introduces a new technique for dataset condensation that creates synthetic input samples by comparing model gradients with those from the original input samples. This method avoids the time-consuming unrolling of the computational graph by considering the gradients of the real and artificial training losses with respect to the model parameters.

Based on this work, Differentiable Siamese Augmentation(DSA)[7] was proposed, which uses data augmentation to create more informative synthetic images and improve performance in training neural networks. The resulting synthetic training set can be used with data augmentation to outperform state-of-the-art methods.

Another technique to improve efficiency is Dataset Meta-Learning from Kernel Ridge-Regression (KRR) [8]. This method condenses datasets by approximating neural networks with kernel ridge regression, using a novel algorithm called Kernel Inducing Point (KIP). KIP is a meta-learning algorithm that is based on recent findings linking infinitely wide neural networks to KRR, which have been successful in generating high-quality datasets. KIP has been shown to compress datasets by one or two orders of magnitude for KRR tasks, outperforming previous methods of dataset distillation.

Unlike DC and DSA, Dataset Condensation with Distribution Matching (DM) [9] learns condensed datasets by directly comparing the output characteristics of actual and synthetic samples. In a large number of sampled embedding spaces, Dataset Condensation with Distribution Matching presents a straightforward yet efficient approach for creating condensed images by matching the feature distributions of synthetic and real training images. This method avoids costly bi-level optimization by framing the task as a distribution matching problem with the maximum mean discrepancy (MMD).

2.1.1 Dataset Condensation on Medical Dataset

A gradient descent-based soft-label anonymous gastric X-ray image distillation technique achieves great classification performance by condensing each class into a single image for training [10]. This work has been further extended where the whole dataset of stomach X-ray images are compressed into a single anonymous soft-label patch image for maximum compression rate [11]. The COVID-19 chest X-ray image dataset condensation demonstrates that DC may achieve great detection performance even with a limited number of anonymized chest X-ray images [12]. Here, the student network's training parameters match those of the teacher network trained on the original dataset. Another technique [13] does a more efficient distillation of the COVID-19 chest X-ray images and enhances distillation performance by pruning difficult-to-match parameters.

The majority of the aforementioned techniques have only been tested on small datasets such as CIFAR, MNIST, and X-ray images, which have mostly been used in medical image condensation. Therefore, dataset condensation with distribution matching on labeled images of HyperKvasir[14] is performed here in light of the inefficiencies related to sharing large medical datasets. DM [9] showed how several classes of synthetic data can be learned independently and concurrently. Additionally, it demonstrated its effectiveness on TinyImageNet, a more difficult dataset. Hence, dataset condensation with distribution matching is proposed for a complex GI endoscopy image dataset.

3. Proposed Methodology

Figure 1 shows the system block diagram. The first step is the data collection, here the HyperKvasir[5] dataset which is the collection of the GI tract endoscopic images will be used. Major tasks include synthesizing condensed images and using them to train classifiers. Models trained on the original dataset and the condensed dataset are then compared and evaluated.

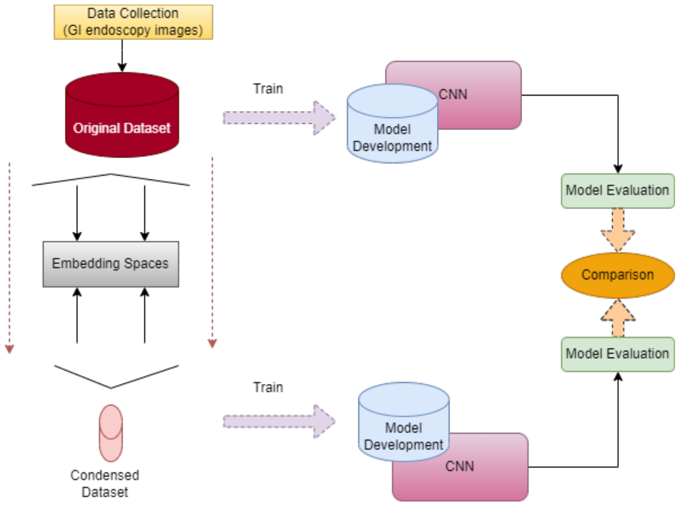

Figure 1: System block diagram

3.1 Preprocessing

The images for gastrointestinal endoscopy included in the dataset range in size from 720x576 to 1920x1072. The size of all images was reduced to 416X416 and the metadata was also removed beforehand because the size of the images may also affect training time. Images were once again resized to 128X128 due to computational complexity, memory limitations, and to enable more effective processing. Then, dataset is normalised which involves transforming the images so that their mean and standard deviation become 0 and 1 respectively. This also helps to optimize the training process and improve the accuracy of the neural network.

3.2 Data Augmentation

As there is significant imbalance in the data which can be seen in figure 4, the result could be the biased models which perform poorly for the minority class. So, oversampling of the minority classes is done. This involves adding more samples of the minority class to balance the class distribution. So, it can be effective in increasing the representation of the minority class in the model. So common augmentation techniques like random transformations is applied. This involves randomly transforming the existing samples in the minority class, such as flipping, rotating, scaling, and adding noise, to generate new samples.

3.3 Synthesizing condensed dataset

The process starts with synthesizing a condensed dataset from the original dataset, which consists of GI endoscopy images. The algorithm utilizes distribution matching to match the original dataset and synthetic dataset, where the distance between the original and condensed images in a lower-dimensional embedding space is estimated using the maximum mean discrepancy. The goal is to minimize the difference between the actual and synthetic images belonging to the same class since this is an image classification task. Here, the large dataset is given as

$$T=\left\{\left(x_1,y_1\right),\ldots,\left(x_{|T|},y_{|T|}\right)\right\} \tag{1}$$

with $|\mathscr{T}|$ images and label pairs. Dataset Condensation condenses this to the small synthetic set

$$S=\left\{\left(x_1,y_1\right),\ldots,\left(x_{|S|},y_{|S|}\right)\right\} \tag{2}$$

with $|S|$ synthetic/condensed image and label pairs. Then the model trained on each T and S obtain the comparable performance on unseen testing data. Since training images are often high-dimensional, it can be both expensive and imprecise to estimate the true data distribution P_D. Instead, in distribution matching [9], it is assumed that each training image $x \in \mathbb{R}^d$ is embedded into a lower-dimensional space using a set of parametric functions, denoted as $\psi_v : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$. Each embedding function ψ provides a partial interpretation of the input, and their combination gives a full interpretation. The Maximum Mean Discrepancy (MMD) [15] is used to estimate the separation between the real and synthetic data distributions. Since ground-truth data distributions are not available to us, the empirical estimate of the MMD:

$$\mathbb{E}_{\vartheta\sim P_\vartheta}\left\|\frac{1}{|T|}\sum_{i=1}^{|T|}\psi_\vartheta\left(\left(x_i\right)\right)-\frac{1}{|S|}\sum_{j=1}^{|\S|}\psi_\vartheta\left(\left(s_j\right)\right)\right\|^2 \tag{3}$$

In each iteration, a random mini-batch of the training data is used to calculate the error gradient and update the model parameters. The initialization of the synthetic images is done using either random real training images or Gaussian noise.

The Differential Siamese Augmentation (DSA) [7] is also applied to both real and synthetic batches, so that the resulting condensed dataset can effectively handle augmented images, leading to better performance. Common augmentation strategies include cropping, flipping, color jittering, and rotations, among others is used.

The figure 2 illustrates the steps involved in generating a condensed dataset using distribution matching. The training algorithm involves splitting the training data into smaller mini-batches to efficiently compute the model parameters in each iteration. In each iteration, a random mini-batch is used to calculate the error gradient and update the model parameters. The model is sampled with parameter v in each iteration, and the network parameters are sampled to reduce complexity and improve training efficiency. For each class, pairs of real and synthetic data batches and augmentation parameters are taken, and the mean discrepancy between the augmented batches is calculated to determine the loss function. Stochastic gradient descent is then used to update the synthetic data with a learning rate η by minimizing the loss.

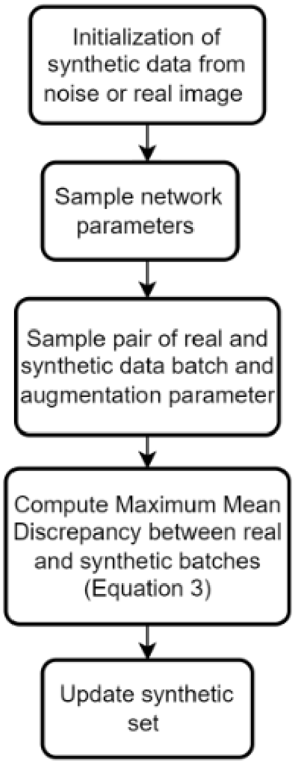
Initialization of synthetic data from noise or real image
Sample network parameters
Sample pair of real and synthetic data batch and augmentation parameter
Compute Maximum Mean Discrepancy between real and synthetic batches (Equation 3)
Update synthetic set

Figure 2: Flow diagram illustrates the steps involved in generating a condensed dataset using distribution matching

3.4 Training Classifier

The Convolutional Neural Network is a well-known machine learning approach used for image classification. In this work, the ConvNet architecture [16] is employed, which consists of blocks with a convolutional layer with W (3x3) filters, a normalization layer N, an activation layer A, and a pooling layer P, designated as [W, N, A, P]xD, with D duplicate blocks in total. The ConvNet architecture comprises of N = 3 repeated blocks, each of which contains a group normalization layer, a ReLU activation layer, and a 128-kernel convolutional layer with 3x3 filters. The linear classifier is placed after the final block. During training, the weights of the ConvNet architecture are initialized using the Kaiming initialization to accelerate convergence, improve the stability of the training process, and reduce overfitting. Additionally, the ConvNet is implemented with Batch Normalization, which further helps in improving the training and generalization performance.

4. Dataset Collection and Description

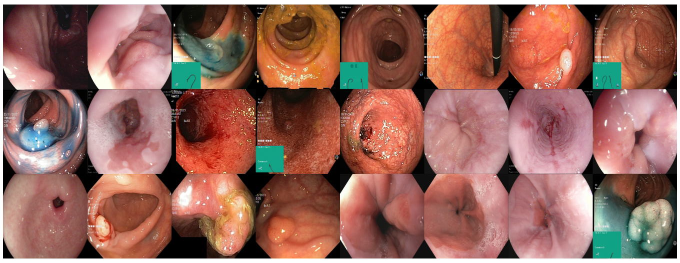

Figure 3: Sample images in the dataset

The HyperKvasir dataset [14] is an improved version of the Kvasir dataset making it a largest repository for GI tract endoscopy images.

The data used in the study was collected during real colonoscopy and gastroscopy procedures at Baerum Hospital in Norway, and was partly labeled by expert gastrointestinal endoscopists.

The dataset consists of several categories that include both pathological signs such as esophagitis, polyps, and ulcerative colitis, as well as anatomical signs like z-line, pylorus, and cecum. Additionally, there are typical signs like normal colon mucosa and stool, along with cases where polyps were removed after treatment, including dyed and lifted polyps and dyed resection margins.

There are 23 classifications and 10,662 images on HyperKvasir. The dataset includes 110,079 (10,662 labeled and 99,417 unlabeled images) images in total, representing both pathological and normal findings as well as anatomical landmarks. Figure 3 shows the sample of images in the dataset and figure 4 shows the images per class for the labeled images in original dataset before augmentation. Here, only labeled images (110,079) are used for the dataset condensation. The dataset was splitted to training, validation and testing set in ratio of 70/20/10.

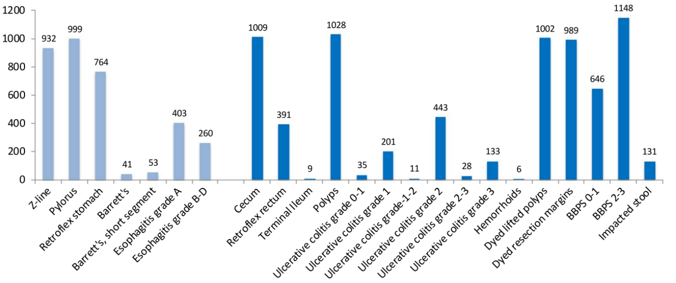

Figure 4: Images per class for the labeled images in original dataset

5. Results and Discussion

The objective of CAD (computer-aided diagnosis) systems is to identify abnormal indicators more efficiently than a human expert. This study aims to assist the CAD system in diagnosing gastrointestinal disorders by creating a synthetic dataset from a large dataset. Obtained output is condensed dataset which still is informative. Here, the first step involves learning 1/5/10 images per class synthetic sets using a ConvNet architecture [16].

The synthetic sets are used to train ConvNet from the beginning, and their performance is assessed on real test data. Figure 5 displays the condensed dataset obtained for 10 images per class with real image initialization, while figure 6 displays the condensed image obtained from noise initialization. Figure 7 shows the visualisation for 5 images per class when batch normalisation is implemented. In each experiment, a single synthetic set is learned and used to train 5 randomly initialized ConvNets. The average accuracy of the 5 trained networks is displayed in all cases. The classification of gastrointestinal image gave the accuracy of 85.31% on whole or original dataset. The obtained accuracy can be considered as the upper bound or maximum accuracy that can be achieved on the condensed dataset as well.

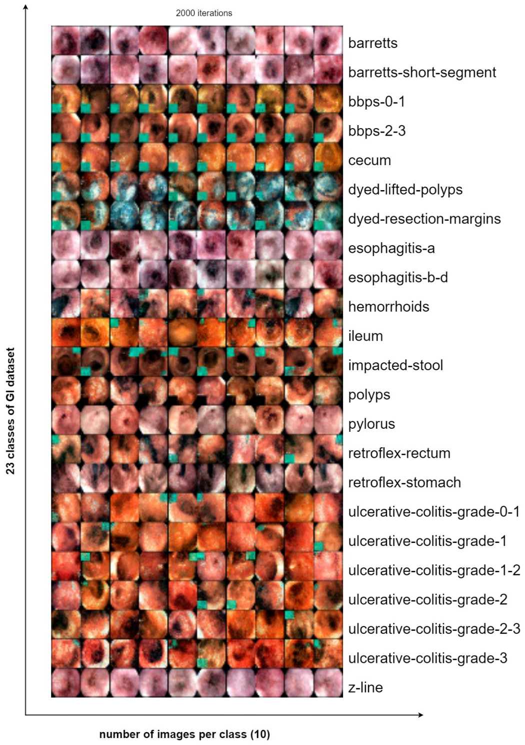

Figure 5: Visualisation of condensed 10 images per class (real image initialisation)

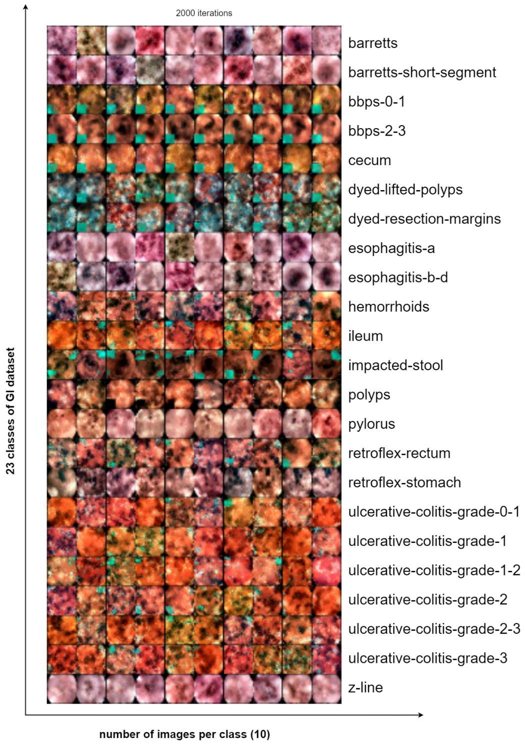

Figure 6: Visualisation of condensed 10 images per class (noise initialisation)

Hyperparameters To generate 1, 5 and 10 images per class, the number of iterations for learning synthetic data was set to 2000. The learning rate for updating synthetic images was set to

1.0 with the batch size of 16 for both real and synthetic dataset. Maximum mean discrepancy between the real and synthetic data distribution was summed as matching loss with stochastic gradient descent (SGD) as the optimizer to update synthetic set. To evaluate the learned dataset, it was trained for 200 epochs with a batch size of 16. The learning rate for updating network parameters was set to 0.001 and the loss function used was categorical cross entropy with the optimizer as stochastic gradient descent (SGD).

Table 1: Performance of the number of images per class when different initialization methods and batch normalisation(BN) are used

ipc	initialisation	accuracy	accuracy with BN
1	noise	30.63 %	31.91%
1	real	31.50 %	34.73%
5	noise	57.48%	60.25%
5	real	60.82 %	65.62%
10	noise	69.05%	73.59%
10	real	70.21%	75.13%

The results of the evaluation of the generated synthetic images are presented in Table 1. The table displays the mean accuracy of the synthetic images generated from both noise and real data initialization for 1, 5, and 10 images per class. The table provides a comparison between the accuracy of the synthetic images generated from different initialization and number of images per class. For 1, 5, 10 images per class, the accuracy of the synthetic images generated from noise initialization is 30.63%, 57.48% and 69.05% without batch normalisation. It can be seen that the accuracy has been increasing with the increasing number of images per classes.

Qualitative Analysis of the results It can be visually inspected that the synthetic images of the Gastrointestinal tract dataset obtained from the noise initialisation contain noticeable noise and unnatural strokes, while the synthetic images generated from the real data are more clear and free of noise. Additionally, the synthetic images of the GI tract dataset obtained from the real data are more visually recognizable and distinct.

Table 1 presents the comparison of the synthetic datasets obtained for 1, 5 and 10 images per class when batch normalization is utilized in the ConvNet model. It is evident that the model with batch normalisation has a higher accuracy. Additionally, with a higher number of images per class, the accuracy increases, with 34.73% , 65.62% and 75.13% in 1, 5, 10 images per class respectively for initialisation from real images with batch normalisation. The figure 7 depicts the synthetic images generated when batch normalization was applied to the convnet model while obtaining a condensed dataset of 5 images per class initialized from real images.

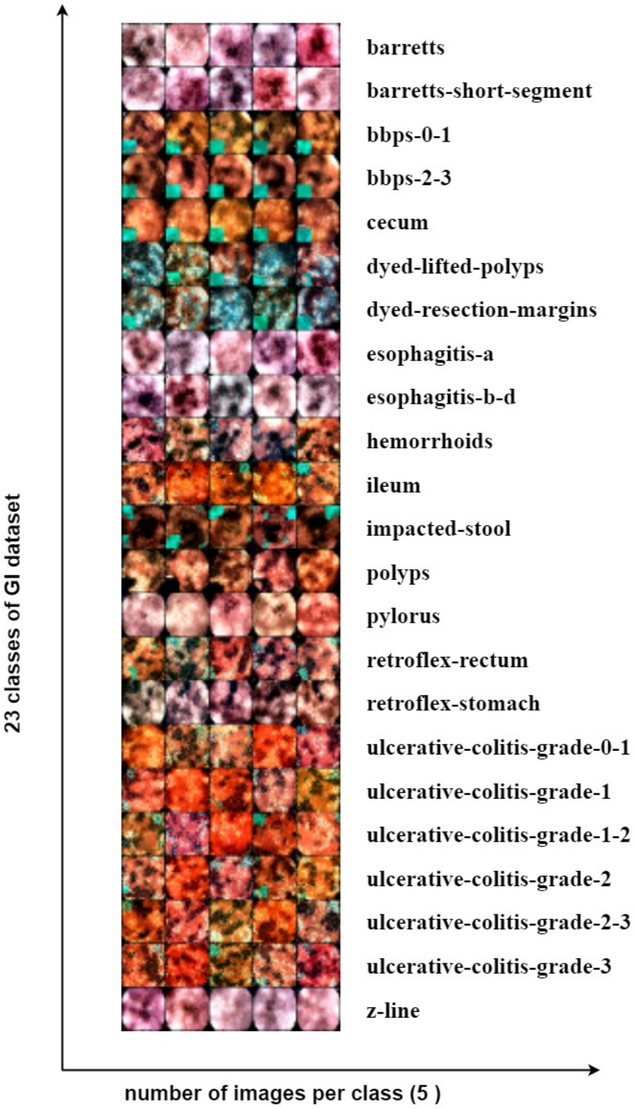

Figure 7: Visualisation of condensed 5 ipc (BN)

6. Conclusions and Future Work

The study generated a condensed dataset of gastrointestinal endoscopic images using feature distribution matching, which can reduce computational complexity. The performance of this condensed dataset was evaluated by applying a convolutional neural network (CNN) for image classification, demonstrating its potential for computer-assisted diagnosis of gastrointestinal diseases. As a potential future improvement, the proposed system could be validated using other datasets and real-time endoscopic images which could improve the system's reliability and utility for detecting and treating digestive tract diseases.

Acknowledgments

This work is carried out under the Department of Electronics and Computer Engineering, IOE, Thapathali Campus. The authors are grateful to the Msc. Coordinator of MSIISE, Er. Dinesh Baniya Kshatri for his valuable support and guidance.

References

[1] Chung-Ming Chen, Yi-Hong Chou, Norio Tagawa, and Younghae Do. Computer-aided detection and diagnosis in medical imaging, 2013.

[2] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[3] Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? *arXiv preprint arXiv:2206.00240*, 2022.

[4] Subhashree Mohapatra, Girish Kumar Pati, Manohar Mishra, and Tripti Swarnkar. Gastrointestinal abnormality detection and classification using empirical wavelet transform and deep convolutional neural network from endoscopic images. *Ain Shams Engineering Journal*, page 101942, 2022.

[5] Yogendra Singh, Jayant Sah, and Bikal Ghimire. Presentation and outcomes of gastric cancer at a university teaching hospital in nepal. *Asian Pacific journal of cancer prevention: APJCP*, 16:5385–5388, 08 2015.

[6] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *ICLR*, 1(2):3, 2021.

[7] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685. PMLR, 2021.

[8] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. *arXiv preprint arXiv:2011.00050*, 2020.

[9] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching, 2021.

[10] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Soft-label anonymous gastric x-ray image distillation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 305–309. IEEE, 2020.

[11] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Compressed gastric image generation based on soft-label dataset distillation for medical data sharing. *Computer Methods and Programs in Biomedicine*, 227:107189, 2022.

[12] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Dataset distillation for medical dataset sharing. *arXiv preprint arXiv:2209.14603*, 2022.

[13] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Dataset distillation using parameter pruning. *arXiv preprint arXiv:2209.14609*, 2022.

[14] Borgli H, Thambawita V, Jha DL Smedsrud PH, Hicks S, et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, 2020.

[15] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[16] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.