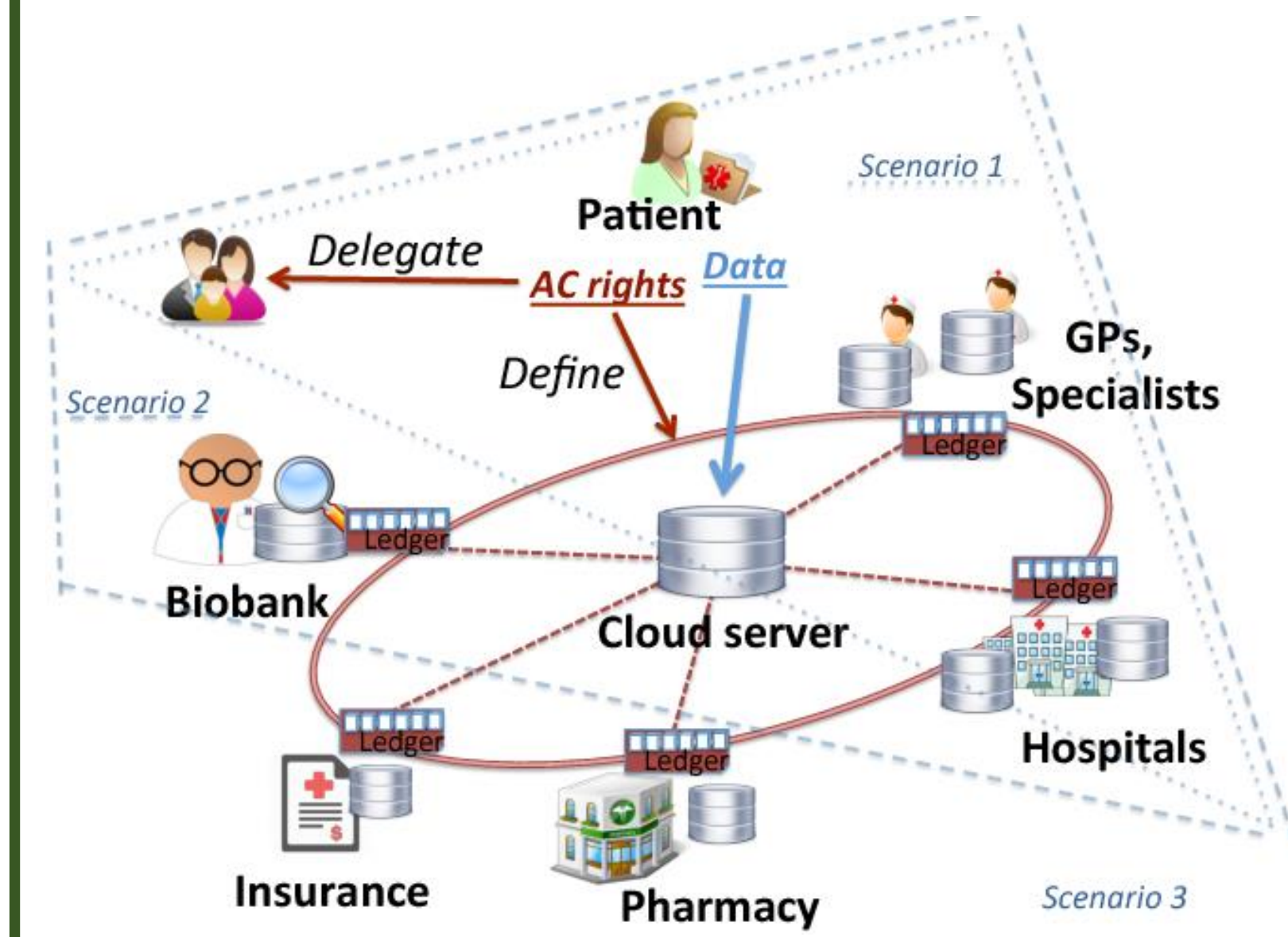


INTRODUCTION

Medical data sharing



- Although a large medical facility may provide adequate healthcare information, a single small medical facility does not have the capacity.
- It is necessary to obtain adequate healthcare information through transmission between multiple hospitals.
- The sharing of medical data plays an essential role in enabling the cross-agency flow of healthcare information and improves the quality of medical services [1].

Medical dataset sharing between hospitals is needed.

Problems

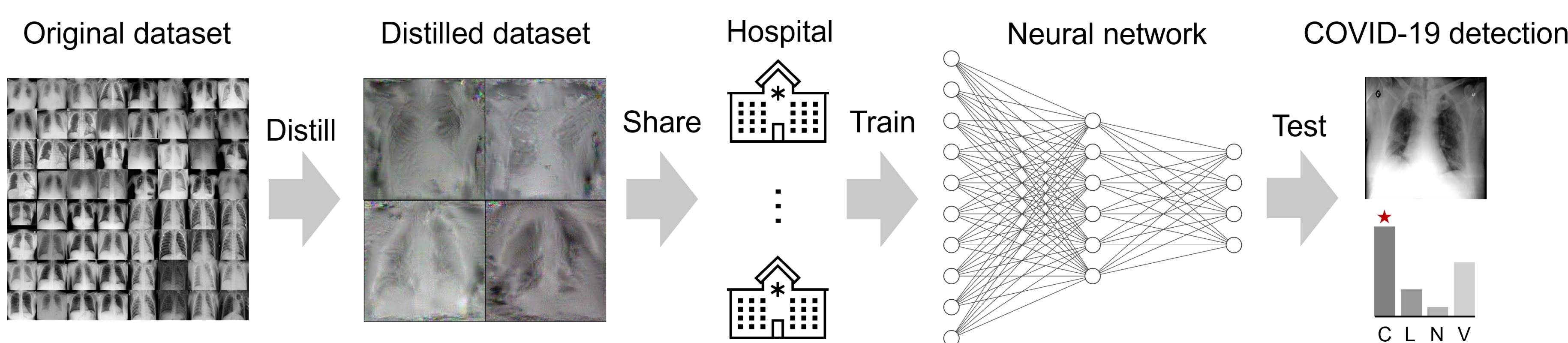


- **Problem 1:** Privacy protection has been a severe issue hindering the process when sharing medical image datasets from different hospitals.
- **Problem 2:** Sharing large-scale high-resolution medical image datasets increases transmission and storage costs.
- The solution to these problems will significantly promote the development of medical dataset sharing [2].

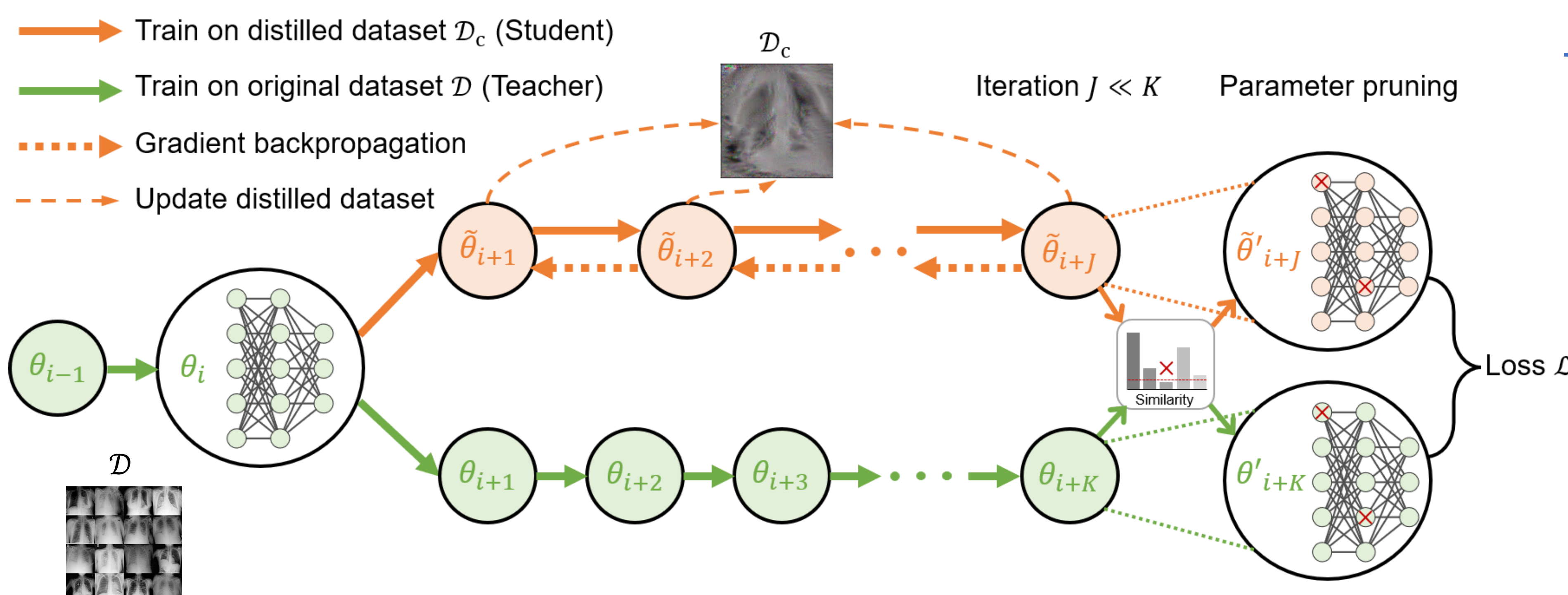
A method to solve existing problems is needed.

PROPOSED METHOD

Concept



Method



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.

Before the distillation process, we first train many teacher networks on the original COVID-19 dataset \mathcal{D} and obtain their parameters.

Then we perform gradient descent updates on the student parameters with respect to the cross-entropy loss of the distilled dataset \mathcal{D}_c .

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 will be pruned.

The final loss \mathcal{L} calculate the normalized L_2 error between the pruned student parameters $\tilde{\theta}'_{i+J}$ and teacher parameters θ'_{i+K} .

Finally, we minimize the loss and backpropagate the gradient through all updates to the student network for obtaining the optimized distilled dataset \mathcal{D}_c .

Novelty: We improve the distillation performance by introducing **parameter pruning**; With the proposed method, the sharing of medical datasets between different hospitals will be more **efficient** and **secure**.

Our method can achieve high COVID-19 detection accuracy even when using scarce distilled chest X-ray images.

EXPERIMENTAL RESULTS

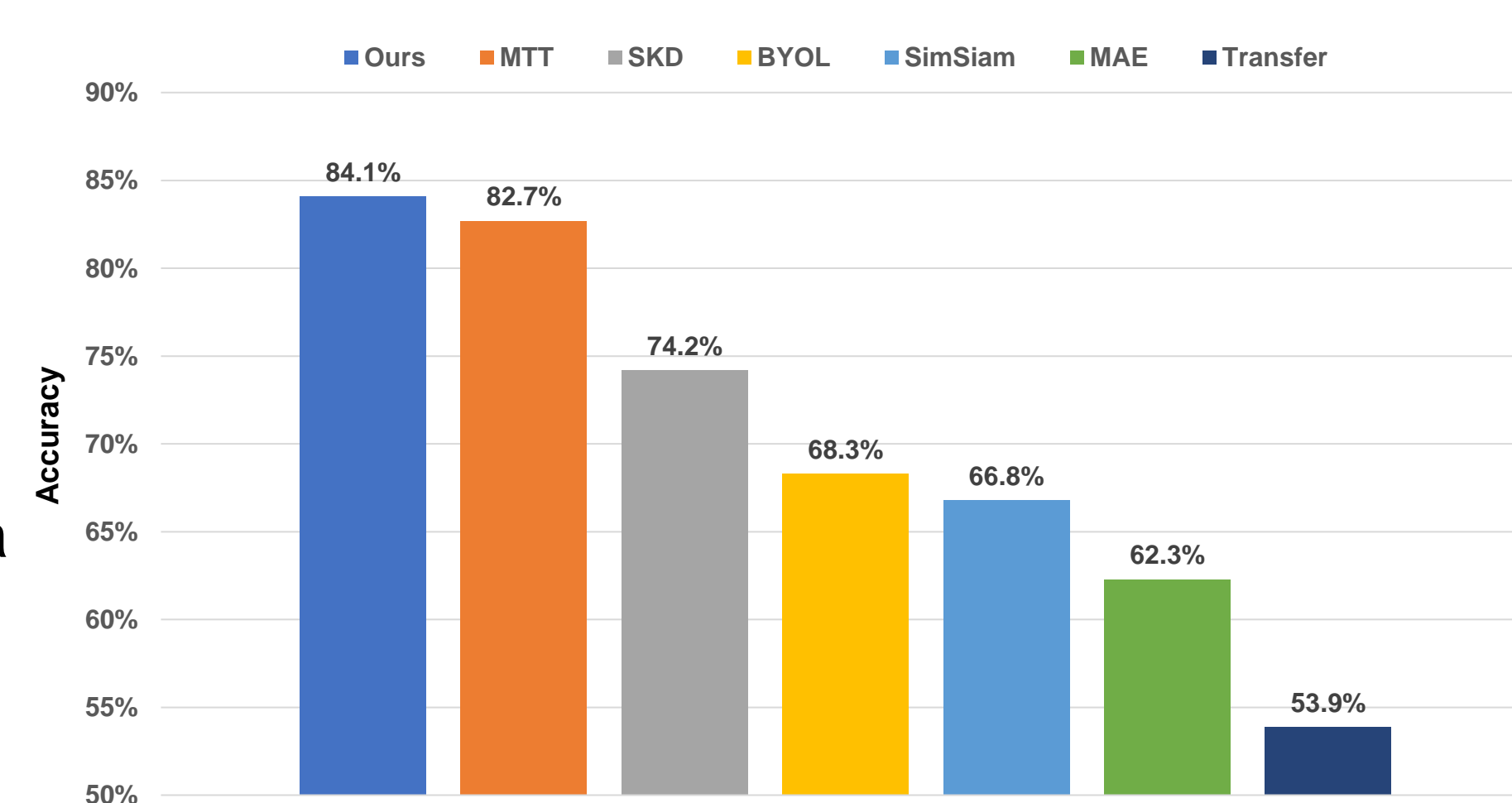
Dataset

The largest open COVID-19 CXR dataset [3]

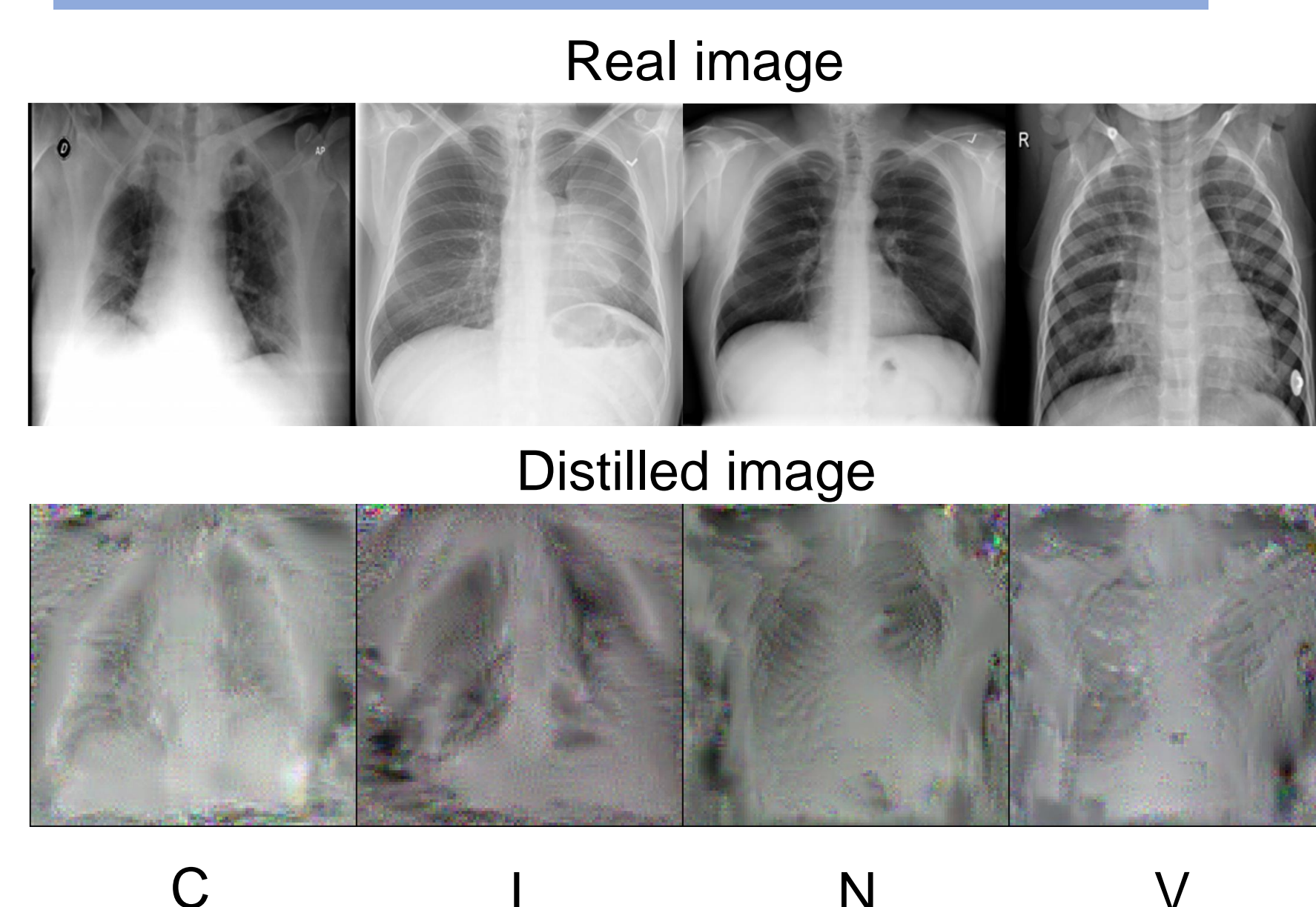
Class	Total	Train	Test
C	3,616	2,893	723
L	6,012	4,810	1,202
N	10,192	8,154	2,038
V	1,345	1,076	269

Ground Truth
C: COVID-19
L: Lung Opacity
N: Normal
V: Viral Pneumonia

Quantitative evaluation



Qualitative evaluation



Settings

Dataset distillation:

- The number of teacher networks: 100
- Threshold σ : 0.1
- The number of distilled data: 20 images per class
- Network structure: 128-width 5-depth ConvNet
- Evaluation: 4-class accuracy

Comparative methods:

- MTT: Matching training trajectories
- SKD: Self-knowledge distillation based self-supervised learning
- BYOL: Bootstrap your own latent
- SimSiam: Simple Siamese representation learning
- MAE: Masked Autoencoder
- Transfer learning from ImageNet

How to Make Artificial Intelligence More Democratic

A new type of learning model uses far less data than conventional AIs, allowing researchers with limited resources to contribute

SCIENTIFIC AMERICAN



This year, GPT-3, a large language model capable of understanding text,

For example, soft distillation techniques have already impacted medical AI research, which trains its models using sensitive health information. In one recent paper [1], researchers used soft distillation in diagnostic x-ray imagery based on a small, privacy-preserving data set.

- Covered by Scientific American, Deccan Herald, 日本経済新聞, 环球科学, 机器之心, 腾讯新闻, 网易新闻
- Cited at TPAMI, NeurIPS, ICML, ICLR, CVPR
- Awesome-Dataset-Distillation (GitHub: 540 stars)
- Most Popular AI Research Aug 2022



[1] Guang Li, et al., "Soft-label anonymous gastric X-ray image distillation," in IEEE ICIP, pp. 305-309, 2020.

[2] Guang Li, et al., "Compressed gastric image generation based on soft-label dataset distillation for medical data sharing," Elsevier CMPB, vol. 227, 107189, 2022.

[3] Guang Li, et al., "Self-knowledge distillation based self-supervised learning for COVID-19 detection from chest X-ray images," in IEEE ICASSP, pp. 1371-1375, 2022.

This study was partly supported by AMED Grant Number JP21zf0127004, the Hokkaido University-Hitachi Collaborative Education and Research Support Program, and the MEXT D-Drive-HU Program.