IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 May 2023, accepted 26 May 2023, date of publication 5 June 2023, date of current version 16 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3283216

RESEARCH ARTICLE

Evaluating Knowledge Transfer in the Neural Network for Medical Images

S. AKBARIAN[1,2], L. SEYYED-KALANTARI[2,3], F. KHALVATI[4,5], (Member, IEEE), AND E. DOLATABADI[2,6,7]

[1]Klick Applied Sciences, Klick Inc., Toronto, ON M4W 3R8, Canada
[2]Vector Institute, Toronto, ON M5G 1M1, Canada
[3]Electrical Engineering and Computer Science Department, Lassonde School of Engineering, York University, Toronto, ON M3J 1P3, Canada
[4]The Hospital for Sick Children, Toronto, ON M5G 1E8, Canada
[5]Institute of Medical Science, University of Toronto, Toronto, ON M5S 1A1, Canada
[6]Faculty of Health, School of Health Policy and Management, York University, Toronto, ON M3J 1P3, Canada
[7]The Institute of Health Policy, Management and Evaluation, University of Toronto, ON M5S 1A1, Toronto, Canada

Corresponding author: E. Dolatabadi (elham.dolatabadi@mail.utoronto.ca)

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant PDF-516984, the NSERC Discovery grant, and by the Vector Institute and Public Health Ontario through Pathfinder Projects. The authors would like to thank Vanessa Allen and Samir Patel.

ABSTRACT The performance of deep learning models, such as convolutional neural networks (CNN)s, is highly dependent on the size of the training dataset. Consequently, it can be challenging to achieve satisfactory performance when training models from scratch in low-data environments. To address this issue, using knowledge transfer approaches from pre-trained networks can be particularly useful. In this study, we implement different experiments for standard transfer learning approaches as our baseline and introduce a novel knowledge transfer approach, called teacher-student learning, to improve the performance of predictive models in diagnostic medical imaging. Specifically, we investigate various configurations in the teacher-student learning framework inspired by the activation attention transfer in computer vision models to help address some challenges faced in medical imaging, such as the limited availability of annotated data and limited computing resources. We show that the teacher-student learning approach holds great promise in significantly enhancing the performance of diagnostic models. The implications of our findings could be instrumental in improving healthcare accessibility and affordability as they may enable the development of cost-effective and widely accessible medical imaging technologies, particularly in limited data environments.

INDEX TERMS Deep learning, medical image, transfer learning, attention transfer, small dataset, teacher-student.

I. INTRODUCTION

Artificial Intelligence has gained attention as essential support for rapid clinical decisions from medical images in several life-threatening diseases [1], [2]. In this context, special attention is given to deep learning algorithms, namely convolutional neural network (CNN) models, which have become the backbone of various medical image computing platforms and clinical diagnosis applications [3], [4], [5]. However, training CNN models from random initialization is compute-intensive, memory-demanding, and requires a large amount of annotated data that is not often accessible in the clinical setting [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues.

These shortcomings have stimulated a large amount of research in the field of knowledge transfer for CNN networks [7], [8], [9], [10], [11], [12], [13], [14], [15]. The most popular knowledge transfer approach is transfer learning which involves using a pre-trained model as a starting point for a new task and fine-tuning it on a smaller dataset [13]. Transfer learning may restrict the network's capacity to prevent overfitting on the small dataset. However, this approach can still be a powerful technique to achieve good performance on the new task with a few data points and computational resources [14], [16]. Transfer learning has been the basis for CNN-based diagnostic imaging such as skin cancer [17], [18], chest X-rays [19], [20], [21], [22], [23], [24], Diabetic Retinopathy [25], [26], [27], [28], [29], Alzheimer's Disease [30], [31], and sleep monitoring [32].

Transfer learning makes a balance between the capacity of the network and the size of the dataset that avoids overfitting. In an empirical study conducted by Raghu et al. [19], it has been shown that the transfer learning from ImageNet to medical images obtained a similar performance in comparison to smaller architectures trained on medical image datasets from scratch. Moreover, Jang et al. [33] also reported that transfer learning may not help if the two tasks and/or datasets are semantically distinct (e.g., transfer learning from and ImageNet to the medical imaging domain). To address these shortcomings and improve performance, researchers have actively studied another type of knowledge transfer in CNNs called the teacher-student learning framework [7], [11], [12], [15]. In this framework, the network providing knowledge is called the teacher, and the network learning the knowledge is called the student. During training, a student network learns to imitate the output of a teacher network or ensemble of networks. Teacher-student learning frameworks have been widely used for performance improvement (especially for small datasets regimes) and/or model compression [7].

Inspired by the growing interest in applying CNN to diagnostic imaging and how to quickly reuse and adapt previously acquired knowledge on new medical tasks and domains, our overarching goal is to leverage knowledge transfer approaches to develop more accurate models requiring less training data. In this study, we conduct experiments with different configurations and training strategies in knowledge transfer and compare their performance on diagnostic labels from medical image datasets. We aim to identify the best setting that yields the highest and most efficient performance where data is scarce or difficult to obtain. The novelty of our work is in the experimentation of various training strategies in knowledge transfer and the introduction of the adoption of a teacher-student learning framework based on an *attention transfer* [12] mechanism for medical images. While attention transfer and knowledge distillation through teacher-student learning framework have been widely used in computer vision, the application to medical images is still relatively new. We focus on four main questions that we found to be fundamental in deriving our experimental analysis:

- How much training data is needed to achieve high performance in knowledge transfer?
- How does knowledge transfer perform when the domains are distinct (i.e., for student initialization and pre-trained teachers)?
- What is the effect of a network's size (parameters) on learning in-domain and cross-domain tasks with various dataset sizes?
- Does knowledge transfer help with overfitting in a low-data environment?

The findings of our experiments have significant implications for healthcare, as they could lead to faster and more accurate diagnoses, which in turn could improve patient outcomes and reduce healthcare costs. Medical image analysis is a critical component of many diagnostic and treatment processes, and any improvements in this area could have a significant impact on healthcare as a whole. This paper is organized as follows: Section II summarizes related works. Section III describes the datasets used in this study. Section IV presents our proposed approach to building knowledge transfer, including transfer learning and teacher-student framework. Section V presents our experiments and results. Section VI discusses the takeaways, addresses the limitations of the current work, and proposes potential future work.

II. RELATED WORKS

A. DIAGNOSTIC MEDICAL IMAGING

The CNN networks are trained on chest X-ray images to provide diagnostic labels and to produce the probability of multiple diseases per image. Chest X-rays are utilized to diagnose a wide range of diseases, such as thorax disease [34], Tuberculosis [3], Pneumonia [4], and COVID-19 [35]. Enriched with access to large public hospital-scale datasets [20], [22], [36], [37], CNNs have been utilized for abnormality classification on medical chest X-rays images [4], [20], [21], [22], [23], [24]. Transfer learning is also widely used to expand the application of CNNs as a chest X-ray diagnostic tool on small datasets tools [4], [20], [21], [22], [23], [24]. The popular CNN network architecture that researchers trained for X-ray diagnostic classifiers [4], [20], [21], [23], [24], [38] is DenseNet [39]. In addition to DenseNet, Irvin et al. [20] have applied several other CNN models, including ResNet-152, Inception-v4, and SE-ResNeXt-101, on X-ray images; however, the DenseNet-121 architecture produces the best results in practice.

B. KNOWLEDGE TRANSFER

Transfer learning showed poor performance for semantically distinct datasets and prevented the CNN parameters from significant updates. To address these shortcomings, researchers proposed several advanced approaches, such as Knowledge Distillation (KD) in the neural network that transfers knowledge between a teacher and a student network [7]. The original idea behind the KD came from Bucilua et al. [8], who proposed compressing the knowledge of several large ensemble base-level classifiers into a single smaller and faster model. This idea reduced the computation and memory complexity of their models. The KD was later generalized by Hinton et al. [7] in which a piece of knowledge is transferred from a large CNN (teacher) to a small network (student) by minimizing the difference between the logits (the inputs to the final softmax) produced by the teacher model and those produced by the student model. Yim et al. [15] proposed an approach that minimized the distance between the intermediate layers of the teacher and student networks. This method helped with faster optimization and better performance of the student network compared to a trained CNN from scratch. Moreover, using their approach, the student CNN can learn the distilled knowledge from a teacher CNN trained for a different task. Romero et al. [11] also proposed another teacher-student framework called FitNet, where they

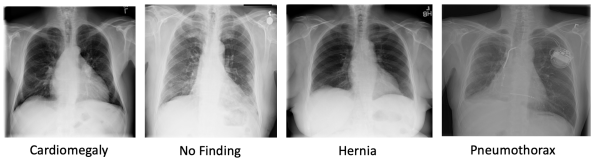

FIGURE 1. Four images from the ChestX-ray14 [22] dataset show Cardiomegaly, no finding, Hernia, and pneumothorax.

introduced intermediate-level hints from the teacher's hidden layers beside output layers to facilitate the training process of the student network. Using FitNet, the student network can also learn from the intermediate representation of the teacher network. FitNet can train deep student models with fewer parameters, which can generalize better and/or run faster than their teachers. Zagoruyko et al. [12] proposed another teacher-student knowledge transfer using teacher's feature maps, called Attention transfer. Using this approach, given the spatial attention maps of a teacher network, the student network is trained to learn the exact behavior of the teacher network by trying to replicate its output at a layer receiving attention from the teacher. The number of attention transfers and position of the layers depends on whether low-, mid-, and high-level representation information is required.

III. MATERIALS AND METHODS

In this study, we implemented various configurations and training strategies in the knowledge transfer for training CNN classifiers and compared their performance on diagnostic labels from medical image datasets. More specifically, we investigated networks' initialization, network size, and networks' knowledge domain on classification performance using standard transfer learning and teacher-student learning framework, as shown in Fig. 2.

A. DATASETS

Three different publicly available medical imaging datasets were used in this study: CheXpert [20], ChestX-ray14 (a sample of 5k 1), and MIMIC-CXR [36]. Fig. 1 shows some sample images included in the ChestX-ray14 dataset.

CheXpert. The CheXpert [20] is a chest X-ray dataset comprising 223,648 frontal and lateral images of 64,740 patients. Each image in the dataset has 14 multilabel annotations associated with diagnostic labels for 13 diseases: Enlarged Cardiomediastinum, Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices, and No Finding.

ChestX-ray14. The original ChestX-ray14 [22] includes 112,120 frontal X-ray images from 30,805 unique patients. However, in this study, we use a small sample (5%) of the dataset, including 5,606 images[1]. The ChestX-ray14 dataset includes 15 multiclass annotations for 14 diseases: Hernia, Pneumonia, Fibrosis, Edema, Emphysema, Cardiomegaly, Pleural Thickening, Consolidation, Pneumothorax, Mass, Nodule, Atelectasis, Effusion, Infiltration, and No Finding.

MIMIC-CXR. The MIMIC-CXR [36] is a chest X-ray dataset composed of 371,858 frontal and lateral images of 65,079 patients. The annotation of each X-ray image is 14 diagnostic diseases similar to CheXpert.

The labels were automatically extracted from the radiologist reports, using natural language processing techniques for all chest X-ray datasets we used in our study, including CheXpert, MIMIC-CXR, and ChestX-ray14. For CheXpert and MIMIC-CXR, in particular, the disease labels were from the set of {positive, negative, not mention, or uncertain} conditions. In this study, all ''non-positive'' labels were mapped to zero, similar to the ''U-zero'' study in [20]. In all three chest X-ray datasets, the ''No Finding'' label indicates the absence of any diseases. The demographic of datasets is shown in Table 1, and a sample of images used for this study is shown in Fig 1.

B. MODEL DESCRIPTIONS

We used DenseNet as the backbone for the CNN classifiers. We implemented two versions of DenseNets, a larger (DenseNet-121) and a lighter (DenseNet-40), to explore the impact of the network size on the performance. The DenseNet-40 is the lighter version of DenseNet-121 where the last two blocks of DenseNet-121 were removed, which we call DenseNet-40 for the rest of this paper (detail of network architecture is shown in Appendix -C). For the DenseNet-40 ($1.4m$ trainable parameters), we didn't freeze any layers of the network during the classification tasks. For the DenseNet-121, depending on the experiment, we let either all the 121 layers ($7.0m$ trainable parameters) or the last 34 layers ($2.4m$ trainable parameters) of the network be tuned during the tasks. For the weights initialization effect, CNN networks were initialized with either pre-trained ChestX-ray14 (in-domain) or ImageNet (cross-domain) weights. We illustrate the training configurations of knowledge transfer for CNN models in Fig. 2.

For the teacher-student learning framework, we built an activation-based attention transfer [12] to transfer knowledge from a layer of the teacher network to the student network. In our setting, the knowledge was transferred between the one layer before the last layer of the last DenseBlock of both the teacher and student networks with similar spatial resolutions. The student network's new loss is as below:

$$L_{tot} = CE_S + \beta L_{AT}, \tag{1}$$

where CE_S is the standard cross-entropy loss, L_{AT} is the attention loss, and β is a coefficient that controls the contribution of L_{AT} in the L_{tot}. The L_{AT} is the l_2 norm between the student's and teacher's normalized attention maps averaged across all feature planes, C.

$$L_{AT} = \frac{1}{C}\sum_{j=1}^{C} || \frac{Q_S^j}{||Q_S^j||_2} - \frac{Q_T^j}{||Q_T^j||_2} ||_2, \tag{2}$$

[1] https://www.kaggle.com/nih-chest-xrays/data

TABLE 1. A summary of medical imaging datasets used in this study.

Dataset	# Labels	Labeling Method	Images view	# Images	# Patients
MIMIC-CXR [36]	14	Automatic	Frontal/Lateral	371,858	65,079
CheXpert [20]	14	Automatic	Frontal/Lateral	223,648	64,740
ChestX-ray14 [22]	15	Automatic	Frontal	112,120	30,805

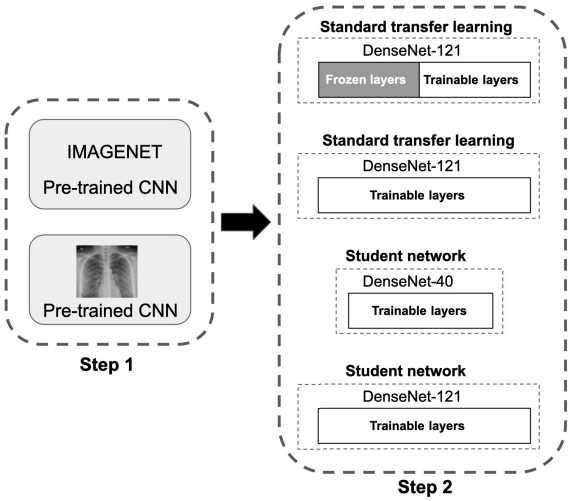

FIGURE 2. Illustration of various configurations for training CNN classifiers using knowledge transfer approaches for medical image tasks, including adjusting the network's initialization, network size, and the pre-trained teachers' domain. In step 1, we select the domain of the CNN model as a teacher in teacher-student learning or network initialization in standard transfer learning. In step 2, we choose the network size and trainable layers.

where Q_T^j and Q_S^j are the *j-th* attention maps of teacher and student networks. Please refer to the appendix -C for more details on the Teacher-Student Learning framework using the Attention Transfer map for knowledge transfer from a powerful CNN teacher to a CNN student.

For the teacher-student learning framework, the knowledge is transferred from a teacher either pre-trained on ImageNet ($\text{Teacher}_{\text{ImageNet}}$) or MIMIC-CXR ($\text{Teacher}_{\text{MIMIC-CXR}}$) [23].

IV. RESULTS

Here we report our observations across the four dimensions fundamental to this work: (1) Training Dataset size, (2) Teacher's Domain, (3) Student Networks' Trainable and Initialization Weights, and (4) Overfitting.

Adam [40] was used to optimize the loss function which was multilabel binary cross-entropy with and without attention transfer loss, Eq. (1), in all of the experiments. The learning rate was decreased by a factor of 2 over every 16 epochs from an initial value of 5×10^{-5} as suggested in [23]. For all the experiments, the CNN models were trained for a maximum of 128 epochs with a batch size of 32 to fit data in Nvidia Titan XP 12 *GB* GPU used for training the CNN models. All evaluations have been made based on three repetitions of each model. The best models were selected based on the performance of the average of multilabel AUC on the validation set. To find an optimal value for the β coefficient, we performed a grid search in the range of 1 to 2000 on the validation set. The β coefficient is reversely related to total loss - that is, when the β decreases, the impact of attention loss increases.

All the images were resized (256×256), centered, and cropped. Additionally, $-15°$ to $+15°$ random rotation and random horizontal flip were applied to the training dataset. Following [4], [20], and [21], images were normalized using the mean and standard deviation of the ImageNet. All datasets split into the train-validation-test-set with no patient shared across the splits.

Table 2 shows the averaged AUC scores ($\pm$ the 95% confidence intervals (CI)) calculated over three runs across all diagnostic labels. For these experiments, the CNN models were trained on the entire CheXpert training data, including $178k$ samples (CheXpert_{178k}) and the three subsets of $1k$, $5k$, and $50k$ data (Dataset column), randomly sampled from the CheXpert_{178k} dataset. As evidenced by Table 2, more training data increases classification performance, whereby performance on the CheXpert_{178k} was the highest. However, based on our observation, the performance improved slightly (less than 1%) after using 50k training data.

Table 3 presents the classification performance (AUC scores) of teacher-student learning with in-domain (MIMIC-CXR) and cross-domain (ImageNet) teachers. We can observe that in-domain pre-trained teachers perform reasonably on small training data (1k, 5k, and 50k). Another observation is that the impact of the teacher's domain on classification performance decreases as the sample size increases.

A glance at Table 2 and Table 3 from the perspective of network size (*tp) and network initialization (ChestX-ray14 and ImageNet) reveals that the best performance in all the experiments was achieved when the networks had 7.0 million trainable parameters. Moreover, in-domain initialization for students' networks improved the performance in the low training environment. But as the training sample size increased, the performance was not dependent on initialization anymore.

Fig. 3 illustrates the AUC learning curves of (a) large and (b) light CNNs trained on CheXpert_{5k} as well as (c) large and (d) light CNNs trained on $\text{ChestX-ray14}_{5.6k}$. As evidenced in Fig. 3, teacher-student learning improves performance and acts as a regularizer reducing overfitting.

To further explore the performance of teacher-student learning in a low data environment, we conducted a similar experiment on a different dataset, a small sample from $\text{ChestX-ray14}_{5.6k}$, and the results are presented in

TABLE 2. The area under the receiver operating characteristic curve (AUC) score ± the 95% confidence intervals (CI). The best scores are in bold and *tp denotes trainable parameters. Here St $_{\text{DenseNet-121}}$ and St $_{\text{DenseNet-40}}$ denote the student networks which are DenseNet-121 in Table A and DenseNet-40 in Table B, respectively. In addition, Teacher $_{\text{MIMIC-CXR}}$ is a network pre-trained on the MIMIC-CXR dataset.

A) Large CNN (DenseNet-121)

Dataset$_{Size}$	Initialization	**Transfer Learning**		**Attention Transfer**
		DenseNet-121		Teacher $_{\text{MIMIC-CXR}}$ St $_{\text{DenseNet-121}}$
		tp*	7.0m	2.7m	7.0m
CheXpert$_{1k}$	ChestX-ray14		68.26±0.07	70.07±0.03	**73.05±0.10**
	ImageNet		68.06±0.14	67.50±0.24	**72.33±0.12**
CheXpert$_{5k}$	ChestX-ray14		73.16±0.14	73.47±0.56	**76.67±0.03**
	ImageNet		72.86±0.18	71.23±0.12	**76.60±0.03**
CheXpert$_{50k}$	ChestX-ray14		78.23±0.11	76.99±0.03	**79.20±0.06**
	ImageNet		78.20±0.13	76.01±0.14	**79.36±0.06**
CheXpert$_{178k}$	ChestX-ray14		**80.11±0.05**	78.40±0.04	80.05±0.04
	ImageNet		**80.47±0.14**	78.47±0.07	80.25±0.39

B) Light CNN (DenseNet-40)

Dataset$_{Size}$	Initialization	**Transfer Learning**		**Attention Transfer**
		DenseNet-40		Teacher $_{\text{MIMIC-CXR}}$ St $_{\text{DenseNet-40}}$
		tp*	1.4m	1.4m
CheXpert$_{1k}$	ChestX-ray14		68.01±0.01	**69.31±0.02**
	ImageNet		68.58±0.09	**70.63±0.22**
CheXpert$_{5k}$	ChestX-ray14		72.47±0.05	**73.05±0.02**
	ImageNet		71.83±0.23	**75.45±0.30**
CheXpert$_{50k}$	ChestX-ray14		76.99±0.03	**77.82±0.13**
	ImageNet		76.01±0.14	**78.78±0.04**
CheXpert$_{178k}$	ChestX-ray14		79.72±0.06	**79.74±0.24**
	ImageNet		78.47±0.07	**80.01±0.05**

TABLE 3. The AUC score ± 95% CI. The best scores are in bold and *tp denotes trainable parameters. Here, St $_{\text{DenseNet-X}}$ denotes the student network which is DenseNet-X (X=121, 40). Also, Teacher $_{\text{MIMIC-CXR}}$ and Teacher $_{\text{ImageNet}}$ are networks pre-trained on MIMIC-CXR and ImageNet datasets, respectively.

Dataset	**Attention Transfer**				
		Teacher $_{\text{ImageNet}}$		Teacher $_{\text{MIMIC-CXR}}$	
	Initialization	St $_{\text{DenseNet-121}}$	St $_{\text{DenseNet-40}}$	St $_{\text{DenseNet-121}}$	St $_{\text{DenseNet-40}}$
		tp*	7.0m	1.4m	7.0m	1.4m
CheXpert$_{1k}$	ChestX-ray14	68.67±0.13	68.00±0.10	**73.05±0.10**	69.31±0.02
	ImageNet	68.02±0.05	68.42±0.32	**72.33±0.12**	70.63±0.22
CheXpert$_{5k}$	ChestX-ray14	73.30±0.11	72.44±0.05	**76.67±0.03**	73.05±0.02
	ImageNet	72.64±0.12	72.22±0.10	**76.60±0.03**	75.45±0.30
CheXpert$_{50k}$	ChestX-ray14	78.10±0.11	77.42±0.04	**79.20±0.06**	77.82±0.13
	ImageNet	78.27±0.19	78.23±0.18	**79.36±0.06**	78.78±0.04
CheXpert$_{178k}$	ChestX-ray14	**80.06±0.12**	79.67±0.03	80.05±0.04	79.74±0.24
	ImageNet	**80.47±0.19**	80.40±0.06	80.25±0.39	80.01±0.05

Appendix -D. We reached similar findings indicating knowledge transfer from an in-domain teacher, Teacher $_{\text{MIMIC-CXR}}$, holds the promise of model improvement for small datasets.

V. DISCUSSION

Integrating advanced deep learning-based computer vision models with the skills and expertise of radiologists will bring significant medical benefits to patients. However, the reliability and accuracy of these models require accessing and training on large datasets, which is a challenge in the healthcare setting. Therefore, researchers are developing new knowledge transfer approaches to train deep CNN models on small datasets (e.g., a few hundred/thousands). In this study, we presented experiments providing further insights into the

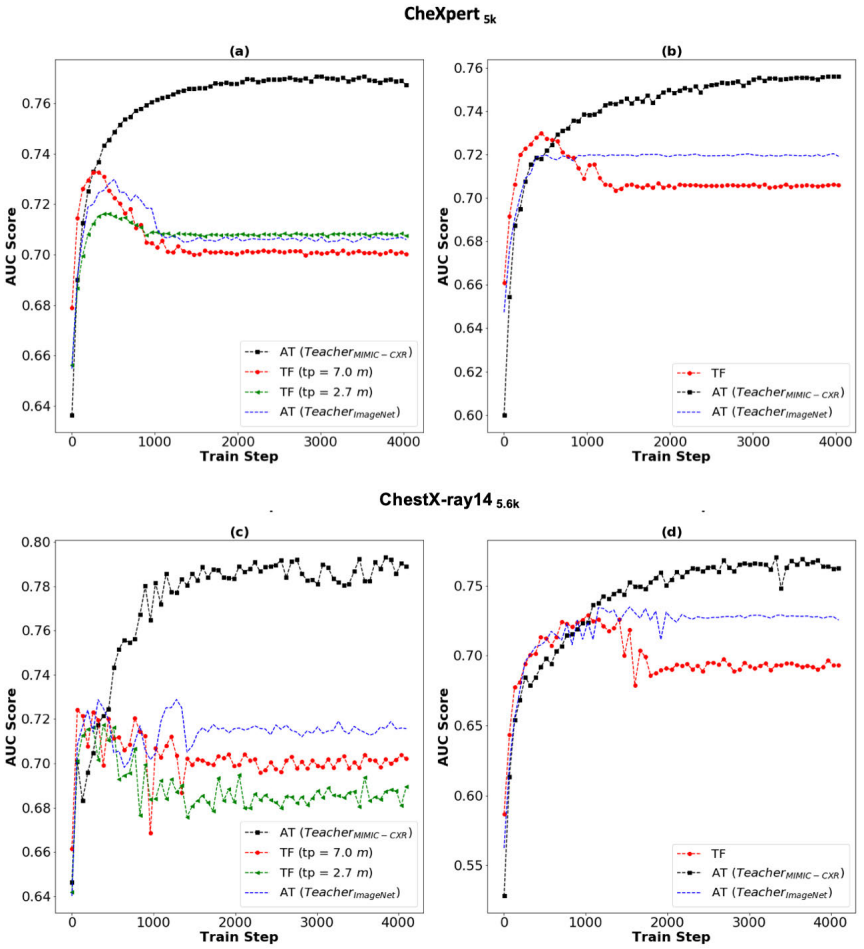

FIGURE 3. The vertical and horizontal axes indicate the AUC score and training steps, respectively, for training CNN models on CheXpert$_{5k}$ (top) and ChestX-ray14$_{5.6k}$ (bottom). Student network is DenseNet-121 in (a) and (c), and DenseNet-40 in (b) and (d). Teacher-student learning framework using Attention Transfer (AT) as a regularizer delays the overfitting and makes the training more robust in comparison to the transfer learning (TF) approach. However, it might slow down the convergence, but it allows the student network to continue training.

effectiveness of two knowledge transfer methods: standard transfer learning and teacher-student learning framework for training CNN models for medical image classification tasks. Below we discuss the highlights of our experiments.

Dataset Size: Based on our findings, teacher-student learning outperformed standard transfer learning by a wide margin in most experiments, especially in a low-data environment. However, as the number of training data increases, teacher-student learning and standard transfer learning perform the same. In fact, with a large dataset, the network can learn the task by optimizing the cross-entropy loss from the data alone, and there is less need to reuse the previously learned information through knowledge transfer.

Network Size: Our experiments show that the size of the network and the number of trainable parameters can have a significant impact on classification performance. It is interesting to note that small student networks still perform well in teacher-student learning settings due to the guidance of a powerful teacher. This fact highlights the importance of choosing the right training approach for a given task and the potential benefits of leveraging pre-trained teachers.

In/Cross Domain: As expected, in-domain knowledge transfer outperforms cross-domain for small datasets, regardless of the student network's initialization. However, as the training dataset size increased (178k), the model performance was less dependent on the teacher domain. This fact indicated that the network could learn the details from the dataset directly without using the teacher's knowledge. It highlights the importance of considering the size of the training dataset and the choice of teacher domain when performing knowledge transfer. It also suggests that as the training dataset size increases, the network becomes less reliant on the teacher's knowledge, which is an important consideration when designing any knowledge transfer learning approaches.

Regularization: In this study, we showed that teacher-student learning improves performance and serves as a

regularizer to delay overfitting regardless of network size and the pre-trained teacher domain. This fact suggests that teacher-student learning could be a useful technique for improving the performance and stability of CNNs.

Significance: Our findings will push forward the democratization of CNN models and accelerate their adoption in small medical imaging regimes and limited computing resources. The findings of this study are comparable with the results of previously published studies showing standard transfer learning has limited performance gains. In particular, [19] and [33] showed that transfer learning fails to improve the performance where the ratio of network parameters to data size is large, the network is initialized with cross-domain weights, and a large number of network parameters are frozen while training the model. By addressing these shortcomings and enabling more efficient training without freezing parameters, teacher-student learning has the potential to significantly improve performance and accelerate the adoption of ML-driven predictive models in medical imaging. This can ultimately lead to better and faster diagnoses, more personalized treatments, and improved patient outcomes. Additionally, this work can contribute to the broader field of transfer learning by exploring novel techniques for training deep models with small datasets in various fields beyond medical imaging.

A. LIMITATIONS AND FUTURE DIRECTIONS

Our work raises several opportunities to elaborate knowledge transfer in diagnostic imaging further. In this study, we focused on image classification in medical imaging; however, the same technique could also apply to image segmentation [41], other types of medical images (e.g., microorganism, histopathology images, etc.) [42], [43], [44], [45], video analysis [46], and object detection [47]. Also, Attention transfer is one of the commonly used teacher-student learning frameworks. But a potential future work could be investigating more sophisticated knowledge transfer approaches, especially the knowledge distillation on medical imaging. The other direction remaining is exploring the rationale of a model decision by explaining CNN's attention maps for each disease. Using the same technique, researchers could obtain the location of the medical images that the teacher network pays more attention to for the final decision and compare it with the regions a physician used for diagnosis to justify the teacher's capability. In addition, inspired by Tian et al. [48], potential future research is to combine attention transfer with few-shot learning techniques to learn a good embedding that can generalize well for a novel class.

VI. CONCLUSION

The limited availability of large annotated datasets usually restricts researchers from building ML-driven medical image diagnostic tools. This fact highlights the importance of exploring novel techniques for training ML models with access to small training datasets. Here, we adopted an advanced knowledge transfer technique based on attention transfer concept, teacher-student learning framework, to train medical image classifiers for chest X-ray pathology
diagnosis. We then compared the performance of the teacher-student learning framework with the widely used transfer learning approach through a series of experiments. Our analysis revealed that the teacher-student learning framework outperforms transfer learning regardless of the number of training parameters and model initialization in a low-data environment. Finally, the ability of teacher-student learning to serve as a regularizer during the training process and delay overfitting is also noteworthy. In summary, teacher-student is a more efficient and reliable knowledge transfer approach, especially when access to large training data is limited.

TABLE 4. The architecture of light DenseNet called DenseNet-40. In DenseNet-40, we removed the last two blocks of DenseNet-121 for the purpose of our exploration.

Layers	Output size	Filters
Convolution	128 × 128	7 × 7 conv, stride 2
Pooling	64 × 64	3 × 3 max pool, stride 2
Dense Block (1)	64 × 64	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	64 × 64	1 × 1 conv
	32 × 32	2 × 2 average pool, stride 2
Dense Block (2)	32 × 32	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Pooling	8 × 8	8 × 8 Adaptive Average Pool
Convolution	8 × 8	1 × 1 conv
Convolution	8 × 8	1 × 1 conv
Pooling	1 × 1	1 × 1 Adaptive Average Pool
Classification Layer		fully-connected, Softmax / Sigmoid

APPENDIX

We organize the appendix into four sections:

A. DETAILS OF LIGHT NETWORK (DenseNet-40)

We explored the performance of the attention transfer framework in the medical imaging setting for both large and light CNN students. For the large student network, we used DenseNet-121, and for the light student network called DenseNet-40, we removed the last two blocks of DenseNet-121, and the architecture details are shown in Table 4.

B. DATA AVAILABILITY

This work used three publicly available datasets that were subject to data use agreements, which we followed all protocols associated with. These datasets were observational and retrospective, and their sources were referenced in the paper. The MIMIC-CXR [36] dataset can be accessed at https://physionet.org/content/mimic-cxr/2.0.0/, the CheXpert

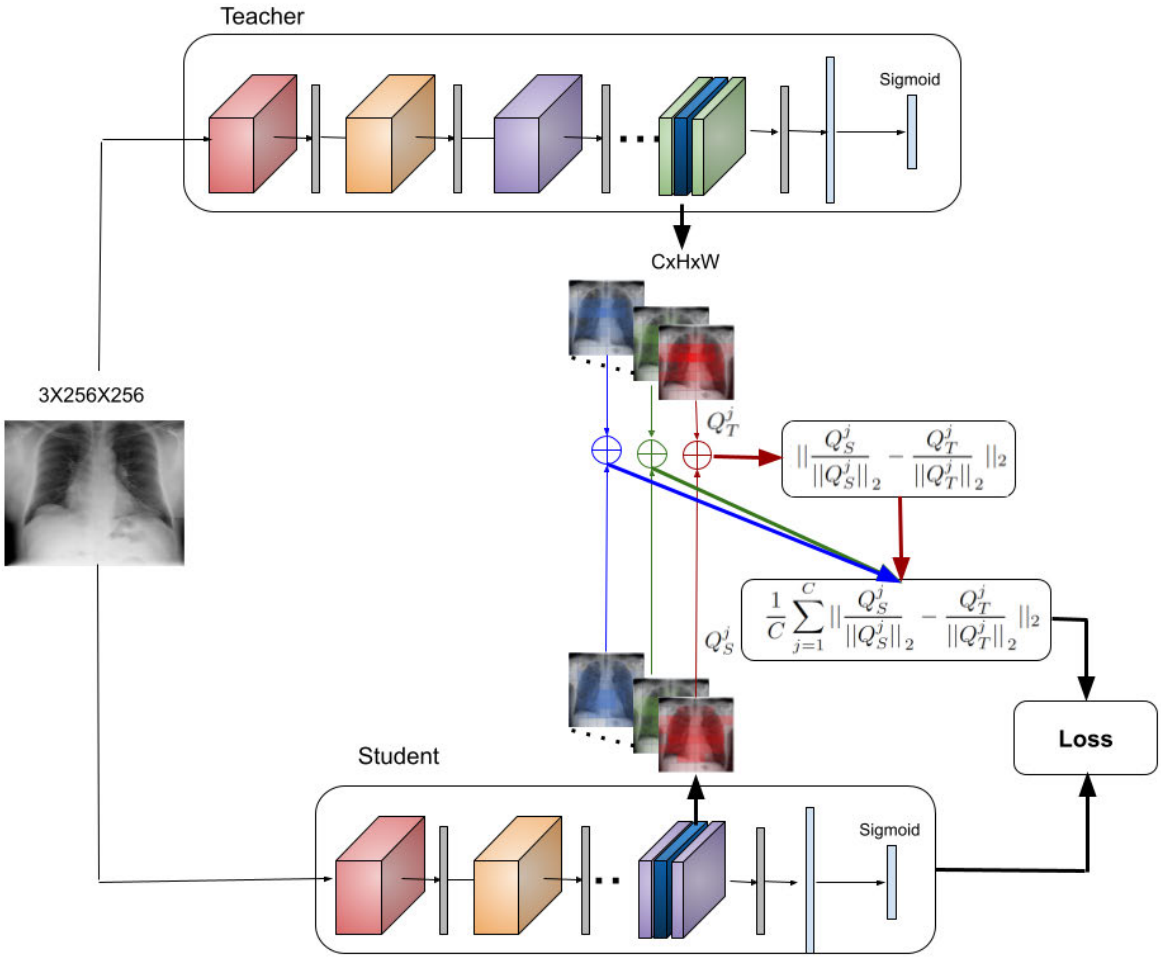

FIGURE 4. Illustration of Teacher-Student Learning framework using Attention Transfer map for knowledge transfer from a powerful CNN teacher to a CNN student. During training, the student network learns similar spatial attention maps to those of an already pre-trained teacher in order to make a good prediction. In our setting, the transfer of knowledge occurs between the one layer before the last layer of the last dense blocks of both the teacher and student networks. In the shown example, the spatial attention map ($H \times W$) is 8 × 8, and there are 32 feature planes (C).

TABLE 5. The AUC score ± 95% confidence intervals (CI). The best is in bold and *tp is the trainable parameters numbers. Here St $_{\text{DenseNet-X}}$ denote the student network DenseNet-X, X=121, 40.

Dataset	Imagenet Initialization							
	Transfer Learning			Attention Transfer				
	DenseNet-121		DenseNet-40	Teacher $_{\text{ImageNet}}$		Teacher $_{\text{MIMIC-CXR}}$	
				St $_{\text{DenseNet-121}}$	St $_{\text{DenseNet-40}}$	St $_{\text{DenseNet-121}}$	St $_{\text{DenseNet-40}}$
tp*	7.0m	2.7m	1.4m	7.0m	1.4m	7.0m	1.4m
ChestX-ray14$_{5.6k}$	71.45± 0.98	70.07 ± 0.60	71.55± 0.19	71.66± 1.18	72.45 ± 1.09	**80.45± 0.38**	75.79± 1.17

[20] dataset at https://stanfordmlgroup.github.io/competitions/chexpert/, and the ChestX-ray14 [22] dataset at https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community. Access to all three datasets required user registration and a signed data use agreement. Only the MIMIC-CXR dataset required an additional credentialing process, which can be completed through the PhysioNet website. The MIMIC-CXR project page on PhysioNet describes the data access procedure.

C. TEACHER-STUDENT LEARNING FRAMEWORK USING ATTENTION TRANSFER MAP

For the teacher-student learning framework, we built an activation-based attention transfer to transfer knowledge from a powerful CNN teacher to a CNN student. The detail of the overall method is shown in Fig 4

D. RESULT OF ChestX-ray14$_{5.6k}$

Table 5 shows the performance of the attention transfer and transfer learning on small datasets of ChestX-ray14$_{5.6k}$.

Overall, attention transfer outperformed transfer learning in all scenarios. At a high level, we observed that the teacher network pre-trained on MIMIC-CXR substantially improves the performance on ChestX-ray14$_{5.6k}$ (AUC = 80.45±10.38) datasets. In terms of student size, for in-domain knowledge transfer (Teacher $_{\text{MIMIC-CXR}}$), the larger student network (DenseNet-121) with 6,968,206 trainable parameters outperforms a lighter student network (DenseNet-40) with 1,364,142 trainable parameters. However, for cross-domain attention transfer (Teacher $_{\text{ImageNet}}$), the lighter student performed better than the larger student. Similar to our previous experiment, the in-domain teacher increased student performance in comparison to the cross-domain teacher.

REFERENCES

[1] A. Rimmer, "Radiologist shortage leaves patient care at risk, warns royal college," *BMJ*, vol. 359, p. j4683, Oct. 2017.

[2] S. Bastawrous and B. Carney, "Improving patient safety: Avoiding unread imaging exams in the national VA enterprise electronic health record," *J. Digit. Imag.*, vol. 30, no. 3, pp. 309–313, Jun. 2017.

[3] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, Aug. 2017, doi: 10.1148/radiol.2017162326.

[4] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*.

[5] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift Für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, May 2019.

[6] G. Torres-Mejía, R. A. Smith, M. D. L. L. Carranza-Flores, A. Bogart, L. Martínez-Matsushita, D. L. Miglioretti, K. Kerlikowske, C. Ortega-Olvera, E. Montemayor-Varela, A. Angeles-Llerenas, S. Bautista-Arredondo, G. Sánchez-González, O. G. Martínez-Montañez, S. R. Uscanga-Sánchez, E. Lazcano-Ponce, and M. Hernández-Ávila, "Radiographers supporting radiologists in the interpretation of screening mammography: A viable strategy to meet the shortage in the number of radiologists," *BMC Cancer*, vol. 15, no. 1, p. 410, May 2015, doi: 10.1186/s12885-015-1399-2.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[8] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 535–541.

[9] Y. Jang, H. Lee, S. J. Hwang, and J. Shin, "Learning what and where to transfer," 2019, *arXiv:1905.05901*.

[10] S. Gutstein, O. Fuentes, and E. Freudenthal, "Knowledge transfer in deep convolutional neural nets," *Int. J. Artif. Intell. Tools*, vol. 17, no. 3, pp. 555–567, Jun. 2008.

[11] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.

[12] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*.

[13] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, pp. 1345–1359, Nov. 2010.

[14] F. Knoll, K. Hammernik, E. Kobler, T. Pock, M. P. Recht, and D. K. Sodickson, "Assessment of the generalization of learned image reconstruction and the potential for transfer learning," *Magn. Reson. Med.*, vol. 81, no. 1, pp. 116–128, Jan. 2019.

[15] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7130–7138.

[16] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2712–2721.

[17] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017. [Online]. Available: https://www.nature.com/articles/nature21056

[18] N. C. F. Codella, Q.-B. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, and J. R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM J. Res. Develop.*, vol. 61, no. 4/5, pp. 1–15, Jul. 2017.

[19] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Proc. 33rd Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–11.

[20] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," 2019, *arXiv:1901.07031*.

[21] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, and B. N. Patel, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002686. [Online]. Available: http://dx.plos.org/10.1371/journal.pmed.1002686

[22] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3462–3471.

[23] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "CheXclusion: Fairness gaps in deep chest X-ray classifiers," 2020, *arXiv:2003.00827*.

[24] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," 2017, *arXiv:1710.10501*.

[25] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Proc. Comput. Sci.*, vol. 90, pp. 200–205, Jan. 2016.

[26] X. Li, T. Pang, B. Xiong, W. Liu, P. Liang, and T. Wang, "Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification," in *Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2017, pp. 1–11.

[27] S. Masood, T. Luthra, H. Sundriyal, and M. Ahmed, "Identification of diabetic retinopathy in eye images using transfer learning," in *Proc. Int. Conf. Comput., Commun. Autom. (ICCCA)*, May 2017, pp. 1183–1187.

[28] C. Lam, D. Yi, M. Guo, and T. Lindsey, "Automated detection of diabetic retinopathy using deep learning," *AMIA Summits Transl. Sci.*, vol. 2018, no. 1, p. 147, 2018.

[29] J. Benson, H. Carrillo, J. Wigdahl, S. Nemeth, J. Maynard, G. Zamora, S. Barriga, T. Estrada, and P. Soliz, "Transfer learning for diabetic retinopathy," *Proc. SPIE*, vol. 10574, Mar. 2018, Art. no. 105741Z.

[30] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, C. M. Aparici, S. C. Behr, R. R. Flavell, S.-Y. Huang, K. A. Zalocusky, L. Nardo, Y. Seo, R. A. Hawkins, M. H. Pampaloni, D. Hadley, and B. L. Franc, "A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain," *Radiology*, vol. 290, no. 2, pp. 456–464, Feb. 2019.

[31] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of Alzheimer's disease with deep learning," in *Proc. IEEE 11th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2014, pp. 1015–1018.

[32] S. Akbarian, N. M. Ghahjaverestan, A. Yadollahi, and B. Taati, "Distinguishing obstructive versus central apneas in infrared video of sleep using deep learning: Validation study," *J. Med. Internet Res.*, vol. 22, no. 5, May 2020, Art. no. e17252.

[33] Y. Jang, H. Lee, S. J. Hwang, and J. Shin, "Learning what and where to transfer," in *Proc. ICML*, 2019, pp. 3030–3039.

[34] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018, *arXiv:1801.09927*.
[35] J. P. Cohen, L. Dao, P. Morrison, K. Roth, Y. Bengio, B. Shen, A. Abbasi, M. Hoshmand-Kochi, M. Ghassemi, H. Li, and T. Q. Duong, "Predicting COVID-19 pneumonia severity on chest X-ray with deep learning," *Cureus*, vol. 12, no. 7, 2020. [Online]. Available: https://www.cureus.com/articles/35692-predicting-covid-19-pneumonia-severity-on-chest-x-ray-with-deep-learning#!/
[36] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, *arXiv:1901.07042*.
[37] A. Bustos, A. Pertusa, J.-M. Salinas, and M. De La Iglesia-Vayá, "PadChest: A large chest X-ray image dataset with multi-label annotated reports," 2019, *arXiv:1901.07441*.
[38] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Confounding variables can degrade generalization performance of radiological deep learning models," 2018, *arXiv:1807.00431*.
[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[41] J. Zhang, C. Li, S. Kosov, M. Grzegorzek, K. Shirahama, T. Jiang, C. Sun, Z. Li, and H. Li, "LCU-Net: A novel low-cost U-Net for environmental microorganism image segmentation," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107885.
[42] J. Zhang, C. Li, Y. Yin, J. Zhang, and M. Grzegorzek, "Applications of artificial neural networks in microorganism image analysis: A comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 1013–1070, Feb. 2023.
[43] X. Li, C. Li, M. M. Rahaman, H. Sun, X. Li, J. Wu, Y. Yao, and M. Grzegorzek, "A comprehensive review of computer-aided whole-slide image analysis: From datasets to feature extraction, segmentation, classification and detection approaches," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4809–4878, Aug. 2022.
[44] H. Chen, C. Li, X. Li, M. M. Rahaman, W. Hu, Y. Li, W. Liu, C. Sun, H. Sun, X. Huang, and M. Grzegorzek, "IL-MCAM: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach," *Comput. Biol. Med.*, vol. 143, Apr. 2022, Art. no. 105265.
[45] F. Kulwa, C. Li, J. Zhang, K. Shirahama, S. Kosov, X. Zhao, T. Jiang, and M. Grzegorzek, "A new pairwise deep learning feature for environmental microorganism image analysis," *Environ. Sci. Pollut. Res.*, vol. 29, no. 34, pp. 51909–51926, Jul. 2022.
[46] A. Chen, C. Li, S. Zou, M. M. Rahaman, Y. Yao, H. Chen, H. Yang, P. Zhao, W. Hu, W. Liu, and M. Grzegorzek, "SVIA dataset: A new dataset of microscopic videos and images for computer-aided sperm analysis," *Biocybernetics Biomed. Eng.*, vol. 42, no. 1, pp. 204–214, Jan. 2022.
[47] P. F. Jaeger, S. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, H.-P. Schlemmer, and K. H. Maier-Hein, "Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection," in *Proc. Mach. Learn. Health Workshop*, 2020, pp. 171–183.
[48] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" 2020, *arXiv:2003.11539*.

S. AKBARIAN received the B.S. degree in biomedical (bio-electrical) engineering from the Amirkabir University of Technology, Tehran, Iran, in 2016, and the M.A.Sc. degree in biomedical engineering from the University of Toronto, Toronto, Canada, in 2020. He is currently a Data Scientist with Public Health Ontario and a Researcher with the Vector Institute. His research interest includes the development of automated technologies for health monitoring.

L. SEYYED-KALANTARI received the Ph.D. degree in electrical engineering from McMaster University, in 2017. She is currently an Assistant Professor with the Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University. Before that, she was an Associate Scientist with the Lunenfeld Tanenbaum Research Institute, Toronto, Canada. She was an NSERC Postdoctoral Fellow with the Vector Institute and the University of Toronto (2019–2022). Her research interests include responsible AI and developing AI diagnostic tools that focus on their fairness. She has received several highly competitive national, provincial, and institutional awards, such as Banting Postdoctoral Fellowship, in 2022 (declined), and NSERC Postdoctoral Fellowship, in 2018. Her research interest includes AI model fairness in medical imaging has been featured in many technologies news.

F. KHALVATI (Member, IEEE) is currently an Assistant Professor with the Department of Medical Imaging with cross-appointment to the Department of Mechanical and Industrial Engineering, University of Toronto, and an Associate Scientist with The Hospital for Sick Children. His research interests include artificial intelligence (transfer learning, attention networks, CNN-based survival analysis, and multi-task learning), computer vision, discovery radiomics, computational diffusion MRI, decision support systems, high-performance computing, and human-AI interface with wide applications in medical imaging and precision medicine.

E. DOLATABADI is currently a Scientist in applied machine learning with the Vector Institute. She has also been an Assistant Professor with the Institute of Health Policy, Management, and Evaluation (IHPME), University of Toronto. Her research interests include the adoption of machine learning (ML) and deep learning technologies for real-world needs. Her mission is to bring the power of ML and data science to health to improve human health and address the challenges facing our healthcare system.