The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R^2HCAI)

Dataset Distillation for Medical Dataset Sharing

Guang Li[1] **Ren Togo**[2] **Takahiro Ogawa**[2] **Miki Haseyama**[2]

[1] Graduate School of Information Science and Technology, Hokkaido University, Japan
[2] Faculty of Information Science and Technology, Hokkaido University, Japan
E-mail: {guang, togo, ogawa, mhaseyama}@lmd.ist.hokudai.ac.jp

Abstract

Sharing medical datasets between hospitals is challenging because of the privacy-protection problem and the massive cost of transmitting and storing many high-resolution medical images. However, dataset distillation can synthesize a small dataset such that models trained on it achieve comparable performance with the original large dataset, which shows potential for solving the existing medical sharing problems. Hence, this paper proposes a novel dataset distillation-based method for medical dataset sharing. We also found that a few parameters in the distillation process are difficult to match, which harms the distillation performance. Based on this observation, we improve the distillation performance by introducing parameter pruning. Experimental results on a COVID-19 chest X-ray image dataset show that our method can achieve high detection performance even using scarce anonymized images. The proposed method may make sharing medical datasets between hospitals more efficient and secure.

Introduction

The sharing of medical datasets is essential in enabling the cross-hospital flow of medical information and improving the quality of medical services (Kumar et al. 2021). However, sharing healthcare datasets between different hospitals faces several thorny issues. Firstly, privacy protection has been a severe issue hindering the process when sharing medical image datasets from different hospitals (Kaissis et al. 2020). Second, sharing large-scale high-resolution medical image datasets increases transmission and storage costs (Dash et al. 2019). Therefore, the solution to these problems will significantly promote the development of medical dataset sharing.

Dataset distillation can synthesize a small dataset such that models trained on it achieve comparable performance with the original large dataset (Wang et al. 2018). Although dataset distillation has been proposed for distilling some simple datasets, such as MNIST and CIFAR10, its effectiveness in high-resolution complex medical datasets has not yet been proved (Zhao and Bilen 2021c). Medical dataset distillation may have potential advantages for solving the existing medical dataset sharing problems (Li et al. 2020, 2022a). For example, the size of distilled medical image datasets can be significantly compressed, and distilled images generated from noise are automatically anonymized (Dong, Zhao, and Liu 2022). Therefore, it is desirable to explore the potential of dataset distillation for medical dataset sharing and contribute to real-world applications.

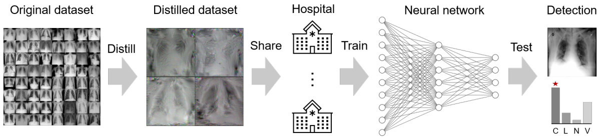

Figure 1: Concept of this study. Our method can improve the efficiency and security of the sharing of medical datasets between different hospitals.

COVID-19 and its variants have rapidly spread worldwide, influencing the health and life of billions of people (Mofijur et al. 2021). Many medical facilities are facing the challenges of the increasing numbers of COVID-19 infections, including a critical shortage of medical resources, and many healthcare providers have themselves been infected (Grover et al. 2020). X-ray is widely used in clinical because of its high speed and low cost. Detecting COVID-19 from chest X-ray (CXR) images is perhaps one of the fastest and easiest ways (Minaee et al. 2020). However, sharing COVID-19 datasets between different hospitals also has the above-mentioned problems.

In this paper, we propose a novel dataset distillation-based method for medical dataset sharing. The concept of this study is shown in Figure 1. The recently proposed dataset distillation method (Cazenavette et al. 2022) by matching network parameters has been proven effective for several datasets. However, the dimension of network parameters is usually large. And we found that a few parameters in the distillation process are difficult to match, which harms the distillation performance. Based on this observation, we improve the distillation performance by introducing parameter pruning. We perform experiments on a COVID-19 CXR image dataset to prove the effectiveness of the proposed method. Experimental results show that we can achieve high COVID-19 detection performance even using scarce anonymized CXR images, hopeful of solving existing problems of medical dataset sharing.

Our main contributions can be summarized as follows:

- We propose a novel dataset distillation-based method for

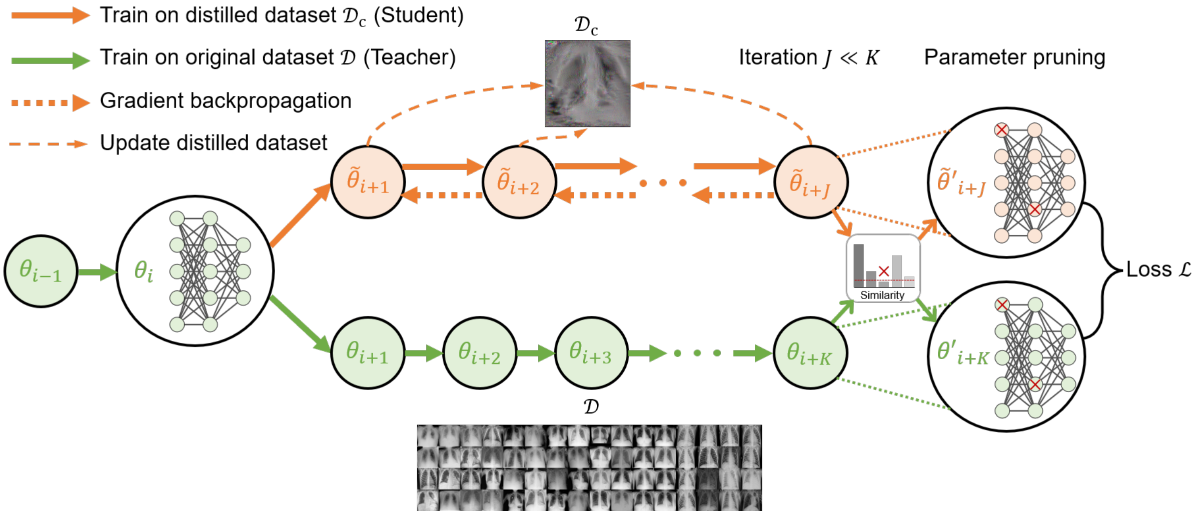

Figure 2: Overview of the proposed method. Our method uses a teacher-student architecture, and the objective is to make the student network parameters $\tilde{\theta}'_{i+J}$ match the teacher network parameters θ'_{i+K}. Our method can avoid the influence of the difficult-to-match parameters on the distilled dataset by pruning the parameters in teacher and student networks.

medical dataset sharing.

- We improve the distillation performance by introducing parameter pruning.
- We verify the effectiveness of the proposed method on a real-world COVID-19 CXR dataset.

Related Work

The acquisition of advanced models relies on large datasets in many fields, which makes storing datasets and training models expensive. An effective way to solve these problems is data selection which identifies representative training samples of large datasets (Bachem, Lucic, and Krause 2017). However, since some of the original data cannot be discarded, there is an upper limit on the compression rate of the data selection method. As a solution, dataset distillation can synthesize a small dataset that preserves most information of the original large dataset. The algorithm of dataset distillation takes a sizeable real dataset as the input and synthesizes a small distilled dataset. Unlike the data selection method that uses actual data from the original dataset, dataset distillation generates synthetic data with a different distribution from the original one (Dong, Zhao, and Liu 2022). Therefore, the dataset distillation method can distill the whole dataset into several images, or even only one image (Li et al. 2022a). Dataset distillation has many application scenarios, such as privacy protection (Li et al. 2020; Song et al. 2022), continual learning (Wiewel and Yang 2021; Sangermano et al. 2022), neural architecture search (Such et al. 2020; Zhao and Bilen 2021c), etc.

Since the dataset distillation task was first introduced in 2018 by Wang et al. (Wang et al. 2018), it has gained increasing attention in the research community[1]. The original dataset distillation algorithm is based on meta-learning and optimizes the distilled images with gradient-based hyperparameter optimization. Subsequently, many works have significantly improved the distillation performance with label distillation (Bohdal, Yang, and Hospedales 2020), gradient matching (Zhao and Bilen 2021c), differentiable augmentation (Zhao and Bilen 2021a), kernel methods (Nguyen, Chen, and Lee 2021; Nguyen et al. 2021; Zhou, Nezhadarya, and Ba 2022), and distribution/feature matching (Zhao and Bilen 2021b; Wang et al. 2022). The recently proposed dataset distillation method by matching network parameters has been the new state-of-the-art (SOTA) on several datasets (Cazenavette et al. 2022). However, a network usually has a large number of parameters. And we found that a few parameters are difficult to match in the distillation process and harm the distillation performance, which could be improved.

Methodology

An overview of the proposed method is shown in Figure 2. The objective of our method is to have the parameters of the student network trained on the distilled dataset match the parameters of the teacher networks trained on the original dataset. Our method consists of three steps, teacher-student architecture training, dataset distillation using parameter pruning, and optimized distilled COVID-19 dataset generation, which we will show details of in the following subsections.

Teacher-Student Architecture Training

Before the distillation process, we first train T teacher networks on the original COVID-19 dataset $\mathcal{D}$ and obtain their parameters. These time sequences of parameters $\{\theta_i\}_0^I$ are defined as teacher parameters. Also, network parameters trained on the distilled dataset $\mathcal{D}_\mathrm{C}$ at each training step i are defined as student parameters $\tilde{\theta}_i$. Our method aims to

[1] https://github.com/Guang000/Awesome-Dataset-Distillation

Algorithm 1: Dataset Distillation using Parameter Pruning

```
Require: $\{\theta_i\}_0^I$ : teacher parameters trained on $\mathcal{D}$; $\alpha_0$: initial value for $\alpha$; $\mathcal{A}$: differentiable augmentation function; $\sigma$: threshold for pruning; $T$: number of distillation step; $J$: number of updates for student network; $K$: number of updates for teacher network; $I^+$: maximum start epoch.
Ensure: optimized distilled dataset $\mathcal{D}_{\text{C}}^*$ and learning rate $\alpha^*$.
Initialize distilled dataset: $\mathcal{D}_{\text{C}} \sim \mathcal{D}$
Initialize trainable learning rate: $\alpha = \alpha_0$
for each distillation step $t = 0$ to $T - 1$ do
    Choose random start epoch $i < I^+$
    Initialize student network with teacher parameter: $\tilde{\theta}_i = \theta_i$
    for each distillation step $j = 0$ to $J - 1$ do
        Update student network with cross-entropy loss: $\tilde{\theta}_{i+j+1} = \tilde{\theta}_{i+j} - \alpha \nabla \ell(\mathcal{A}(\mathcal{D}_{\text{C}}); \tilde{\theta}_{i+j})$
    end for
    if parameter similarity in $\tilde{\theta}_{i+J}$ and $\theta_{i+K}$ is less than $\sigma$ then
        Prune network parameters:
        $\tilde{\theta}'_{i+J}, \theta'_{i+K}, \theta'_i = \text{Prune}(\tilde{\theta}_{i+J}, \theta_{i+K}, \theta_i)$
    end if
    Compute loss between pruned parameters:
    $\mathcal{L} = ||\tilde{\theta}'_{i+J} - \theta'_{i+K}||_2^2 \ / \ ||\theta'_i - \theta'_{i+K}||_2^2$
    Update $\mathcal{D}_{\text{C}}$ and $\alpha$ with respect to $\mathcal{L}$
end for
```

distill CXR images that induce network parameters similar to those learned from the original COVID-19 dataset (given the same initial values). In the distillation process, student parameters are initialized as $\tilde{\theta}_i = \theta_i$ by sampling from one of the teacher parameters at a random step i. We set an upper bound I^+ on the random step i to ignore the less informative later parts of the teacher parameters. Then we perform gradient descent updates on the student parameters $\tilde{\theta}$ with respect to the cross-entropy loss ℓ of the distilled dataset $\mathcal{D}_{\text{C}}$ as follows:

$$\tilde{\theta}_{i+j+1} = \tilde{\theta}_{i+j} - \alpha \nabla \ell(\mathcal{A}(\mathcal{D}_{\text{C}}); \tilde{\theta}_{i+j}), \tag{1}$$

where j and α represent the number of gradient descent updates and the trainable learning rate, respectively. $\mathcal{A}$ represents a differentiable data augmentation module that can improve the distillation performance, which was proposed in (Zhao and Bilen 2021a). Since the data augmentation used during distillation is differentiable, it can be propagated back through the augmentation layers to the distilled dataset.

Dataset Distillation Using Parameter Pruning

Next, we get the student parameters $\tilde{\theta}_{i+J}$ trained on the distilled dataset $\mathcal{D}_{\text{C}}$ from J updates after initializing the student network. Meanwhile, we can get the teacher parameters θ_{i+K} trained on the original COVID-19 dataset $\mathcal{D}$ from K updates, which are the known parameters that have been pre-trained. If the similarity of parameters in $\tilde{\theta}_{i+J}$ and θ_{i+K} is less than a threshold σ, these parameters are recognized as difficult-to-match parameters and are pruned as follows:

$$\tilde{\theta}'_{i+J}, \theta'_{i+K}, \theta'_i = \text{Prune}(\tilde{\theta}_{i+J}, \theta_{i+K}, \theta_i), \tag{2}$$

where Prune represents a function that transforms the parameters to a one-dimensional vector and prunes the parameters under the threshold at each last distillation step. By pruning difficult-to-match parameters in teacher and student networks, the proposed method can avoid the influence of these parameters on the distilled dataset, which can improve the distillation performance. The final loss $\mathcal{L}$ calculates the normalized squared L_2 error between pruned student parameters $\tilde{\theta}'_{i+J}$ and teacher parameters θ'_{i+K} as follows:

$$\mathcal{L} = \frac{||\tilde{\theta}'_{i+J} - \theta'_{i+K}||_2^2}{||\theta'_i - \theta'_{i+K}||_2^2}, \tag{3}$$

where we normalize the L_2 error by the distance $\theta'_i - \theta'_{i+K}$ moved by the teacher so that we can still obtain proper supervision from the late training period of the teacher network even if it has converged. In addition, the normalization eliminates cross-layer and neuronal differences in magnitude.

Optimized Distilled COVID-19 Dataset Generation

Finally, we minimize the loss $\mathcal{L}$ using momentum stochastic gradient descent (SGD) and backpropagate the gradients through all J updates to the student network for updating the pixels of the distilled COVID-19 dataset $\mathcal{D}_{\text{C}}$ and trainable learning rate α. Note that the process of searching the optimized learning rate α^* can act as an automatic adjustment for the number of student and teacher updates (i.e., hyperparameters J and K). Since the distilled CXR images have different visual similarities from the original images, they are automatically anonymized. The distillation process of the proposed method is summarized in Algorithm 1. After obtaining the distilled dataset $\mathcal{D}_{\text{C}}^*$, we can share it with different hospitals and train neural networks for high-accuracy COVID-19 detection.

Experiments

Dataset and Experimental Settings

The dataset used in our study has four classes, i.e., COVID-19 (C), Lung Opacity (L), Normal (N), and Viral Pneumonia (V) (Rahman et al. 2021). The number of images in each class is 3616, 6012, 10192, and 1345, respectively. The resolution of CXR images is 224 × 224, and we resized it to 112 × 112 for distillation.

The network used in this study is a sample 128-width ConvNet (Gidaris and Komodakis 2018) with depth-5, which is often used in current dataset distillation methods. The number of pre-trained teacher networks T was set to 100. We found that pruning too many parameters would cause the model training to crash. Hence, the parameter pruning threshold σ was set to 0.1, which performed well in all experiments. And we set the number of distilled images as 1, 2, 3, 5, 10, and 20 images per class.

For comparative methods, we used the SOTA dataset distillation method MTT (Cazenavette et al. 2022), we also

Table 1: COVID-19 detection accuracy when using different numbers of distilled images. IPC denotes images per class.

IPC	1	2	3	5	10	20	Full Dataset
Ours	**54.2%**	**77.3%**	**78.9%**	**81.6%**	**83.5%**	**84.1%**	88.9%
MTT	52.5%	76.4%	77.0%	79.3%	82.2%	82.7%	

Table 2: COVID-19 detection accuracy of different methods.

Method	**Ours**	MTT	SKD	BYOL	SimSiam	MAE	Transfer	From Scratch
Accuracy	**84.1%**	82.7%	74.2%	68.3%	66.8%	62.3%	53.9%	28.4%

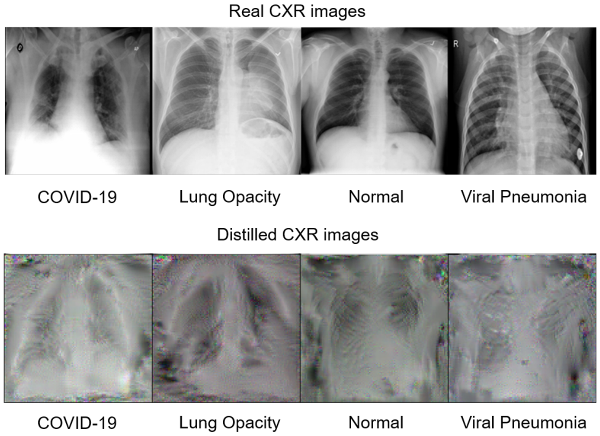

Figure 3: Examples of real and distilled CXR images.

used several SOTA self-supervised learning methods, including SKD (Li et al. 2022b), BYOL (Grill et al. 2020), SimSiam (Chen and He 2021) and MAE (He et al. 2022). Transfer learning from ImageNet (Deng et al. 2009) and training from scratch were used as baseline methods. For MTT, the experimental settings are exactly the same as our method. Except for the MAE method used ViT-Large (Dosovitskiy et al. 2021), all other self-supervised learning methods used ResNet-50 (He et al. 2022) as the backbone network. We randomly selected 42 images per class (1% of the training set) for these self-supervised learning methods. All experiments were conducted using the PyTorch framework with an NVIDIA Tesla P100 GPU with 16G memory.

Results and Discussion

The test accuracy of COVID-19 detection are shown in Tables 1 and 2. From Table 1, we can see that the accuracy of our method increased accordingly as the number of distilled images grew. Furthermore, the proposed method has higher COVID-19 detection accuracy than MTT, which shows the effectiveness of parameter pruning. We also show the upper bound accuracy of 88.9% when training on the full dataset. Even with a compression rate of 0.0047, no significant accuracy degradation is exhibited. Table 2 shows that our method achieved high COVID-19 detection accuracy even when using scarce distilled CXR images. Furthermore, our method drastically outperformed other SOTA methods with a simpler network and fewer training images. Figure 3 shows some examples of real and distilled images. We can see that the distilled images are entirely visually different from the original images, which shows the anonymization effectiveness of the proposed method.

The findings of this paper show the effectiveness of dataset distillation for medical dataset sharing. Although the experimental results are promising, the proposed method should be verified on other medical datasets of different diseases for any potential bias. Since the computational overhead of training and storing teacher parameters is relatively high, which may not necessarily be available in low-resource settings. In addition to MTT, we also did some experiments to verify the usefulness of early distillation algorithms for medical images, but the results were not very effective and computationally intensive. Hence we did not present these results. Furthermore, verifying the validity of distilled medical images on other network structures and in terms of differential privacy will be our future work.

Conclusion

We have proposed a novel dataset distillation-based method for medical dataset sharing. Since the size of the distilled medical image dataset has been significantly compressed and the images are also anonymized, the sharing of medical datasets between different hospitals will be more efficient and secure. We also found that a few parameters in the distillation process are difficult to match, which harms the distillation performance. Based on this observation, we improve the distillation performance by introducing parameter pruning. Experimental results show that we can achieve high COVID-19 detection performance even using scarce anonymized CXR images, hopeful of solving existing problems of medical dataset sharing.

Acknowledgement

This study was supported in part by AMED Grant Number JP21zf0127004, the Hokkaido University-Hitachi Joint Cooperative Support Program for Education and Research, and the MEXT Doctoral Program for Data-Related InnoVation Expert Hokkaido University (D-DRIVE-HU) Program. This study was conducted on the Data Science Computing System of Education and Research Center for Mathematical and Data Science, Hokkaido University.

References

Bachem, O.; Lucic, M.; and Krause, A. 2017. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*.

Bohdal, O.; Yang, Y.; and Hospedales, T. 2020. Flexible Dataset Distillation: Learn Labels Instead of Images. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Workshop*.

Cazenavette, G.; Wang, T.; Torralba, A.; Efros, A. A.; and Zhu, J.-Y. 2022. Dataset Distillation by Matching Training Trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4750–4759.

Chen, X.; and He, K. 2021. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15750–15758.

Dash, S.; Shakyawar, S. K.; Sharma, M.; and Kaushik, S. 2019. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1): 1–25.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Dong, T.; Zhao, B.; and Liu, L. 2022. Privacy for Free: How does Dataset Condensation Help Privacy? In *Proceedings of the International Conference on Machine Learning (ICML)*, 5378–5396.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Gidaris, S.; and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 4367–4375.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 21271–21284.

Grover, S.; Dua, D.; Sahoo, S.; Mehra, A.; Nehra, R.; and Chakrabarti, S. 2020. Why all COVID-19 hospitals should have mental health professionals: The importance of mental health in a worldwide crisis! *Asian Journal of Psychiatry*, 51: 102147.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

Kaissis, G. A.; Makowski, M. R.; Rückert, D.; and Braren, R. F. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311.

Kumar, R.; Wang, W.; Kumar, J.; Yang, T.; Khan, A.; Ali, W.; and Ali, I. 2021. An integration of blockchain and AI for secure data sharing and detection of CT images for the hospitals. *Computerized Medical Imaging and Graphics*, 87: 101812.

Li, G.; Togo, R.; Ogawa, T.; and Haseyama, M. 2020. Soft-Label Anonymous Gastric X-Ray Image Distillation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 305–309.

Li, G.; Togo, R.; Ogawa, T.; and Haseyama, M. 2022a. Compressed Gastric Image Generation Based on Soft-Label Dataset Distillation for Medical Data Sharing. *Computer Methods and Programs in Biomedicine*.

Li, G.; Togo, R.; Ogawa, T.; and Haseyama, M. 2022b. Self-knowledge distillation based self-supervised learning for covid-19 detection from chest x-ray images. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1371–1375.

Minaee, S.; Kafieh, R.; Sonka, M.; Yazdani, S.; and Soufi, G. J. 2020. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. *Medical image analysis*, 65: 101794.

Mofijur, M.; Fattah, I. R.; Alam, M. A.; Islam, A. S.; Ong, H. C.; Rahman, S. A.; Najafi, G.; Ahmed, S. F.; Uddin, M. A.; and Mahlia, T. M. I. 2021. Impact of COVID-19 on the social, economic, environmental and energy domains: Lessons learnt from a global pandemic. *Sustainable production and consumption*, 26: 343–359.

Nguyen, T.; Chen, Z.; and Lee, J. 2021. Dataset Meta-Learning from Kernel Ridge-Regression. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Nguyen, T.; Novak, R.; Xiao, L.; and Lee, J. 2021. Dataset Distillation with Infinitely Wide Convolutional Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 5186–5198.

Rahman, T.; Khandakar, A.; Qiblawey, Y.; Tahir, A.; Kiranyaz, S.; Kashem, S. B. A.; Islam, M. T.; Al Maadeed, S.; Zughaier, S. M.; Khan, M. S.; et al. 2021. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in Biology and Medicine*, 132: 104319.

Sangermano, M.; Carta, A.; Cossu, A.; and Bacciu, D. 2022. Sample Condensation in Online Continual Learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Song, R.; Liu, D.; Chen, D. Z.; Festag, A.; Trinitis, C.; Schulz, M.; and Knoll, A. 2022. Federated Learning via Decentralized Dataset Distillation in Resource-Constrained Edge Environments. *arXiv preprint arXiv:2208.11311*.

Such, F. P.; Rawal, A.; Lehman, J.; Stanley, K.; and Clune, J. 2020. Generative Teaching Networks: Accelerating Neural Architecture Search by Learning to Generate Synthetic Training Data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 9206–9216.

Wang, K.; Zhao, B.; Peng, X.; Zhu, Z.; Yang, S.; Wang, S.; Huang, G.; Bilen, H.; Wang, X.; and You, Y. 2022. CAFE:

Learning to Condense Dataset by Aligning Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12196–12205.

Wang, T.; Zhu, J.-Y.; Torralba, A.; and Efros, A. A. 2018. Dataset Distillation. *arXiv preprint arXiv:1811.10959*.

Wiewel, F.; and Yang, B. 2021. Condensed Composite Memory Continual Learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Zhao, B.; and Bilen, H. 2021a. Dataset condensation with Differentiable Siamese Augmentation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 12674–12685.

Zhao, B.; and Bilen, H. 2021b. Dataset Condensation with Distribution Matching. *arXiv preprint arXiv:2110.04181*.

Zhao, B.; and Bilen, H. 2021c. Dataset Condensation with Gradient Matching. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhou, Y.; Nezhadarya, E.; and Ba, J. 2022. Dataset Distillation using Neural Feature Regression. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.