Deep Supervised Domain Adaptation for Pneumonia Diagnosis From Chest X-Ray Images

Yangqin Feng, Xinxing Xu, Yan Wang, Xiaofeng Lei, Soo Kng Teo, Jordan Zheng Ting Sim, Yonghan Ting, Liangli Zhen, Joey Tianyi Zhou, Yong Liu, and Cher Heng Tan

***Abstract*—Pneumonia is one of the most common treatable causes of death, and early diagnosis allows for early intervention. Automated diagnosis of pneumonia can therefore improve outcomes. However, it is challenging to develop high-performance deep learning models due to the lack of well-annotated data for training. This paper proposes a novel method, called Deep Supervised Domain Adaptation (DSDA), to automatically diagnose pneumonia from chest X-ray images. Specifically, we propose to transfer the knowledge from a publicly available large-scale source dataset (ChestX-ray14) to a well-annotated but small-scale target dataset (the TTSH dataset). DSDA aligns the distributions of the source domain and the target domain according to the underlying semantics of the training samples. It includes two task-specific sub-networks for the source domain and the target domain, respectively. These two sub-networks share the feature extraction layers and are trained in an end-to-end manner. Unlike most existing domain adaptation approaches that perform the same tasks in the source domain and the target domain, we attempt to transfer the knowledge from a multi-label classification task in the source domain to a binary classification task in the target domain. To evaluate the effectiveness of our method, we compare it with several existing peer methods. The experimental results show that our method can achieve promising performance for automated pneumonia diagnosis.**

***Index Terms*—Domain adaptation, pneumonia diagnosis, transfer learning, chest X-ray images.**

Manuscript received December 26, 2020; revised April 15, 2021 and July 13, 2021; accepted July 17, 2021. Date of publication July 27, 2021; date of current version March 7, 2022. This work was supported by A*STAR under its Gap fund. *(Corresponding author: Xinxing Xu.)*

Yangqin Feng, Xinxing Xu, Yan Wang, Xiaofeng Lei, Soo Kng Teo, Liangli Zhen, Joey Tianyi Zhou, and Yong Liu are with the Institute of High Performance Computing, Agency for Science, Technology, and Research (A*STAR), Singapore 138632, Singapore (e-mail: fengyq@ihpc.a-star.edu.sg; xuxinx@ihpc.a-star.edu.sg; nickyan@gmail.com; leixf@ihpc.a-star.edu.sg; sookng@gmail.com; zhenll@ihpc.a-star.edu.sg; Joey_Zhou@ihpc.a-star.edu.sg; liuyong@ihpc.a-star.edu.sg).

Jordan Zheng Ting Sim and Yonghan Ting are with the Department of Diagnostic Radiology, Tan Tock Seng Hospital (TTSH), Singapore 138632, Singapore (e-mail: jordansim92@gmail.com; yonghan_ting@ttsh.com.sg).

Cher Heng Tan is with the Department of Diagnostic Radiology, Tan Tock Seng Hospital (TTSH), Singapore 308433, Singapore, and also with the Lee Kong Chian School of Medicine, Singapore 308232, Singapore (e-mail: cher_heng_tan@ttsh.com.sg).

Digital Object Identifier 10.1109/JBHI.2021.3100119

I. Introduction

PNEUMONIA is a form of an acute respiratory infection that affects the lungs and can be caused by viruses, bacteria, fungi, or other pathogens [1]. Recently, a novel coronavirus disease 2019 (COVID-19) caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has resulted in a global pandemic [2]. There have been more than 185 million confirmed cases, including more than 4 million COVID-19-related deaths globally since the outbreak has occurred [3]. Rapid diagnosis and early treatment of pneumonia can potentially reduce the mortality rate of COVID-19. Pneumonia is usually diagnosed through imaging, either through chest radiography, computed tomography (CT) or even ultrasonography (USG). Chest radiography is one of the most common and cost-effective medical imaging examinations available [4]. Traditionally, pneumonia is diagnosed through careful assessment and evaluation of chest radiographs (also known as chest X-rays or CXRs) by trained radiologists. However, it is often time-consuming, and human radiologists occasionally fall prey to fatigue and errors in interpretation. Fortunately, computer-assisted pneumonia diagnosis methods can help make the diagnosis workflow more efficient and accurate.

Recently, many advancements in deep learning approaches have demonstrated the promise of deep models for a variety of medical image analysis tasks [5]–[8]. Deep learning models have also been developed for the analysis of CXRs. For instance, Wang *et al.* released a large-scale CXRs database (ChestX-ray14), which contained 112 120 frontal-view X-ray images collected from 30 805 subjects and each image labelled with one or more of 14 disease labels [9]. They adopted some widely-used deep learning models, *e.g.*, AlexNet [10], GoogLeNet [11] and ResNet50 [12], to detect lung diseases. Yao *et al.* used LSTMs to leverage inter-dependencies among target labels for predicting 14 pathologic patterns from the ChestX-ray14 dataset and achieved promising results without pre-training [13]. Pranav *et al.* proposed a deep learning model called ChexNet [14]. It adopts a 121-layer convolutional neural network to detect lung diseases from raw images, and achieved a higher F1 score than four board-certified radiologists. Based on the location-aware Dense Networks (DNetLoc), Guendel *et al.* [15] proposed a novel approach that incorporates both high-resolution image data and spatial information for abnormality diagnosis.

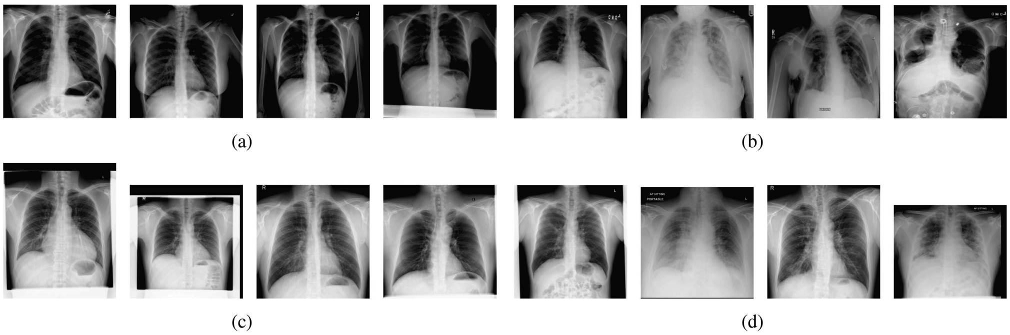


Fig. 1. Chest X-ray images from the source dataset (top row) and the target dataset (bottom row). (a) Four samples without pneumonia from the ChestX-ray14 dataset. (b) Four positive pneumonia samples from the ChestX-ray14 dataset. (c) Four samples without pneumonia from the TTSH dataset. (d) Four positive pneumonia samples from the TTSH dataset.

Despite their promising performance, these deep learning methods usually require a large-scale dataset with reliable ground-truth for training. However, large sets of well-labelled data remain a challenge in many institutions, especially for the purpose of medical data analysis. A common technique for tackling this issue is pre-training the deep learning model on a publicly available large-scale dataset, *e.g.*, ImageNet [10], and then fine-tuning the model on the target dataset to improve the accuracy. For instance, Baltruschat *et al.* used ResNet50 as the backbone network and proved that the pre-training on ImageNet outperforms the training from scratch. However, this approach does not take into account the difference between the distributions of the two domains, which may lead to the difficulty of knowledge transfer, also known as domain-shift problem [16]. The largest publicly available pneumonia diagnosis dataset, ChestX-ray14 [9], is a multi-label dataset (the source domain). Each sample from ChestX-ray14 is labelled with up to 14 different disease labels (*i. e.*, Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia and Pneumothorax). In contrast, our target dataset at hand, collected from Tan Tock Seng Hospital (TTSH), is a binary labelled dataset, in which each image is labelled with a single value to indicate if it is a positive or negative case of pneumonia. Transferring knowledge from a multi-label dataset domain to a target binary domain may result in more noise than signal. Thus, this study aims to solve the problem of transferring knowledge between heterogeneous tasks. Fig. 1 shows some images from the publicly available ChestX-ray14 dataset and the TTSH dataset in the first row and the second row, respectively. From the figure, we can see that the images from different datasets look different even though the image labels are similar. This results in domain shift for deep learning algorithms.

To conquer the two challenges above, we present a new transfer learning approach called deep supervised domain adaptation (DSDA) to detect pneumonia from CXRs automatically. Specifically, to address the domain-shift problem, we propose to align distributions of the two domains progressively. To overcome the challenge of transferring knowledge between heterogeneous tasks, we design a new neural network that contains two task-specific sub-networks. One of them is for the multi-label classification in the source domain, and another one is for the binary classification in the target domain. These two sub-networks share the feature extraction layers and are trained in an end-to-end manner. In this work, we conducted a multi-label classification task instead of a binary classification in the source domain. The reason is that the classifications of other diseases, which share the feature extraction layers, can help to improve the accuracy of pneumonia diagnosis (see the experimental results in Section IV-D). Furthermore, we align the distributions of the samples from the two datasets in the shared space in an interactive way to progressively transfer the knowledge from the source domain to the target domain. The target dataset contains 4185 chest X-ray images from TTSH; each CXR was individually and manually reviewed by a radiologist with 14 years of experience. These images are split into two classes (*i.e.*, pneumonia or non-pneumonia). Even if there are other abnormalities in the CXR (e.g. mass, cardiomegaly etc.) they will still be labelled as non-pneumonia. Furthermore, we take the publicly available ChestX-ray14 dataset [9] as the source domain. Pneumonia is one of the 14 lung diseases labelled in the ChestX-ray14 dataset and can be difficult to be detected in CXR, especially in mild infections [17]. To the best of our knowledge, this work is the first one that adopts domain adaptation between heterogeneous tasks for pneumonia diagnosis.

The novelty and contributions of this work are summarised as follows:

- We propose a domain adaptation method to transfer knowledge from a multi-label classification task in the source domain to a binary classification task in the target domain.
- A novel domain alignment strategy is proposed to align the source domain and the target domain according to the underlying semantics of the samples progressively. It explicitly minimises the inter-class similarity and maximises the intra-class similarity across different domains.
- A new CXR dataset (the TTSH dataset) is collected and well-annotated to verify the effectiveness of our proposed pneumonia diagnosis method. Extensive experiments are conducted and show that our proposed domain alignment and the multi-task learning strategy can improve the pneumonia diagnosis performance significantly.

II. Related Work

Deep learning for chest X-ray analysis: Inspired by the recent success of deep learning algorithms in many real-world applications, researchers in digital healthcare have made great efforts to develop automated disease diagnosis methods using chest X-ray images [9], [18]–[20]. For example, to help alert clinicians and radiologists of potential abnormal findings in the lungs and to help triage and prioritise reporting of certain radiographs, Tang *et al.* employed various deep learning models to detect the abnormal cases from CXRs [18]. They investigated the performance of different deep convolutional neural networks (such as AlexNet [10], VGGNet [21], ResNet [12], Inception-v3 [22], and DenseNet [23]) on the ChestX-ray14 dataset [9]. To evaluate the generalisation ability of deep learning models on data from different sources, Pan *et al.* used DenseNet and MobileNetV2 [24] to classify CXRs into normal or abnormal categories on the ChestX-ray14 dataset and the Rhode Island Hospital chest radiograph dataset. To overcome the problem of difficulty in learning from a small-scale dataset, Stephen *et al.* proposed a convolutional neural network model trained from scratch to classify and detect the presence of pneumonia [25]. They adopted several data augmentation algorithms to improve the performance of the deep learning model. Liang *et al.* proposed a transfer learning framework that combines residual structure and dilated convolution to detect childhood pneumonia [26]. A large-scale dataset is exploited for improving performance. In recent years, several studies [27]–[29] have been conducted to investigate the modality-specific transfer learning for CXRs, which consider exploiting the source dataset that shares the same modality with the target dataset. For instance, Rajaraman *et al.* proposed to construct ensembles based on the modality-specific knowledge transfer to perform the detection of COVID-19, which has achieved promising results. However, these methods do not sufficiently bridge the domain gap between the target dataset and the source dataset in an iterative manner.

Domain adaptation: Deep learning methods usually require massive amounts of labelled data for training to achieve a high prediction accuracy [30]. While structured and well-annotated data are difficult to collect in many real-world applications, transferring the learnt knowledge from a label-rich source domain to a label-poor or an unlabelled target domain can address the lack of well-annotated data [31]–[33]. However, the source domain and the target domain are often drawn from different distributions, which may result in a domain-shift problem [16], [34], [35]. In such a challenging case, many domain adaptation approaches have been developed to address this issue by aligning the two domains [36]–[41] in the feature space. For example, Kang *et al.* proposed a contrastive adaptation network (CAN), which models the intra-class and inter-class domain discrepancy explicitly [42]. You *et al.* proposed a universal adaptation network (UAN) to quantify sample-level transferability and discover the common label set for learning, thus encouraging the adaptation automatically [43]. Long *et al.* proposed a framework named conditional domain adversarial networks [39]. The model contains two novel conditioning strategies: multi-linear conditioning and entropy conditioning. Multi-linear conditioning helps capture the cross-covariance between feature representations and classifier predictions, and entropy conditioning helps control the uncertainty of classifier predictions. Zhang *et al.* proposed a novel measurement, margin disparity discrepancy (MDD), to measure the distribution difference and achieve the domain adaptation in an adversarial learning manner [44].

Note that these approaches all consider the scenarios where the two domains have homogeneous tasks. In this work, we consider the case where the source dataset has been labelled for multi-label classification while the target dataset is labelled for binary classification. The domain adaptation methods mentioned above cannot be applied to solve this task directly. Unlike these existing domain adaptation approaches, ours considers heterogeneous tasks, and we align the distributions of the two domains according to the underlying semantics of the training samples progressively.

III. Our Proposed Method

A. Problem Formulation

In this work, we aim to improve the performance of a model on the target domain by transferring knowledge from a source domain [45]. Assuming that we have a source data matrix of n_s samples $\mathbf{X}^s = [\mathbf{x}_1^s, \mathbf{x}_2^s, \ldots, \mathbf{x}_{n_s}^s]$. Each data sample $\mathbf{x}_i^s$ has a semantic label vector $\mathbf{y}_i^s = [y_{1i}^s, y_{2i}^s, \ldots, y_{ci}^s]^t \in \{0,1\}^{c_s}$, where c_s is the number of categories. If $\mathbf{x}_i^s$ belongs to the jth category, then $y_{ji}^s = 1$, otherwise $y_{ji}^s = 0$. Note that $\mathbf{y}_i^s$ may contain several non-zero elements as it can belong to multiple categories. We denote the data labels of the source domain in a matrix form as $\mathbf{Y}^s = [\mathbf{y}_1^s, \mathbf{y}_2^s, \ldots, \mathbf{y}_{n_s}^s]$.

In the target domain, we have the data matrix of n_t samples $\mathbf{X}^t = [\mathbf{x}_1^t, \mathbf{x}_2^t, \ldots, \mathbf{x}_{n_t}^t]$. Its associated label matrix is $\mathbf{Y}^t = [\mathbf{y}_1^t, \mathbf{y}_2^t, \ldots, \mathbf{y}_{n_t}^t]$, where $\mathbf{y}_i^t \in \{0,1\}^{c_t}$ and c_t is the number of categories of the target dataset. In this work, we consider the binary classification in the target domain, *i.e.*, each sample $\mathbf{x}_i^t$ is annotated into a pneumonia or non-pneumonia category as $\mathbf{y}_i^t = [0,1]^T$ or $\mathbf{y}_i^t = [1,0]^T$. Let us denote the probability distributions of the source and target domains as $D^s(\mathbf{x}^s, \mathbf{y}^s)$ and $D^t(\mathbf{x}^t, \mathbf{y}^t)$, respectively, and $D^s \neq D^t$. We trained a deep neural network on the union set of the source dataset and target dataset and attempted to align these two distributions (*i.e.*, D^s and D^t) in an interactive way. This work considers the scenario where the samples from the source dataset and the target dataset are all CXRs, i.e., the two domains have the same modality but with different distributions as they cover different populations.

B. Framework of DSDA

Our method aims to accurately diagnose pneumonia in the target dataset by leveraging the knowledge from a publicly available multi-label dataset as the source domain. The task for the source dataset is a multi-label classification task, while the task for the target dataset is a binary classification task, *i.e.*, these two tasks are heterogeneous. For such a case, we design two sub-networks for the multi-label classification and the binary

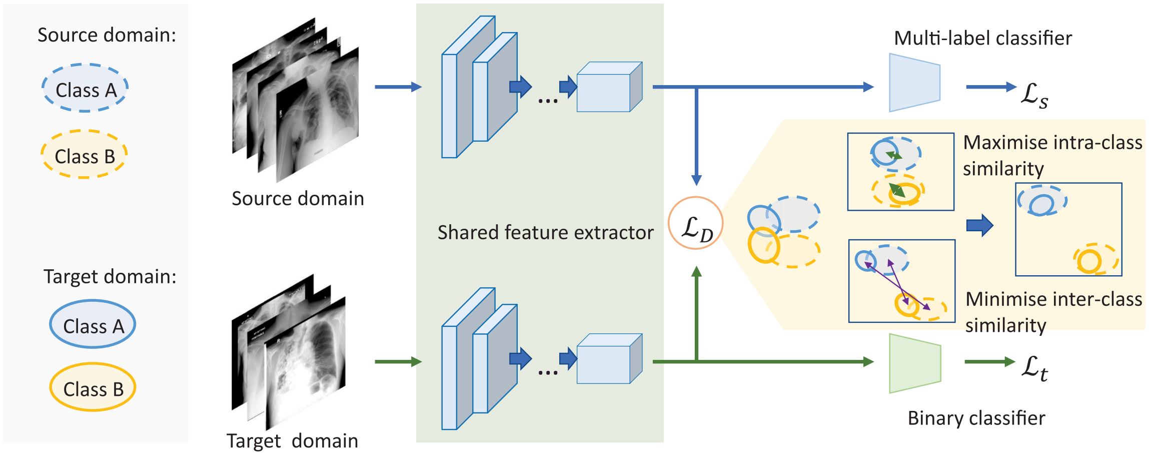


Fig. 2. The architecture of the proposed method. DSDA takes samples from both the source domain and the target domain as inputs at the same time. It contains two task-specific sub-networks: 1) a multi-label classifier with a linear activation function for the source domain and 2) a binary classifier with a softmax activation function for the target domain. The proposed architecture shares the convolutional feature extraction layers. Furthermore, we explicitly impose a new distribution alignment constraint on the samples from the source domain and the target domain in the shared space.

classification, respectively. They share the feature extraction layers, which can be from off the shelf existing convolutional neural networks (CNNs), such as ResNets [12] and DenseNets [23].

The architecture of the proposed domain adaptation model is shown in Fig. 2, from which we can see that our DSDA has two task-specific sub-networks. One is to discriminate 14 diseases (including pneumonia) for the source domain, and another one is to distinguish pneumonia and non-pneumonia images for the target domain. The shared feature extraction layers enforce DSDA to map the input samples into a shared common space and learn the shared features for the source domain and the target domain. More importantly, we explicitly align the distributions of the feature vectors of the source domain and that of the target domain according to their underlying semantics. Specifically, we enforce the intra-class samples to be close, while the inter-class samples to be far away, no matter which domain they are come from. Thus, unlike the pre-training strategy, the proposed model can leverage the knowledge from both the source domain and the target domain to improve the model's performance by interactively learning discriminative features. Thus, it leads to better performance than the widely-used pre-training strategy.

C. Objective Function

To achieve accurate pneumonia diagnosis, our method proposes to minimise the following objective function:

$$\mathcal{L}(\Theta|\mathbf{X}^s, \mathbf{X}^t, \mathbf{Y}^s, \mathbf{Y}^t) = \mathcal{L}_t + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_D, \qquad (1)$$

where Θ denotes the parameters of our model, that is, the weights and biases, $\mathcal{L}_s$ and $\mathcal{L}_t$ are the classification losses of the source domain and the target domain, respectively. $\mathcal{L}_D$ is the domain adaptation loss, and λ_1 and λ_2 denote the hyper-parameters that trade-off the contributions of the three different terms. The details of $\mathcal{L}_s$, $\mathcal{L}_t$ and $\mathcal{L}_D$ will be illustrated in the following.

1) Classification Loss for the Target Domain: Since the dataset of the target domain is class-imbalanced, we optimise the weighted cross entropy loss (WCEL) for the classifier. For a single sample in the training dataset, WCEL can be calculated as:

$$\mathcal{L}_t(x_i^t, y_i^t) = -\left(y_{1i}^t \log(\hat{y}_{1i}^t) + u y_{2i}^t \log(\hat{y}_{2i}^t)\right), \qquad (2)$$

where u is a manual rescaling weight given to the positive class, and we set u as the ratio of the number of negative training samples to the number of positive training samples in this work. $\hat{\mathbf{y}}_i^t$ denotes the network output for the ith input image $\mathbf{x}_i^t$ from the target domain.

2) Classification Loss for the Source Domain: For the multi-label classifier of source domain, we optimise the BCEWithLogitsLoss, which combines a sigmoid layer and the BCELoss in each class. For a single sample in the training dataset, the BCEWithLogitsLoss can be described as:

$$\mathcal{L}_s(x_i^s, y_i^s) = \frac{1}{c_t}\sum_{j=1}^{c_t} -w_j(y_{ji}^s \log(\sigma(\hat{y}_{ji}^s)) - (1 - y_{ji}^s)\log(1 - \sigma(\hat{y}_{ji}^s))), \qquad (3)$$

where w_j is the weight for the jth class, and we set it as the ratio of the number of negative samples to the number of positive samples for the jth class, $\sigma(\cdot)$ is the sigmoid function and $\hat{y}_i^s$ denotes the network output for the input image $\mathbf{x}_i^s$ from the source domain.

3) Domain Adaptation Loss: For the given source data matrix $\mathbf{X}^s$ and the target data matrix $\mathbf{X}^t$, we use the shared feature extractor to obtain their high-level representations in the shared common space, denoted as $\mathbf{Z}^s = [\mathbf{z}_1^s, \mathbf{z}_2^s, \ldots, \mathbf{z}_{n_s}^s]$ and $\mathbf{Z}^t = [\mathbf{z}_1^t, \mathbf{z}_2^t, \ldots, \mathbf{z}_{n_t}^t]$. The proposed method attempts to align the distributions of $\mathbf{Z}^s$ and $\mathbf{Z}^t$ by minimising the inter-class similarity and maximising the intra-class similarity across the two domains.

The main idea of the proposed domain adaptation strategy is shown in Fig. 2. In the training process, we computed the similarity between the samples across domains. If the samples belong to the same class, we maximise their similarity, and if they are in different classes, we minimise their similarity. In this manner, we can align the distributions of different domains, resulting in different classes being far away. Specifically, to measure the similarity of the input images $\mathbf{x}_i^s$ and $\mathbf{x}_j^t$, we firstly compute their high-level representations $\mathbf{z}_i^s$ and $\mathbf{z}_j^t$ in the shared space. Then we use the Gaussian kernel metrics to calculate the similarity:

$$k(\mathbf{z}_i^s, \mathbf{z}_j^t) = e^{-\frac{\|\mathbf{z}_i^s - \mathbf{z}_j^t\|_2^2}{b}} \text{ and } b = \frac{1}{n_s n_t} \sum_i^{n_s} \sum_j^{n_t} \|\mathbf{z}_i^s - \mathbf{z}_j^t\|_2^2.$$

Next, we perform distribution alignment by minimising the mean inter-class similarity and maximising the mean intra-class similarity across two domains using the following loss function:

$$\mathcal{L}_D = \frac{1}{n_s n_t} \sum_i^{n_s} \sum_j^{n_t} \delta(x_i^s, x_j^t) k(\mathbf{z}_i^s, \mathbf{z}_j^t), \tag{4}$$

where $\delta(x_i^s, x_j^t)$ is an indicator function to show whether x_i^s and x_j^t belong to the same class. If x_i^s and x_j^t belong to the same class (*i.e.*, both have pneumonia or non-pneumonia), then $\delta(x_i^s, x_j^t) = -1$, otherwise $\delta(x_i^s, x_j^t) = 1$.

The advantage of our method is that our method can transfer knowledge during the interactions between the source and target domains progressively. Moreover, our proposed model can be optimised in an end-to-end manner via the commonly used back-propagation algorithms, such as the stochastic gradient descent method.

D. Implementation Details

In this work, we employ the convolutional layers of the ResNets [12] or DenseNets [23] as the backbone of our model. On the top of the CNN backbone, we add a fully connected layer of 14 neurons with the linear activation function for the source domain sub-network and add a fully connected layer of 2 neurons with the Softmax activation function for the target domain sub-network. The weights of the feature extractor in our model are initialised with the weights from a model pre-trained on ImageNet [46]. The network is trained end-to-end using Adam [47] with default parameters (beta1=0.9, beta2=0.999). The batch size of our model is 32. We use a learning rate of 0.0001, $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$. In addition, we trained our model for 20 epochs and selected the model with the highest AUC score on the validation set for testing. Our model is trained on two NVIDIA Titan XP GPUs in PyTorch.

IV. Experimental Study

We conducted the experiments using the ChestX-ray14 dataset (as the source dataset) [9] and our own TTSH dataset (as the target dataset). We first compare the DSDA method with state-of-the-art methods, including traditional machine learning methods, the deep learning methods for CXRs, and some domain adaptation methods, to evaluate its performance. We then evaluate the importance of learning from the multi-labelled source domain, the influence of different transfer strategies, and the impact of three different terms of our objective function in Equation (1).

TABLE I
Characteristics of the TTSH Dataset, Where n_{train}, $n_{validation}$ and n_{test} Are the Numbers of Images in the Training, Validation, and Test Sets, Respectively

Category	n_{train}	$n_{validation}$	n_{test}
Non-pneumonia	2,115	302	604
Pneumonia	825	117	222
Global	2,940	419	826

TABLE II
Characteristics of the ChestX-Ray14 Dataset

Disease	n_{train}	$n_{validation}$	n_{test}
Atelectasis	7,996	1,119	2,420
Cardiomegaly	1,950	240	582
Effusion	9,261	1,292	2,754
Infiltration	13,914	2,018	3,938
Mass	3,988	625	1,133
Nodule	4,375	613	1,335
Pneumonia	978	133	242
Pneumothorax	3,705	504	1,089
Consolidation	3,263	447	957
Edema	1,690	200	413
Emphysema	1,799	208	509
Fibrosis	1,158	166	362
Pleural Thickening	2,279	372	734
Hernia	144	41	42

A. Datasets

1) The TTSH dataset: we collected this dataset from Tan Tock Seng Hospital (TTSH) in Singapore, which contains 4185 CXRs. Each CXR was individually and manually reviewed by a radiologist with 14 years of experience. The images in the TTSH dataset are split into two classes (pneumonia or non-pneumonia). The dataset comprises 3021 non-pneumonia and 1164 pneumonia cases. We divided the dataset into three subsets: training set, validation set and test set. The characteristics of the target dataset are summarised in Table I.

2) ChestX-ray14 [9]: ChestX-ray14 is compiled by the National Institutes of Health (NIH) and is currently one of the largest public repository of CXRs. It includes 112 120 front-view X-ray images of 30 805 patients. Each image in the ChestX-ray14 dataset is labelled with up to 14 different labels that are chosen based on the frequency of observation and diagnosis in clinical practice [9]. The labels for each CXR are obtained by an automatic extraction method on radiology reports with an estimated accuracy of 90%; multiple diseases can be presented in one image. The nature of the ChestX-ray14 dataset is oriented to the formulation of a multi-label classification problem. ChestX-ray14 is also divided into three subsets, and its characteristics are summarised in Table II.

3) The RSNA dataset: The Radiological Society of North America (RSNA) Pneumonia Detection Challenge[1] dataset (Stage I) is a collection of 26 684 samples taken from ChestX-ray14 dataset. It contains 20 672 non-pneumonia and 6012

[1] [Online]. Available: https://www.kaggle.com/c/ rsna-pneumonia-detection-challenge

pneumonia[2]. The nature of the ChestX-ray14 dataset is oriented to the formulation of a multi-label classification problem, whereas the RSNA dataset formulates a multi-class classification problem. Although the RSNA dataset is a subset of the ChestX-ray14 dataset, each sample in the RSNA dataset will be categorised with only one class. Therefore, it is an improved dataset focused on pneumonia cases, providing accurate and useful information to be used in classification and detection tasks.

B. Comparison With Other Methods

We verify the effectiveness of DSDA by comparing it with three different types of methods, including six traditional methods, namely support-vector machine (SVM) [50]), linear discriminant analysis (LDA) [51], adaBoost classifier [52], decision tree (DT) [53], random forest (RF) [54], and logistic regression (LR) [55], and five deep learning methods for CXRs, ChexNet [14] and the methods proposed by Baltruschat *et al.* [56], Tang *et al.* [18], Varma *et al.* [19], and Pan *et al.* [20]. We also compare our method with four different domain adaptation methods, namely deep adaptation neural network (DANN) [35], beyond sharing weights (BSW) [57], margin disparity discrepancy (MDD) [44], and conditional domain adversarial network (CDAN) [39]. We implement each compared domain adaption method in both supervised (S) and unsupervised (U) learning settings. In addition, for all the traditional methods, we extract two different types of features, respectively. The first type is the local descriptors over the oriented fast and rotated brief (ORB) [58] feature using the bag-of-words (BoW) method [59]. It has been proven to be well-adapted to recognition and matching tasks, as they are robust to partial visibility. The second type is the CNN-based feature, and we used the feature extractor of DenseNet121 [23], which was pre-trained on the ChestX-ray14 dataset and fine-tuned on the TTSH dataset to extract the feature representations of the TTSH dataset. Lastly, ChexNet is pre-trained on ImageNet as in the original paper; and the method in [56] is pre-trained on the ChestX-ray14 dataset. Original MDD, CDAN, and DAN are unsupervised domain adaptation methods, and we implemented a variant of these methods to exploit the label information on the target dataset by adding a target domain classifier that minimises Equation (2). All the compared domain adaptation methods require the classifiers for the source dataset and the target dataset to have the same architecture; we constructed a binary (pneumonia cases and non-pneumonia) classification dataset based on the labels and the data from ChestX-ray14 for them in our experiments.

Table III reports the $Precision$, $F1$, AUC, $Sensitivity$ and $Specificity$ scores of the proposed method and other peer methods, from which we have the following observations:

- All the traditional methods that used the CNN-based features (DN) outperform its corresponding method with the ORB features (ORB) in terms of AUC score. It indicates that the CNN-based feature is more suitable for our task.

TABLE III

Performance Comparison of the Proposed Method With the Peer Methods in Terms of $Precision$ (PRC) Scores, $F1$ Scores, AUC Scores, $Sensitivity$ (SEN) and $Specificity$ (SPE). (ORB) Denotes the ORB Feature and (DN) Denotes the DenseNet Feature

Method	PRC	$F1$	AUC	SEN	SPE
DT(ORB)	0.7642	0.6270	0.5428	0.5315	0.5541
DT(DN)	0.6522	0.6846	0.8026	0.7205	0.8847
LDA(ORB)	0.7623	0.8544	0.7416	0.7268	0.6622
LDA(DN)	0.8394	0.7678	0.8334	0.7274	0.9094
AdaBoost(ORB)	0.8304	0.8122	0.7349	0.7666	0.5991
AdaBoost(DN)	0.7538	0.8404	0.7555	0.9253	0.6712
SVM(ORB	0.7312	0.8448	0.7406	0.7964	0.5991
SVM(DN)	0.8157	0.7937	0.8603	0.7730	0.9275
RF(ORB)	0.8499	0.6903	0.6509	0.5811	0.7207
RF(DN)	0.7802	0.7852	0.8618	0.7904	0.9332
LR(ORB)	0.7659	0.8515	0.7447	0.8063	0.5856
LR(DN)	0.8027	0.7920	0.8620	0.8017	0.9223
ChexNet [14]	0.7937	0.8439	0.9481	0.9009	0.9139
Baltruschat [56]	0.8065	0.8511	0.9488	0.9009	0.9205
Tang [18]	0.7519	0.8197	0.9444	0.9009	0.8907
Varma [19]	0.7725	0.8260	0.9276	0.8874	0.9040
Pan [20]	0.7069	0.8008	0.9544	0.9234	0.8593
DANN (U) [35]	0.4851	0.6058	0.8339	0.8063	0.6854
DANN (S) [35]	0.7088	0.7968	0.9543	0.9099	0.8626
BSW (U) [57]	0.4208	0.5338	0.7433	0.7297	0.6308
BSW (S) [57]	0.6815	0.7743	0.9496	0.8964	0.8460
MDD (U) [44]	0.6859	0.7615	0.9096	0.8559	0.8560
MDD (S) [44]	0.7672	0.8306	0.9485	0.9054	0.8990
CDAN (U) [39]	0.5430	0.6547	0.8652	0.8243	0.7450
CDAN (S) [39]	0.6769	0.7713	0.9470	0.8964	0.8427
DSDA (ours)	**0.8697**	**0.9000**	**0.9782**	**0.9324**	**0.9487**

- Deep learning methods (deep X-ray methods and deep domain adaptation methods) outperform the traditional methods significantly. Also, the deep learning methods obtain higher scores than the traditional ones in terms of $Sensitivity$ and $Specificity$. It indicates that the X-ray images may need deep neural networks to extract the complex nonlinear structure of the dataset for pneumonia diagnosis.
- Deep X-ray methods are specifically designed to handle X-ray images and work well for the Pneumonia classification from CXRs. They can achieve comparable AUC scores with the domain adaptation methods, especially the method proposed by Pan *et al.* obtains the second-highest AUC score.
- Domain adaptation methods perform well, especially the supervised variants. Even without the label information on the target dataset, the domain adaptation methods of DANN (U), MDD (U), and CDAN (U) can achieve more than 83.3% in terms of the AUC score, and MDD (U) obtains 90.96%. These results show the importance of domain adaptation. When using the label information on the target dataset, the domain adaptation methods achieve an improvement of 12.04% for DANN, 20.63% for BSW, 3.89% for MDD, and 8.18% for CDAN. Note that there are some noisy labels for the ChestX-ray14 dataset (source domain), which has a negative impact for the performance of the domain adaptation methods.
- Our method outperformed all other methods by a large margin. There are two potential reasons why the proposed method outperformed the adversarial-based method. Firstly, although adversarial-based methods have achieved state-of-the-art performance on several natural

[2] RSNA unifies the categories of pneumonia and diseases with similar pathologies such as consolidation and infiltration [48], [49]. It contains three classes: normal, lung opacity and not-normal. We treated the Lung opacity class as pneumonia, Normal and Not-Normal as non-Pneumonia.

TABLE IV
COMPARISON OF THE PROPOSED METHOD UNDER SEVEN DIFFERENT BACKBONES WITH FOUR DIFFERENT PRE-TRAINING STRATEGIES IN TERMS OF THE AUC SCORE

Method	DenseNet121	DenseNet201	ResNet18	ResNet34	ResNet50	ResNet101	ResNet152
P-C	0.9097	0.9061	0.9062	0.9289	0.8849	0.8717	0.8625
P-I-FC	0.9073	0.8972	0.9008	0.8899	0.8967	0.9147	0.9083
P-I-FA	0.9481	0.9482	0.9378	0.9363	0.9266	0.9256	0.9159
P-C-FC	0.9601	0.9430	0.9507	0.9494	0.9344	0.9289	0.9110
P-C-FA	0.9702	0.9583	0.9575	0.9505	0.9488	0.9470	0.9373
DSDA	**0.9782**	**0.9626**	**0.9666**	**0.9599**	**0.9574**	**0.9601**	**0.9396**

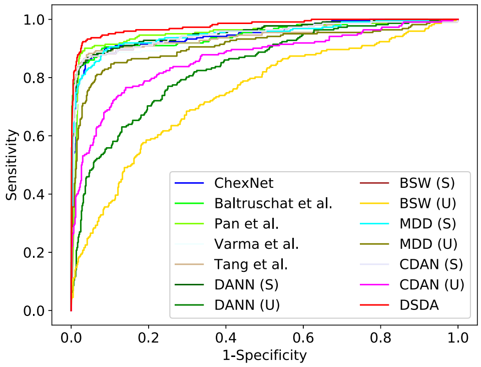


Fig. 3. The ROC curves of our DSDA method and the peer methods.

image classification tasks, they are hard to train, especially on the small-scale medical image dataset. Secondly, our method transfers knowledge between heterogeneous tasks. It conducts a multi-label classification task instead of only considering the pneumonia binary classification in the source domain, which boosts the performance on the binary classification of the target domain significantly.

Fig. 3 shows the ROC curves of our DSDA method, deep X-ray methods and the other domain adaptation methods. From the results, we can find that our DSDA method outperforms other methods. Besides, the examined deep models (deep X-ray methods and deep domain adaptation methods) obtain much better performance than all the examined traditional machine learning methods. These observations are consistent with the findings in Table III.

C. Evaluation of Different Backbones and Training Strategies

To investigate the effectiveness of different backbones to the proposed method, we compare the proposed method under seven different backbones with different pre-training strategies. The backbones used in the experiments are Densenet121, Densenet201 [23], ResNet18, ResNet34, ResNet50, ResNet101 and ResNet152 [12]. We also evaluated four different pre-training strategies: pre-training on the ChestX-ray14 dataset without fine-tuning (P-C); pre-training on ImageNet with fine-tuning all layer (P-I-FA); pre-training on the ImageNet with fine-tuning the fully connected layers and fixed convolution layers (P-I-FC); pre-training on ChestX-ray14 and fine-tuning the fully-connected layers with fixed convolutional layers (P-C-FC); and pre-training on ChestX-ray14 and fine-tuning all layers without freezing any layers (P-C-FA). For pre-training, we adopted $\mathcal{L}_s$ in Equation (3) as the objective function; for fine-tuning, we used $\mathcal{L}_t$ in Equation (2) as the objective function by following our method.

Table IV reports the AUC score of the proposed method with different backbones and training strategies, from which we can see that:

- The models with fine-tuning (P-I-FA, P-C-FC and P-C-FA) significantly outperformed the models without fine-tuning for all the different backbones, which shows the importance of the classification loss in the label space of the target domain.
- The models pre-trained on ChestX-ray14 (P-C-FA) outperformed the models pre-trained on ImageNet (P-I-FA). It demonstrates the importance of the source dataset.
- Our method outperformed the others under different backbones, which indicates that our domain adaption method can better utilize knowledge from the source domain to the target domain and improve the performance on the pneumonia diagnosis.

D. Impact of Different Classification Strategies in the Source Domain

To investigate the impact of different classification strategies in the source domain, we compared our DSDA model under seven different backbones with other two different source datasets. The first one is the RSNA dataset, which is well-annotated for the pneumonia detection task. The second one treats the images in the ChestX-ray14 dataset as a binary labelled dataset (pneumonia and non-pneumonia) (BCX14). The source domain objective function ($\mathcal{L}_s$) for this experiment was just set as WCEL.

Table V reports the AUC scores of the proposed method with different backbones and two other source datasets of 5-fold cross-validation, from which we can see that:

- Our proposed heterogeneous tasks domain adaptation strategy, *i.e.*, conducting multi-label classification in the source domain, significantly outperformed the strategy of adopting binary classification in the source domain. It denotes that the classifications of the other diseases can help to improve the accuracy of pneumonia diagnosis, as we hypothesized previously. This may be because multi-label learning task may enforce the model to obtain more information from shared features;
- Our method (trained on the RSNA dataset as the source domain) outperformed the method trained on the BCX14 dataset as the source dataset. This may be because the annotations of the samples are more accurate and more positive pneumonia cases are contained in the RSNA

TABLE V

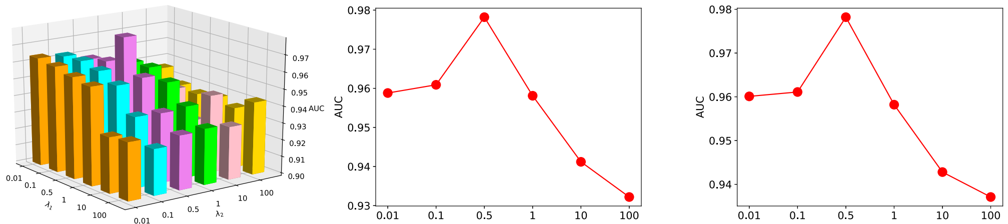
COMPARISON OF THE PROPOSED METHOD UNDER SEVEN DIFFERENT BACKBONES WITH OTHER TWO DIFFERENT SOURCE DATASETS IN TERMS OF THE AVERAGE AUC SCORE AND STANDARD DEVIATION. THE SYMBOL OF "†" INDICATES THAT THE VALUE OF THE PROPOSED METHOD IS SIGNIFICANTLY DIFFERENT FROM ALL OTHER METHODS AT A 0.05 LEVEL BY THE T-TEST

Method	DenseNet121	DenseNet201	ResNet18	ResNet34	ResNet50	ResNet101	ResNet152
RSNA	0.9505±0.0119	0.9469±0.0088	0.9496±0.0087	0.9465±0.0072	0.9383±0.0067	0.9375±0.0067	0.9341±0.0056
BCX14	0.9432±0.0121	0.9257±0.0293	0.9440±0.0102	0.9316±0.0252	0.9261±0.0151	0.9212±0.0167	0.9247±0.0173
DSDA	**0.9672±0.0096†**	**0.9595±0.0080†**	**0.9610±0.0068†**	**0.9611±0.0109†**	**0.9531±0.0095†**	**0.9566±0.0113†**	**0.9470±0.0102†**

TABLE VI

PERFORMANCE COMPARISON OF OUR DSDA METHOD WITH ITS THREE VARIANTS UNDER DIFFERENT BACKBONES IN TERMS OF AUC SCORE

Method	DenseNet121	DenseNet201	ResNet18	ResNet34	ResNet50	ResNet101	ResNet152
w/o $\mathcal{L}_t$	0.8771	0.8693	0.8981	0.8730	0.9150	0.8961	0.8735
w/o $\mathcal{L}_D$	0.9512	0.9579	0.9637	0.9520	0.9522	**0.9615**	0.9364
w/o $\mathcal{L}_s$	0.9627	0.9609	0.9658	0.9594	0.9552	0.9579	0.9350
DSDA	**0.9782**	**0.9626**	**0.9666**	**0.9599**	**0.9574**	0.9601	**0.9396**



(a) (b) (c)

Fig. 4. Performance (20 epoch) of the proposed method with the DenseNet121 backbone. (a) The AUC score of DSDA versus different value of λ_1 and λ_2. (b) The AUC score of DSDA versus different values of λ_1 under $\lambda_2 = 0.5$. (c) The AUC score of DSDA versus different values of λ_2 under $\lambda_1 = 0.5$.

dataset than in the BCX14 dataset. Moreover, the performance of our model can further improve if some of the wrong annotations are corrected or more positive cases are added to the BCX14 dataset.

E. Ablation Study

The objective function of DSDA consists of three terms, including the domain adaptation loss of the feature space, and the classification losses in the label space of the source domain and the target domain, respectively. To investigate the impact of the three different terms, we evaluate three variations of the objective function $\mathcal{L}$: the objective function without $\mathcal{L}_t$ (w/o $\mathcal{L}_t$), the objective function without $\mathcal{L}_D$ (w/o $\mathcal{L}_D$) and the objective function without $\mathcal{L}_s$ (w/o $\mathcal{L}_s$). In addition, without $\mathcal{L}_D$ (w/o $\mathcal{L}_D$), the model is a simple multi-task learning model.

Table VI reports the AUC scores of the proposed method with its three variants, from which we find that:

- The full objective function (DSDA) performed the best on the TTSH dataset. It indicates that all of the three different terms ($\mathcal{L}_s$, $\mathcal{L}_t$, and $\mathcal{L}_D$) in the objective function ($\mathcal{L}$) contribute to the final AUC score.
- The full objective function (DSDA) outperforms the model w/o $\mathcal{L}_t$ with a large margin. It demonstrates the importance of the $\mathcal{L}_t$ term (the classification loss in the label space of the target domain).
- The full objective function (DSDA) is superior to the model w/o $\mathcal{L}_s$ and the model w/o $\mathcal{L}_D$. It demonstrates the significance of the $\mathcal{L}_s$ term and the $\mathcal{L}_D$ term of our DSDA. This illustrates that formulating both the classification loss of the source domain and the domain adaptation loss in the objective function is a meaningful strategy for transfer learning.

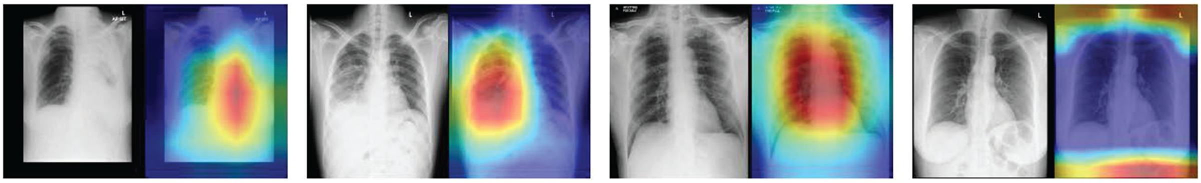
F. Parameter Analysis

DSDA has two hyper-parameters, λ_1 and λ_2. To investigate the sensitivity of λ_1 and λ_2, we conducted the experiment with the DenseNet121 backbone. We set λ_1 and λ_2 both from 0.01 to 100, the results are shown in Fig. 4, from which we have that: 1) DSDAs with the backbone of DenseNet121 obtained the highest AUC score with $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$; and 2) DSDAs with the backbone of DenseNet121 can achieve higher than the AUC score of 95% with λ_1 and λ_1 both from 0.01 to 1.

G. Visualisation of Heatmaps

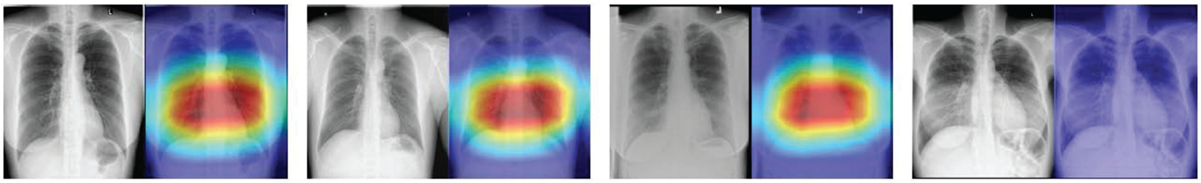
To visually investigate the effectiveness of our DSDA model, we adopted the Gradient-weighted Class Activation Mapping (Grad-CAM) approach [60] to visualise the regions of input that are "important" for predictions from our model in the X-ray images. The Grad-CAM can produce a coarse localisation map highlighting the critical areas of the image for predicting the concept.

Fig. 5 and Fig. 6 show several examples of Grad-CAM visualisation of pneumonia and non-pneumonia X-ray images, respectively. There are two images in each sub-figure of Fig. 5



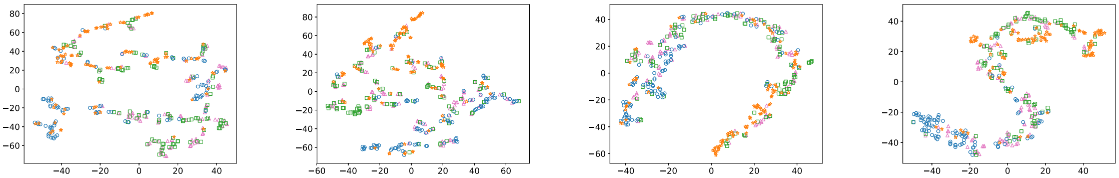
(a) True positive (b) True positive (c) True positive (d) False negative

Fig. 5. Four examples of Grad-CAM visualisation of pneumonia x-ray images. (a), (b) and (c) show three positive cases and are predicted as pneumonia. (d) is a positive case but is predicted as non-pneumonia. The areas in the X-ray that are most important for making the predictions are highlighted.



(a) True negative (b) True negative (c) True negative (d) False positive

Fig. 6. Four examples of Grad-CAM visualisation of non-pneumonia X-ray images. (a), (b) and (c) show three negative cases that are predicted as non-pneumonia. (d) shows a negative case that is predicted as pneumonia.



(a) ResNet152 (b) DenseNet121 (c) DSDA (ResNet152) (d) DSDA (DenseNet121)

Fig. 7. Visualisation of representations using the t-SNE method. In all the sub-figures, (Orange star: pneumonia cases in the target domain. Blue circle: non-pneumonia cases in the target domain. Green square: pneumonia cases in the source domain. Pink triangle: non-pneumonia cases in the source domain.) (a) The feature representations obtained by ResNet152. (b) The feature representations obtained by DenseNet121. (c) The feature representations obtained by training our method with the ResNet152 backbone. (d) The feature representations obtained by our method with the DenseNet121 backbone.

and Fig. 6, the left one is the original X-ray image, and the right one is the heatmap generated by Grad-CAM. From Fig. 5 and Fig. 6, we have that:

- In most cases, the proposed model correctly identifies manifestations of pneumonia in X-ray images as seen in (a), (b) and (c) from Fig. 5. In Fig. 5(a), the model highlights the pneumonia which presents radiopaque region in the right lung area; in Fig. 5(b) the model highlights the pneumonia region in the left lung; in Fig. 5(c) the model highlights pneumonia in the both sides. It indicates that our model can predict pneumonia by focusing on the lesion region presented in the lung area. On the contrary, Fig. 5(d) is a pneumonia case but is misclassified as a non-pneumonia case since the model focuses on the incorrect region;
- Almost all of the true negative samples are highlighted in the similar regions of the input images as shown in (a), (b) and (c) from Fig. 6. Fig. 6(d) is a non-pneumonia case but is misclassified as a pneumonia case as the model cannot focus on the lung area. They demonstrate that our DSDA method has a reasonable interpretation.

H. Visualisation of Representations From Different Domains

We embed the representations of the samples which come from different classes and different domains (in the shared representation space) into a two-dimensional visualisation plane by using the t-SNE method [61]. To visually investigate the effectiveness of DSDA, we randomly chose 100 samples in each class of target domain (in the test set) and randomly selected 100 positive pneumonia samples and 100 samples without pneumonia from the source domain (in the test set).

The representations of the selected cases from different classes of the source domain and the target domain are displayed in Fig. 7. From Fig. 7, we can see that applying our DSDA method makes the target samples more discriminative and the target samples are aligned with each class of source samples. Although the target samples are not separated well in the non-adapted situation (Fig. 7(a) and (b)), they are separated as that of the source samples in the adapted situation (Fig. 7(c) and (d)). Fig. 7(c) and (d) show that our DSDA method effectively separates the samples into two discriminative clusters. It denotes that our formulation of the domain adaptation loss can model the discrimination between the samples from different classes and different domains.

V. Discussion and Conclusion

Deep learning models can learn features that are generically useful across a variety of tasks in various domains [62]. However, due to the domain shift problem [16], deep learning models trained on the source dataset do not perform well on the target dataset [63]. To overcome such a challenge, domain adaptation methods, which minimise the discrepancy between the source domain and the target domain in the feature space, are developed.

This study investigated transferring knowledge from a publicly available dataset as the source domain to improve models' performance on the small-scale or medium-size target dataset. A new framework is designed to transfer knowledge from the source dataset to the target dataset by aligning the samples from domains according to their underlying semantics. Different from the widely-used pre-training strategies, our approach exploits the samples both from the source domain and target domain to transfer the knowledge in an iterative way. Extensive experimental results on the X-ray dataset (our TTSH dataset) and the comprehensive analysis have demonstrated the effectiveness of our DSDA method.

There have been several attempts to diagnose Coronavirus Disease 2019 (COVID-19) using data from CT scans [64]. They have been quoted to be helpful for the early detection of COVID-19. However, chest radiography is performed in significantly larger numbers in clinical practice, especially in resource-constrained health systems where CT is not easily available. Furthermore, several consensus statements by international workgroups have raised concern for the untested specificity of CT for diagnosis where the pre-test probability of COVID-19 infection, and chest radiography is preferred to minimise the risk of nosocomial transmission [65]–[68]. In a recent report, the severity of CXR findings correlates well with clinical severity and could predict severe pneumonia [69]. Thus, deep learning models that aid in automated CXR detection of COVID-19 pneumonia will be valuable for clinical practice. Our proposed approach is very effective to rapid diagnosis pneumonia, and we plan to further develop model that is able to differentiate between COVID-19 pneumonia and non-COVID-19 pneumonia.

In this work, we proposed DSDA to diagnose pneumonia from CXRs. Although it has achieved a promising performance, it does not consider the issue of imbalanced data. In practice, the medical datasets are usually imbalanced, especially for rare diseases. We plan to investigate combinations of domain adaption with imbalanced learning to handle more changing tasks in automated disease diagnoses in our further work.

References

[1] World Health Organization, "Pneumonia," Aug. 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia

[2] Z. Y. Zu *et al.*, "Coronavirus disease 2019 (COVID-19): A perspective from China," *Radiology*, 2020, Art. no. 200490.

[3] World Health Organization, "Weekly epidemiological update-6 Jul. 2021," [Online]. Available: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---6-july-2021

[4] S. Rajaraman, S. Candemir, G. Thoma, and S. Antani, "Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs," in *Proc. Med. Imag. 2019: Comput.-Aided Diagnosis*, vol. 10950, 2019, Art. no. 109500S.

[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, 2019.

[6] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, 2019, Art. no. 101552.

[7] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, 2019.

[8] Z. Gu *et al.*, "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[9] 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.*, 2017, pp. 2097–2106.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[11] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[13] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," 2017. [Online]. Available: https://arxiv.org/pdf/1710.10501.pdf

[14] P. Rajpurkar *et al.* "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017. [Online]. Available: https://arxiv.org/pdf/1711.05225.pdf%202017.pdf

[15] S. Guendel, S. Grbic, B. Georgescu, S. Liu, A. Maier, and D. Comaniciu, "Learning to recognize abnormalities in chest X-rays with location-aware dense networks," in *Iberoamerican Congr. Pattern Recognit.*, 2018, pp. 757–765.

[16] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.

[17] S. Inui *et al.*, "Chest CT findings in cases from the cruise ship Diamond Princess, with Coronavirus Disease 2019 (COVID-19)," *Radiol.: Cardiothoracic Imag.*, vol. 2, no. 2, 2020, Art. no. e200110.

[18] Y.-X. Tang *et al.* "Automated abnormality classification of chest radiographs using deep convolutional neural networks," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–8, 2020.

[19] M. Varma *et al.* "Automated abnormality detection in lower extremity radiographs using deep learning," *Nature Mach. Intell.*, vol. 1, no. 12, pp. 578–583, 2019.

[20] I. Pan, S. Agarwal, and D. Merck, "Generalizable inter-institutional classification of abnormal chest radiographs using efficient convolutional neural networks," *J. Digit. Imag.*, vol. 32, no. 5, pp. 888–896, 2019.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[24] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: https://arxiv.org/pdf/1704.04861.pdf

[25] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *J. Healthcare Eng.*, vol. 2019, pp. 1–7, 2019.

[26] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Comput. Methods Programs Biomed.*, vol. 187, 2020, Art. no. 104964.

[27] S. Rajaraman, S. Sornapudi, P. O. Alderson, L. R. Folio, and S. K. Antani, "Analyzing inter-reader variability affecting deep ensemble learning for COVID-19 detection in chest radiographs," *PLoS One*, vol. 15, no. 11, 2020, Art. no. e0242301.

[28] S. Rajaraman and S. K. Antani, "Modality-specific deep learning model ensembles toward improving TB detection in chest radiographs," *IEEE Access*, vol. 8, pp. 27318–27326, 2020.

[29] O. Yadav, K. Passi, and C. K. Jain, "Using deep learning to classify x-ray images of potential tuberculosis patients," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2018, pp. 2368–2375.

[30] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[31] S. Thrun and L. Pratt, *Learning to Learn*. Berlin, Germany: Springer, 2012.

[32] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, 2010, pp. 242–264.

[33] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 1–46, 2020.

[34] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2010.

[35] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[36] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2960–2967.

[37] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

[38] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3723–3732.

[39] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1640–1650.

[40] Q. Yu, A. Hashimoto, and Y. Ushiku, "Divergence optimization for noisy universal domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2515–2524.

[41] Y. Sawada, Y. Sato, T. Nakada, K. Ujimoto, and N. Hayashi, "All-transfer learning for deep neural networks and its application to sepsis classification," 2017. [Online]. Available: https://arxiv.org/pdf/1711.04450.pdf

[42] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4893–4902.

[43] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Universal domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2720–2729.

[44] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7404–7413.

[45] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[47] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[48] M. V. Enríquez, "A deep learning approach for pneumonia detection in chest X-ray," Master's thesis, Universidade de Vigo, Pontevedra, Spain, 2019.

[49] T. D. Team, "Pneumonia detection in chest radiographs," 2018. [Online]. Available: https://arxiv.org/pdf/1811.08939.pdf

[50] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[51] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[52] X. Li, L. Wang, and E. Sung, "Adaboost with SVM-based component classifiers," *Eng. Appl. Artif. Intell.*, vol. 21, no. 5, pp. 785–795, 2008.

[53] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, May/Jun. 1991.

[54] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, and S. Marshall, "A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers," *Physiol. Meas.*, vol. 35, no. 11, 2014, Art. no. 2191.

[55] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, 2013.

[56] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, 2019.

[57] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 801–814, Apr. 2019.

[58] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[59] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop Stat. Learn. Comput. Vis.*, vol. 1, no. 1–22, 2004, pp. 1–2.

[60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[61] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[62] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[63] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1521–1528.

[64] H. Kang *et al.* "Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2606–2614, Aug. 2020.

[65] American College of Radiology *et al.*, "ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection," Mar. 2020. [Online]. Available: https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection

[66] M.-P. Revel *et al.*, "COVID-19 patients and the radiology department-advice from the European society of radiology (ESR) and the European society of thoracic imaging (ESTI)," *Eur. Radiol.*, vol. 30, no. 9, pp. 4903–4909, 2020.

[67] A. Nair *et al.*, "A british society of thoracic imaging statement: Considerations in designing local imaging diagnostic algorithms for the COVID-19 pandemic," *Clin. Radiol.*, vol. 75, no. 5, pp. 329–334, 2020.

[68] H. Skulstad *et al.* "COVID-19 pandemic and cardiac imaging: EACVI recommendations on precautions, indications, prioritization, and protection for patients and healthcare personnel," *Eur. Heart J.-Cardiovasc. Imag.*, vol. 21, no. 6, pp. 592–598, 2020.

[69] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.