IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 1, February, 2023, pp. 241-248
Journal homepage: http://iieta.org/journals/ts

Towards Domain-Aware Transfer Learning for Medical Image Analysis: Opportunities and Challenges

Marut Jindal*, Birmohan Singh

Department of Computer Science & Engineering, Sant Longowal Institute of Engineering & Technology, Sangrur 148106, India

Corresponding Author Email: marut_pcs1802@sliet.ac.in

https://doi.org/10.18280/ts.400123

Received: 27 August 2022
Accepted: 10 December 2022

Keywords:
transfer learning, medical image analysis, domain-aware learning, deep learning

ABSTRACT

The tremendous success of transfer learning (TL) in natural imaging has also motivated the researchers in biomedical imaging. A lot of methods utilizing TL have been proposed, however, only a few have emphasized on its actual impact in biomedical tasks. In this article, we review the current landscape of TL in medical image analysis, and outlined the existing myths and related findings. We found that there exists substantial lack of medically specialized (domain-specific) pretrained transfer learning models, which can significantly benefit the biomedical imaging. Thus, to further explore our opinion experimentally, we identified three large datasets previously available from different medical areas and pre-trained the standard CNN models on them, both separately and on aggregated dataset. These pre-trained models are then transferred for five different target medical tasks and their performance is compared. The comparison has shown promising benefits of domain-aware learning and aggregated generalized medical TL models along with associated challenges. We believe the outcomes of this work will encourage the community to rethink the existing de-facto ImageNet TL standard, and work for the domain-specific TL.

1. INTRODUCTION

Modern research practices in computer vision and pattern recognition have emphasized the capabilities of deep learning [1], especially Convolutional Neural Networks (CNN) [2], achieving state-of-the-art results and setting benchmarks. Though, the success of these models highly depends upon the quality and the amount of labelled data available for training. The applications with a small data regime face overfitting and underperformance for unseen data. A notable solution to mitigate this small dataset problem in CNNs is transfer learning (TL), i.e., knowledge transfer from one dataset to another. TL has successfully established its dominance in natural vision, thanks to the available fine-trained CNN models on million-sized well-annotated datasets like ImageNet [3] and MS COCO [4]. However, no such large-sized datasets are available in medical imaging, so there is a lack of domain adapting pretrained models.

The current standard practice in Medical Image Analysis (MIA) is the use of ImageNet pre-trained models for the target medical tasks. In this approach, the models pretrained on ImageNet are used instead of training from scratch (randomly initialized parameters). These models can either be applied as feature extractors or used as base models (pretrained weights rather than random initialization) for further fine-tuning on target datasets. Researchers have explored both approaches [5], among which the latter is more favored by the latest.

Despite the broad applicability of ImageNet-trained models, little work exists studying its meticulous effects in medical imaging. Some recent works [6-9] have studied the impact of ImageNet-trained TL in MIA and concluded that it only benefits the early convergence. Additionally, the initial layers of the TL architectures are more responsible for knowledge reuse [7, 8]. Another conclusion was that the smaller and simpler architectures could perform comparably to ImageNet architectures in MIA [8]. Thus, the question arises is whether the ImageNet models are the right option for TL in MIA or does we require some specialized domain-aware medically trained models. Moreover, some commonly held beliefs related to TL in the natural domain have been challenged by the findings of the studies [10-14]. Therefore, in MIA, questions like the above become more critical since the natural and medical imaging domains have considerable differences. Rajpurkar et al. [15] have also addressed the same concern and proposed CheXNet (a model specialized for chest-related diagnosis from X-rays), demonstrating the superiority of domain knowledge. Similarly, Chen et al. [6] trained a biomedically dedicated in-domain transfer model named 'Med3D', though their work was limited to 3D medical images.

A recent work [16] has surveyed the articles on incorporating biomedical domain knowledge while training deep learning models and found domain adaptation as a promising futuristic research direction in MIA. The domain adaptation in MIA is a vast area of research which is not only limited to imaging data but all kinds of information that help the doctors in their excellence. The current article is concise on the need and importance of domain-aware TL for MIA which is a subpart of domain adaptation and could be an immediate action towards better performance of automated models. We have reviewed the latest work studying the impact of knowledge transfer in both natural and biomedical domains to generalize the opportunities and challenges better. We also performed additional TL experiments on biomedical datasets to assess domain-aware learning and its transfer.

The rest of the article is organized as follows. Section 2 reviews the existing work studying the widespread impact of ImageNet TL. Next section describes the work analyzing TL's effect on biomedical tasks. Then, section 4 summarizes some commonly held TL queries and related findings. The following section discusses our experimental results and findings. Lastly, section 6 outlines the challenges and future directions related to domain-aware and unified medical TL.

2. TRANSFER LANDSCAPE OF IMAGENET: OUTCOMES VS. BELIEF

First and foremost work searching for the answers to an obvious question, "What makes ImageNet good for TL?" was done by Huh et al. [12]. They performed several experiments on three tasks: object detection, action classification, and scene classification, to study various related facts like the importance of the number of samples, classes, samples-per-class vs. the number of classes, and the role of coarse-grained vs fine-grained classes. Though the question was still unanswered, they found that a substantial decrease in sample count (50% per class) and class count (only 127 out of 1000) shows a tiny effect on TL performance, which drive the community to rethink the rule "more data better performance". However, the derived conclusions cannot be generalized yet, as much more investigation is required in the field. Or, it may be possible that the derived facts only hold for the studied target tasks, as the object detection-like tasks are entirely different from classification in which localization is the key concern rather than recognition. Likewise, He et al. [11] also hold a similar consent, as they too studied object detection and instance segmentation tasks. They aimed to compare the transfer models with randomly initialized (RI) models. They experimented on the MS COCO dataset, which is considered a sufficiently well-organized dataset. Based on their experiments, the key observations were i) the TL speed-up the convergence in early phase but did not guarantee better regularization, and ii) the TL does not show benefits to localization related tasks. Moreover, they also remarked that, the RI models can perform comparable of TL but need extra training iterations, and the ultimate goal of the community still hold to pursuit for learning universal representations.

Some other studies have also challenged the de-facto ImageNet trained transfer. Geirhos et al. [10] have shown in their work that the ImageNet trained models are more inclined towards the textures of the images rather than the shapes; thus, these are less general than previously thought. Furthermore, they proposed a stylized version of ImageNet on which the models trained can also learn shapes. Similarly, Kornblith et al. [13] have also concluded that, although the ImageNet models generalize well over datasets, their weights are less general than assumed. They also showed that better architectural designs perform better on target tasks.

Unlike others, Ngiam et al. [14] have investigated a different aspect of TL, i.e., choice of source data. They investigated another source dataset JFT [17] along with ImageNet for knowledge transfer to six target classification datasets. Like previous findings, they suggest that improved performance is not always necessary via extensive pretraining samples. However, domain adaptation, i.e., matching the source and target data distribution, is the main idea for better transfer which is also agreed by the outcomes of the studies [18, 19].

3. TRANSFER'S IMPACT IN BIOMEDICAL: IMAGENET VS. DOMAIN-SPECIFIC LEARNING

Knowledge transfer is extremely popular in biomedical imaging because of smaller datasets (a few 100 to thousands). Researchers have utilized various transfer strategies and reported state-of-the-art results. Morid et al. [5] reviewed 102 studies on TL (from ImageNet) in biomedical imaging and found that most studies with data samples <1000 have applied the feature extraction TL strategy while the studies with samples >1000 have applied the fine-tuning based TL approach. However, some recent studies investigating ImageNet transfer on biomedical datasets have propounded the community to reconsider it.

The pioneering work [8] has attempted to acknowledge the open questions, "how much ImageNet features are reusable for medical imaging, and exactly where?", "does there exist differences in the filters learned by TL and RI models?" and "how do model filters get affected with pretraining?". Authors have highlighted various non-trivial differences between natural and medical imaging, on account of which they organized several experiments on two datasets (CheXpert and Retina) and found that i) shallow and simpler models can perform equivalent to standard ImageNet models, ii) feature reuse occurs only at initial layers, iii) larger models (both RI and TL) changes less in starting layers after fine-tuning, i.e., over-parameterization for MIA tasks, iii) converged smaller models show similar filters for both RI and TL, i.e., no feature reuse, iv) TL does offer feature independent benefits like convergence speed, even the scale of pretrained weights adapted for RI helps to converge faster. However, recently [9] has challenged some findings of Raghu et al. [8] and pointed out three profound limitations of the experiments: 1) *Poor evaluation metric*- As the datasets are highly imbalanced, the used metric AUC is not a good choice; 2) *Unrepresentative target datasets*- Both the datasets contain thousands of samples, while MIA datasets usually range from hundreds to few thousands; 3) *Rigid TL methods*- Alternative strategies like truncated models can also be explored. Peng et al. [9] measured AUROC and AURPC performance metrics for classification on two datasets (CheXpert and Covid), and found that TL mostly outperform RI for both shallow and deep models, mainly smaller datasets benefit more. They also found that the truncated TL models perform better than conventional and hybrid TL methods. However, their experimentation still holds the second limitation mentioned above, since, both the datasets are of chest X-rays.

Neyshabur et al. [7] further investigated "what is being transferred in TL?" performing a series of TL comparisons (feature similarity, *l2* distance, and loss basins) on both natural and medical datasets. They arranged the target domains into decreasing similarity (i.e., Real>Clipart>CheXpert>Quickdraw) with the source dataset (ImageNet) and found the highest TL performance boost for similar data (Real domain). Moreover, distant domains also get profited with TL along with convergence speed. These results show the importance of in-domain learning and additional low-level benefits to distant-domain. To further verify the low-level advantages of TL, the authors shuffled the different sized blocks (from full image to 1-pixel) of images to destroy its visual information. They observed that the performance falls with a decrease in block size, but the performance of TL is still better than RI even for the fully shuffled images. These results not only give an idea of how TL

is beneficial for distant domains, but also advocate rethinking the source domain. Following it, many researchers have tried to utilize previously available biomedical datasets as transfer datasets.

Med3D [6] was among the initial medically trained models used as TL on diverse and different medical tasks. A recent work by Zoetmulder et al. [20] too advocated the optimality of domain and task-specific TL for brain MRI lesion segmentation. Similarly, Alzubaidi et al. have shown the dominance of same-domain TL over the conventional approach in their research [21-23], considering skin lesion and breast cancer histopathology images. Other than these, Xie et al. [16] have provided a comprehensive survey of more than 200 papers (mainly ranging in 2017-2020), among which 163 are purely based on domain knowledge incorporation. They partitioned the domain knowledge into two categories, i.e., knowledge incorporation from natural and medical datasets (TL models trained over ImageNet and other similar or dissimilar medical datasets) and knowledge incorporation from doctors (such as tags from health records, training and general diagnosis patterns of doctors, particular patterns or areas that the doctors focus while diagnosing, etc.). Further, they grouped the articles into four medical categories i) disease diagnosis, ii) lesion or abnormality detection, iii) organ or lesion segmentation, and iv) other medical applications (like image reconstruction, retrieval, and generation). The critical realization from their work is that domain adaptation is the path to be followed for the success of automated diagnosis. However, the identification, selection, representation and incorporation of the domain knowledge are challenging and demand a well-organized research in collaboration of the medical and data-science community.

4. SUMMARY

The previous sections have reviewed the work analyzing TL's impact in both natural and medical domains. Many imperative questions related to TL that required extensive experimentation and analytical work were addressed. Here we outline them with related findings. *1) Does TL have benefits over RI?* Yes, TL is advantageous over RI. Benefits are more visible for smaller datasets. It boosts both the performance and convergence speed, though the localization-related tasks may have lesser gain. *2) What type of benefits are offered by TL?* TL offers the guaranteed advantage to model's convergence. TL models are much faster than RI models and requires few iterations to converge. Even the scale adaptation of pretrained weights has shown an advantage in the convergence of RI models. Another benefits are related to feature reuse which depends upon the target domains. Normally, the reuse is limited to initial layers, though, for similar target tasks it can be realized in later layers too. *3) Do ImageNet TL models improve the MIA performance?* Yes, many researchers have shown that even for a far distant problem, TL is better than RI. However, some recent studies have shown the superiority of the domain-specific TL models, but MIA lacks in having medically-aware TL models. Therefore, in the absence of that, ImageNet-trained models are beneficial. *4) What is the impact of domain knowledge in DL? Will it benefit TL approaches in MIA?* Domain knowledge has positively impacted the performance of DL methods in MIA. Researchers have reported the state-of-the-art results utilizing it. However, there exist specific challenges, regarding how and what type of domain knowledge can be utilized. It needs extra attention from the research community. In TL, existing analysis has shown a higher performance boost for similar datasets than distant ones. Moreover, TL models pretrained in the medical domain dominate over ImageNet trained models. Thus, we can infer that domain-aware TL will benefit MIA.

Examining the overall status of TL in MIA, we found that, like ImageNet in natural vision, the medical domain also need unified datasets that can be used as source databases for TL. However, it is exceptionally challenging and laborious to create them, as it will involve the efforts of both data scientists and medical practitioners. Thus, other tranquil and approachable initiatives can be taken at the moment, e.g., the utilization of already available well-annotated medical image databases. Moreover, efforts can be made to explore their combined effect in transfer. These approaches further pose some new questions like; *Does available medical datasets are advantageous over ImageNet in MIA? Is it possible to aggregate them? If yes, what will be its impact compared to ImageNet and in-domain datasets?* In this direction, more tests are needed apart from existing analytical TL research. Thus, to fulfil the gap, we have organized some experiments which are described in the following section along with obtained results.

5. EXPERIMENTAL ANALYSIS

In this section, first, we discuss the type and purpose of the experiments conducted. It also includes the details of CNN network used for analysis. Next, the description of the source, target, and aggregated datasets is given. Then, we show the results and their comparison inferring the insights and related challenges. Finally, the experimental limitations of the current work have been discussed.

5.1 Methodology

As summarized in the previous section, domain-aware learning can significantly benefit TL in MIA. Thus, to verify its generalized potential, we analyzed three areas of medical images consisting of diverse feature space, i.e., chest X-Rays, histopathology whole slide images, and eye fundus images. As the study aims to analyze domain-aware TL, we selected two types of datasets from each area. One is used as the source dataset to pretrain the transfer model, and the other is the target dataset for which the pretrained model transfers. More details about these datasets are discussed in the next section. Further, for every target task/dataset, five types of transfer are performed to precisely analyze their behavior, which is summarized below.

- **IN:** Transfer of conventionally used ImageNet weights.
- **Domain_RI:** Weights transfer of randomly initialized (RI) model pretrained on similar domain source dataset.
- **Domain_IN:** Weights transfer of ImageNet initialized (IN) model pretrained on similar domain source dataset.
- **All_RI:** Weights transfer of randomly initialized (RI) model pretrained on the dataset aggregating all source datasets of considered three areas.
- **All_IN:** Weights transfer of ImageNet initialized (IN) model pretrained on the dataset aggregating all source datasets of considered three areas.

The above-listed transfer experiments are performed to verify the advantage of both the domain-specific learning and

unified medical-aware dataset. Among these, 2nd and 3rd are dedicated domain TL models, while 4th and 5th are generalized medical TL models. Further details of dataset's aggregation method (for generalized TL models) are discussed in the next section.

The base CNN model used in this study for analysis is the DenseNet-121. Besides fewer parameters, its architectural design has shown a significant advantage over other CNN models in both the natural and medical fields. Thus, we have selected this model for all experiments. Apart from its basic architecture, a customized head network is also attached on top of it, consisting of a combination of fully connected (FC), batch normalization (BN), and dropout layers. Figure 1 shows the procedure followed and the network used in this work.

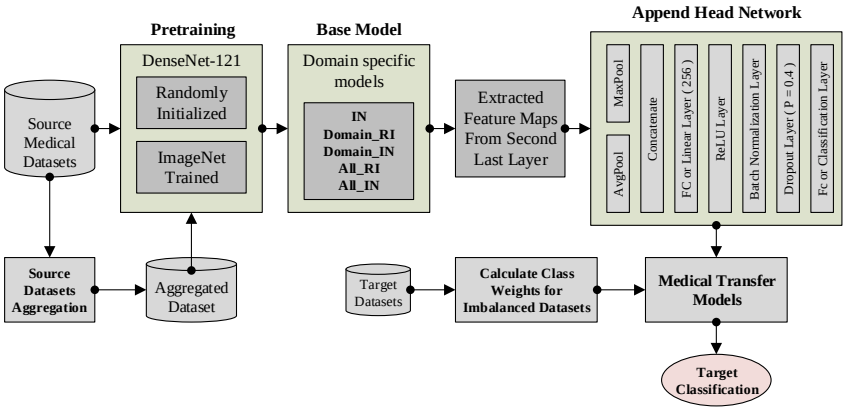

Figure 1. Workflow and network detail

5.2 Datasets description and processing

This work uses three source datasets (one from each area) and five target datasets for experimentation. We have selected those dataset as a source dataset that are used in literature by other authors also, e.g., '*ChestX-ray 14*' from chest X-ray images [15], '*BreakHis*' from histopathology whole slide images [22, 23], and '*EyePACS*' from eye fundus images [24]. Among five target datasets, two are chest X-ray datasets ('*NLM-ChinaCXRSet*' and '*COVID-19 Radiography*'), one is histological dataset of human colorectal cancer images ('*Kather_Texture_2016*'), and two are diabetic retinopathy datasets of eye fundus images ('*Messidor*' and '*DeepDRiD*'). All these datasets are resized to 224×224 resolution in this work. Following sub-sections 5.2.1 and 5.2.2 describes these datasets and their corresponding preprocessing.

Besides the datasets mentioned above, we also pretrained the models on an aggregated dataset (All_three) by collating the samples from all source datasets. Sub-section 5.2.3 describes the details related to the dataset and followed aggregation procedure.

5.2.1 Source datasets

ChestX-ray14**.** ChestX-ray14 [25] is a multi-label frontal-view chest X-ray dataset with fourteen common thorax disease categories. It consists of 112,120 X-rays (resolution of 1024×1024 pixel) of 30,805 unique patients with fourteen disease labels otherwise labelled as 'No finding'. X-rays in this dataset may have multiple labels from the fourteen categories mentioned in the study of Wang et al. [25]. This dataset outputs two pretrained domain-specific models named CXR_RI and CXR_IN.

BreakHis. BreakHis [26] is a dataset of 9109 microscopic WSI images of breast tumor tissues collected from 82 patients. It consists of 5429 malignant and 2480 benign samples of 700×460 resolution at 40X, 100X, 200X, and 400X magnifying factors. Figure 2 shows the samples from these magnifying factors in third row, respectively. In this work, we have pretrained the models on collated images from all magnifying levels to train a generalized classification model for all levels. Before training, the dataset has been resized to 340×224 pixels without affecting the overall L/W ratio. However, only 224×224 sized randomly selected images have been used while training, which produces HWSI_RI and HWSI_IN transfer models. Following Figure 2 shows row-wise sample images from '*ChestX-ray 14*', '*EyePACS*', and '*BreakHis*' datasets before and after processing.

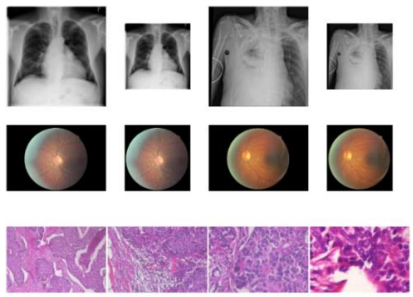

Figure 2. Sample images from three source datasets

EyePACS. EyePACS is a dataset collected from diabetic retinopathy (DR) detection challenge on Kaggle [27] organized by California Healthcare Foundation and eyepacs. It contains 35126 and 53576 images in the training and test sets labelled with five DR grading levels ranging from 0-4. The dataset is highly imbalanced (i.e., 0 level has ~30 times more sample than 3 and 4 levels), which causes hindrance in training. Therefore, samples of 1-4 levels from the test set have also been added in training (i.e., 14043 more samples are added to the training set), and then 10% samples of the dataset are used for validation in this work. Class weights have been used further to balance the distribution while optimizing loss. The samples in this dataset do not have a uniform resolution; thus, the images have been center cropped to extract eye area with 224×224 size (from images resized to 340×224 size keeping aspect ratio). The transfer models pretrained on this dataset are named EFI_RI and EFI_IN in this work.

5.2.2 Target datasets

All the five target datasets used in this work are summarized in the below Table 1.

Table 1. Target datasets details

Sr. No.	Dataset Name	Image Type	Classes	Samples	Resolution
1.	NLM-ChinaCXRSet [28] (**D1**)	CXR (Gray)	Normal Tuberculosis	326 336	Variable
2.	COVID-19 Radiography [29, 30] (**D2**)	CXR (Gray)	COVID Lung_Opacity Normal Viral_Pneumonia	3616 6012 10192 1345	299×299
3.	Kather_Texture_2016 [31] (**D3**)	HWSI (RGB)	Tumor, Stroma, Complex, Lympho, Debris, Mucosa, Adipose, Empty	Each class with 625 samples	150×150
4.	Messidor [32] (**D4**)	EFI (RGB)	Grade_0 Grade_1 Grade_2 Grade_3	541 154 247 254	Variable
5.	DeepDRiD [33] (**D5**)	EFI (RGB)	0-No DR 1-Mild NPDR 2-Moderate NPDR 3-Severe NPDR 4-Proliferate DR	714 186 326 282 92	Variable

5.2.3 Aggregated dataset

To the best of our knowledge, in MIA, literature has most studies on ImageNet transfer and a few on similar domain. However, the combined effect of different medical domains could also be explored. Therefore, apart from similar domain transfer, we also examined the effect of the unified medical transfer model on target medical datasets. To train this model, the dataset is prepared by collecting the samples together from all considered (three) source datasets. Among these datasets, one is a multi-label dataset (ChestX-ray14), while the other two are single-label classification datasets. Thus, to bring them in a single orientation, all are considered multi-label datasets with a total of 22 (i.e., 15+2+5) labels. These labels are inputted as a one-hot encoding vector of length 22. Moreover, the data is inputted to the model with a data-generator at size 224×224, where images of BreakHis dataset are selected randomly from a size of 340×224. After aggregation, the dataset contains 169198 images, partitioned into 90/10% for train/validation sets.

5.3 Results and discussion

We evaluate the five transfer models for every target dataset, one- natural pretrained, two- domain-specific, and two- medically generalized, to compare their applicability over each other. Moreover, two types of parameters initialization have also been considered to measure their impact on overall training and performance. The comparison has been made considering accuracy and cohen-kappa score as performance indicators. In literature, accuracy is the most used indicator for evaluation of classification tasks, but, for some tasks where the classes are in the form of grading, cohen-kappa score is preferred. Therefore, in this work we have used these two metrics accordingly. Figure 3 shows the training and validation accuracy curves of three datasets for a fixed number of epochs, a) NLM-ChinaCXRSet, b) COVID-19 Radiography, and c) Kather_Texture_2016. All three datasets have shown high-performance gain with generalized medical transfer models (All_XX) compared to others. Their learning curves also showed an early and stabilized flow towards convergence. However, the natural trained model (IN) showed the least performance and an unstable convergence curve. Domain-specific models have shown mixed behavior. Though, both Domain_RI and Domain_IN have performed better than IN, Domain_RI showed higher accuracy than Domain_IN for a) and b) datasets, while for c) it is the opposite.

Table 2. Model's performance at last epoch

Dataset		Accuracy (%)				
		IN	**Dom RI**	**Dom IN**	**ALL RI**	**ALL IN**
D1	Training	95.6	98.6	96.8	99.1	99.5
	Validation	86.7	90.0	91.7	90.0	91.7
D2	Training	99.9	99.9	99.9	100	100
	Validation	96.5	95.9	96.0	96.2	96.6
D3	Training	98.7	98.1	98.2	99.6	99.6
	Validation	95.1	93.9	95.2	94.4	95.2
		Kappa Score (%)				
D4	Training	77.5	97.5	95.5	97.8	98.6
	Validation	52.1	38.8	51.0	33.6	45.3
D5	Training	90.0	99.6	99.3	100	100
	Validation	36.1	47.3	46.0	31.6	46.2

Likewise, Figure 4 shows the training and validation curve behavior of eye datasets, i.e., d) Messidor and e) DeepDRiD. The cohen_kappa score has been plotted for these two datasets as a metric measure due to high-class imbalance, and balanced class weights are also used while training. Both the datasets have shown similar performance behavior to a), b), and c) for training sets, though the IN model does not converge in the fixed number of epochs. On the other hand, for both datasets, all models' validation curves do not show a high metric value and oscillate between short ranges. Despite the nearly equal performance, sharp ditches have been shown by the IN model for the d) dataset. However, for the e) dataset, All_IN and domain-specific models (EFI_IN and EFI_RI) have shown better metric values than the other two. Same can also be inferred from the metric values at last training epoch shown in Table 2.

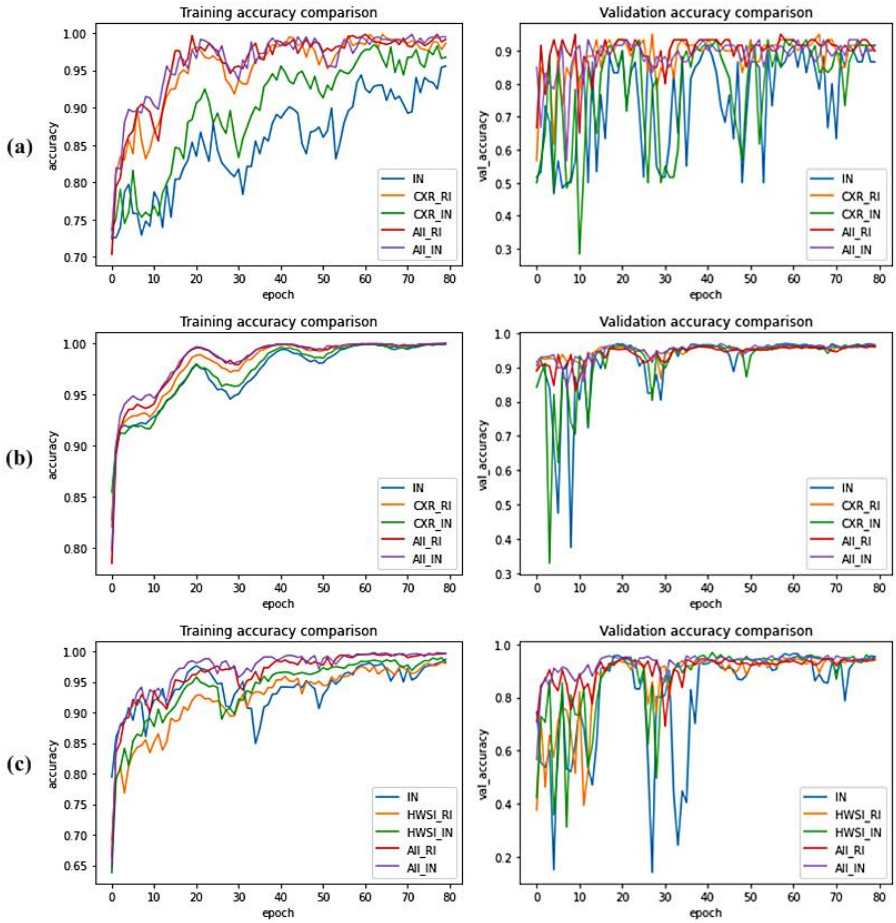

Figure 3. Training and validation curves comparison of three datasets a) NLM-ChinaCXRSet b) COVID-19 Radiography c) Kather_Texture_2016

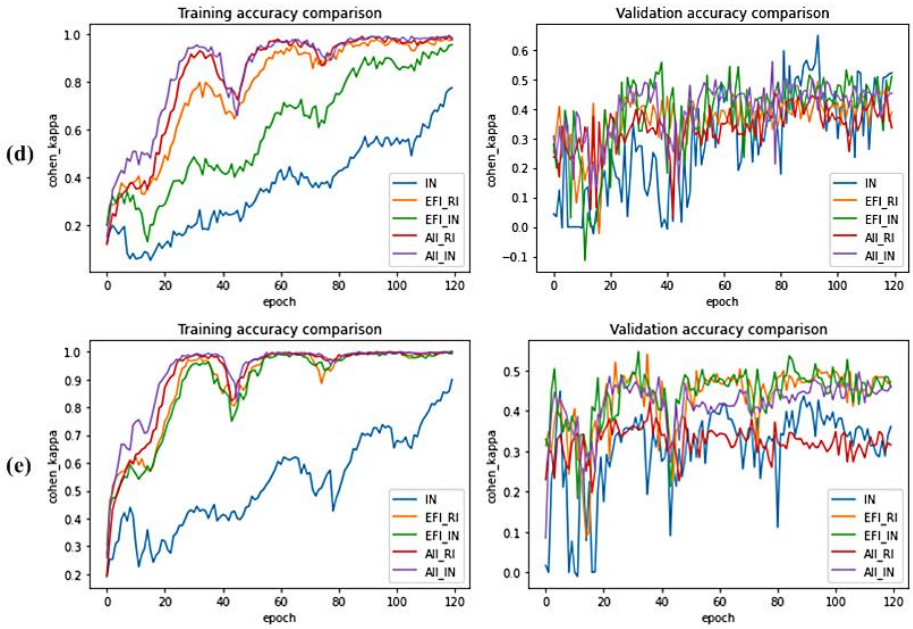

Figure 4. Training and validation curves comparison of eye datasets d) Messidor e) DeepDriD

Analyzing the convergence curves and metric results, we find that generalized medical models (All_RI and ALL_IN) have shown the best results for almost every dataset, and 4 out of 5 datasets have shown competing results by randomly initialized domain-specific models (Domain_RI). Additionally, these models provide early, stable, and smooth transition towards convergence, which IN lacks. Thus, the intuition of domain-aware and unified medical learning seem beneficial and promising, which can be further explored with careful data aggregation and model training.

5.4 Experimental limitations

The main limitations of this work lie with the datasets aggregation and models training, e.g., the aggregated dataset has imbalanced data distribution, and class weights are also not used. As this study's main motive is to show the proficiency of unified medical models and domain-aware learning, no extra efforts have been made toward perfect aggregation or collative training. However, a simplified approach is followed. Thus, this research direction needs further analysis which must include more datasets and heterogeneous training methods over diverse medical data.

6. RESEARCH CHALLENGES AND FUTURE DIRECTIONS

Analyzing the available literature and experimental results, we summarize the challenges and future directives in the path of unified medical-aware TL. These are challenges related to unified medical dataset, challenges related to model training.

- Challenges related to a unified medical dataset: The main challenge lies with universal medical-aware TL models is the dataset construction. As the medical domain involves multiple modalities to examine and capture different body parts, there exist high spatial and intensity differences between them. Moreover, region of interest (ROI) for different domains are also diverse, e.g., small red spots and blood vessels in eye fundus images are indicative of diabetes, nucleus blasts and structural deformations in whole slide images are indicative of cancers, texture and colour patterns of skin lesion images are indicative of melanoma, etc. Thus, the ROIs ranged from small spots (1-5% image area) to large lesion areas (50-90% image area). Other than these, some medical applications involve colored images (skin, eye, and whole slide images) while others do not (X-rays, MRI, CT, mammograms, etc.). Additionally, if we consider segmentation datasets, they are too ambiguous [6]. Some domains have segmented annotations of only lesions but not surrounding organs, while others provide only organ segmentations. Further, all the organs are not annotated in an image. For example, the liver dataset has only the liver area's annotation but not the surrounding pancreas, and the pancreas/lesion dataset has only the pancreas/lesion area annotated. Therefore, these diversities pose a significant challenge to the research community, and initiatives can be taken to find constructive paths. However, to avoid it, another option may be the separate enrichment of fully-annotated samples in all domains and having separate TL models for every domain, but the motive of the work is to utilize already available datasets and train ImageNet-like unified models.
- Challenges related to model training: Another challenge associated with unified TL models is finding reliable ways of model training over diverse datasets. The available annotated datasets from different areas vary in the number of samples available (like in the present study). Also, the datasets themselves contain unbalanced class distributions (e.g., EyePACS). Thus, aggregated unified dataset becomes highly imbalanced, driving the training towards overfitting. Moreover, a high difference in pixels representation and values also causes hindrance to learning. Different modalities and tasks have different intensity ranges, so the histograms show polarized groupings, i.e., confusing the feature learning process (the parameters learned for some parts of the dataset may not be helpful for other parts). Thus, a confined way of training and architectural design is required. Lastly, segmentation datasets are not fully annotated as discussed earlier. Therefore, some specific ways of training, like the study [6] trained a model (single_encoder –eight_decoder branch model) for eight different segmentation datasets, is needed.

7. CONCLUSIONS

In this study, we investigated domain-aware transfer learning in medical image analysis. The main focus of the work was to explore its impact, so we reviewed various analytical studies on TL both in the natural and biomedical fields. We found that ImageNet transfer models are not much advantageous for medical tasks; instead, domain knowledge can provide significant benefits. Moreover, initiatives can be taken to train unified medical models like ImageNet in natural vision. To further verify our intuition, we organized some experiments and found that domain-aware learning is beneficial over traditional ImageNet transfer. Our exploration of unified/generalized medical transfer models showed even better results for all datasets. However, particular challenges need the attention of research community. In conclusion, until well organized and trained common medical-aware TL models are not available, the domain-specific TL is better choice than the natural domain IN models for MIA.

REFERENCES

[1] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553): 436-444. https://doi.org/10.1038/nature14539.

[2] Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. Insights Into Imaging, 9(4): 611-629. https://doi.org/10.1007/s13244-018-0639-9

[3] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Li, F.F. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252. https://doi.org/10.1007/s11263-015-0816-y

[4] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L. (2014). Microsoft coco: Common objects in context. In European Conference on Computer Vision, pp. 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[5] Morid, M.A., Borjali, A., Del Fiol, G. (2021). A scoping review of transfer learning research on medical image analysis using ImageNet. Computers in Biology and Medicine, 128: 104115. https://doi.org/10.1016/j.compbiomed.2020.104115

[6] Chen, S., Ma, K., Zheng, Y. (2019). Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625. https://doi.org/10.48550/arXiv.1904.00625

[7] Neyshabur, B., Sedghi, H., Zhang, C. (2020). What is being transferred in transfer learning? Advances in Neural Information Processing Systems, 33: 512-523. https://doi.org/10.48550/arXiv.2008.11687

[8] Raghu, M., Zhang, C., Kleinberg, J., Bengio, S. (2019). Transfusion: Understanding transfer learning for medical

imaging. Advances in Neural Information Processing Systems, 32: 3342-3352. https://doi.org/10.48550/arXiv.1902.07208

[9] Peng, L., Liang, H., Li, T., Sun, J. (2021). Rethink Transfer Learning in Medical Image Classification. arXiv preprint arXiv:2106.05152. https://doi.org/10.48550/arXiv.2106.05152

[10] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231. https://doi.org/10.48550/arXiv.1811.12231

[11] He, K., Girshick, R., Dollár, P. (2019). Rethinking ImageNet pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4918-4927. https://doi.org/10.48550/arXiv.1811.08883

[12] Huh, M., Agrawal, P., Efros, A.A. (2016). What makes ImageNet good for transfer learning? arXiv preprint arXiv:1608.08614. https://doi.org/10.48550/arXiv.1608.08614

[13] Kornblith, S., Shlens, J., Le, Q.V. (2019). Do better ImageNet models transfer better? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2661-2671. https://doi.org/10.48550/arXiv.1805.08974

[14] Ngiam, J., Peng, D., Vasudevan, V., Kornblith, S., Le, Q.V., Pang, R. (2018). Domain adaptive transfer learning with specialist models. arXiv preprint arXiv:1811.07056. https://doi.org/10.48550/arXiv.1811.07056

[15] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ng, A.Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225. https://doi.org/10.48550/arXiv.1711.05225

[16] Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., Yu, S. (2021). A survey on incorporating domain knowledge into deep learning for medical image analysis. Medical Image Analysis, 69: 101985. https://doi.org/10.1016/j.media.2021.101985

[17] Hinton, G., Vinyals, O., Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint, arXiv:1503.02531. https://doi.org/10.48550/arXiv.1503.02531

[18] Ge, W., Yu, Y. (2017). Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1086-1095. https://doi.org/10.48550/arXiv.1702.08690

[19] Cui, Y., Song, Y., Sun, C., Howard, A., Belongie, S. (2018). Large scale fine-grained categorization and domain-specific transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4109-4118. https://doi.org/10.48550/arXiv.1806.06193

[20] Zoetmulder, R., Gavves, E., Caan, M., Marquering, H. (2022). Domain-and task-specific transfer learning for medical segmentation tasks. Computer Methods and Programs in Biomedicine, 214: 106539. https://doi.org/10.1016/j.cmpb.2021.106539

[21] Alzubaidi, L., Fadhel, M.A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., Oleiwi, S. (2020). Towards a better understanding of transfer learning for medical imaging: a case study. Applied Sciences, 10(13): 4523. https://doi.org/10.3390/app10134523

[22] Alzubaidi, L., Al-Shamma, O., Fadhel, M. A., Farhan, L., Zhang, J., Duan, Y. (2020). Optimizing the performance of breast cancer classification by employing the same domain transfer learning from hybrid deep convolutional neural network model. Electronics, 9(3): 445. https://doi.org/10.3390/electronics9030445

[23] Alzubaidi, L., Al-Amidie, M., Al-Asadi, A., Humaidi, A.J., Al-Shamma, O., Fadhel, M.A., Duan, Y. (2021). Novel transfer learning approach for medical imaging with limited labeled data. Cancers, 13(7): 1590. https://doi.org/10.3390/cancers13071590

[24] Patil, M., Chickerur, S., Bakale, V., Giraddi, S., Roodagi, V., Kulkarni, Y. (2021). Deep hyperparameter transfer learning for diabetic retinopathy classification. Turkish Journal of Electrical Engineering and Computer Sciences, 29(8): 2824-2839. https://doi.org/10.3906/elk-2105-36

[25] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097-2106. https://doi.org/10.1109/CVPR.2017.369

[26] Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L. (2015). A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering, 63(7): 1455-1462. https://doi.org/10.1109/TBME.2015.2496264

[27] Kaggle. (2015). Diabetic Retinopathy Detection. https://www.kaggle.com/competitions/diabetic-retinopathy-detection/overview.

[28] Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., McDonald, C.J. (2013). Automatic tuberculosis screening using chest radiographs. IEEE Transactions on Medical Imaging, 33(2): 233-245. https://doi.org/10.1109/TMI.2013.2284099

[29] Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, M.T. (2020). Can AI help in screening viral and COVID-19 pneumonia? IEEE Access, 8: 132665-132676. https://doi.org/10.1109/ACCESS.2020.3010287

[30] Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Chowdhury, M.E. (2021). Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Computers in Biology and Medicine, 132: 104319. https://doi.org/10.1016/j.compbiomed.2021.104319.

[31] Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Zöllner, F.G. (2016). Multi-class texture analysis in colorectal cancer histology. Scientific Reports, 6(1): 1-11. https://doi.org/10.1038/srep27988

[32] Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Klein, J.C. (2014). Feedback on a publicly distributed image database: the Messidor database. Image Analysis & Stereology, 33(3): 231-234. https://doi.org/10.5566/ias.1155

[33] Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Zhang, P. (2022). DeepDRiD: Diabetic retinopathy-grading and image quality estimation challenge. Patterns, 100512. https://doi.org/10.1016/j.patter.2022.100512