Cats or CAT scans: transfer learning from natural or medical image source data sets?

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biomedical Engineering

Cats or CAT scans: Transfer learning from natural or medical image source data sets?

Veronika Cheplygina

Abstract

Transfer learning is a widely used strategy in medical image analysis. Instead of only training a network with a limited amount of data from the target task of interest, we can first train the network with other, potentially larger source data sets, creating a more robust model. The source data sets do not have to be related to the target task. For a classification task in lung computed tomography (CT) images, we could use both head CT images and images of cats as the source. While head CT images appear more similar to lung CT images, the number and diversity of cat images might lead to a better model overall. In this survey, we review a number of articles that have studied similar comparisons. Although the answer to which strategy is best seems to be 'it depends', we discuss a number of research directions we need to take as a community to gain more understanding of this topic.

Addresses

Eindhoven University of Technology, the Netherlands

Corresponding author: Cheplygina, Veronika (v.cheplygina@tue.nl)

Current Opinion in Biomedical Engineering 2019, **9**:21–27

This review comes from a themed issue on **Future of BME: Digital Heath and BME**

Edited by **George Truskey**

https://doi.org/10.1016/j.cobme.2018.12.005

Keywords

Medical imaging, Deep learning, Transfer learning.

Introduction

In recent years, transfer learning has become a popular technique for training machine learning classifiers [7,17,22]. The idea is to transfer information from one classification problem (the source) to the next (the target), thereby increasing the amount of data seen by the classifier. This is important for medical imaging, where data sets can be relatively small. In this review, we look specifically at a type of transfer learning — training a network on one type of data and then further training it on (a small amount of) possibly unrelated type of data. An illustration of this procedure is shown in Figure 1.

Training a neural network for a target data set is typically achieved by one of the three main strategies:

- Training the network 'from scratch' or 'full training', that is, randomly initializing the weights and only using data from the target domain for training. In this case, no transfer learning is done.
- Using 'off-the-shelf' features, that is, training a network on source data, using this pretrained network to extract features from the target data, and training another classifier, for example, a support vector machine, on the extracted features. This is a type of transfer learning.
- Training with 'fine-tuning', that is, training a network on source data and then using this network to initialize the weights of a network that is further trained with target data. During training, some layers can be 'frozen' so that their weights do not change. This is another type of transfer learning.

More details on each strategy can be found in the studies by Litjens et al. and Yamashita et al. [22,37].

In transfer learning, the source problem may be seemingly unrelated to the target problem that is being solved. For example, ImageNet [30], a large-scale data set for object recognition, has been successfully used as source data for many medical imaging target tasks, with the studies by Bar et al. [2], Ciompi et al. [10], and Schlegl et al. [31] among the earliest examples. Using other medical data sets as source data is less frequent, possibly because pretrained models are not as conveniently available as models trained on ImageNet, which are included in various toolboxes. It is therefore unclear whether pretraining on ImageNet is indeed the best strategy to choose for transfer learning in medical imaging.

In this review article, we review a number of articles that have used multiple source and/or target data sets, where the target data sets are from the medical imaging domain. The articles were selected by searching Google Scholar for 'transfer learning' and 'medical imaging' or 'biomedical imaging'. As of September 2018, this yielded close to 2 K hits, more than 90% of which were from the last five years. We screened the results for articles where the title or abstract suggested that multiple source and/or target data sets have been used. In

addition, we screened the references of recent surveys [7,22] and screened the references and citations of the articles cited within this review. All selected articles were published between 2014 and 2018. A key change that happened in 2014 is being able to transfer from nonmedical data sets, which often yielded good results and gave this area of research a boost.

Our goal is to get insights into what type of considerations should be made when choosing a source data set for transfer learning. We first review the articles that compare different source data (Section Comparisons of source data sets) and provide a summary of publicly available source data sets (Section Public source data sets). We then discuss several gaps in current literature and opportunities for future research in Section Discussion.

Comparisons of source data sets

In this section, we discuss the articles that provide insights into using nonmedical or medical data sets for transfer learning. The articles are sorted by year and then alphabetically.

The study by Schlegl et al. [31] is the earliest reference we are aware of doing transfer learning from nonmedical data. The application is five-class classification of abnormalities in 2D slices of chest computed tomography (CT) images. They pretrain an unsupervised convolutional restricted Boltzmann machine on different source data sets with 20 K patches and fine-tune an entire convolutional neural network (CNN) with varying sizes of lung patches. The target data are from 380 chest CT scans of the Lung Tissue Research Consortium (LTRC) data set [3]. The source data include chest CT scans from LTRC, chest CT scans from a private data set, brain CT scans from a private data set, and natural images from the STL-10 data set [11], a subset of ImageNet. Natural images performed comparably or even slightly better than using only lung images. Brain images were less effective, possibly because of large homogeneous areas present in the scans not present in more texture-rich lung scans.

The study by Tajbakhsh et al. [35] addresses four different applications: polyp detection in colonoscopy, image quality assessment in colonoscopy, pulmonary embolism detection in CT, and intima-media boundary segmentation in ultrasonography. The authors investigate full training and fine-tuning in a layerwise manner with AlexNet pretrained on ImageNet. Overall, they observe that fine-tuning only the last layers performed worse than full training, but fine-tuning more layers was comparable to, or outperformed, full training. Fine-tuning more layers was especially important for polyp detection and intima-media boundary segmentation, which the authors hypothesize are less similar to ImageNet than the other applications they examined.

The study by Shin et al. [34] addresses two tasks: thoracoabdominal lymph node detection and interstitial lung disease classification. CIFAR-10 [18] and ImageNet are used as source data. Three strategies are compared: training from-scratch, off-the-shelf, and fine-tuning strategies for different networks: CifarNet (trained on CIFAR-10), AlexNet (trained on ImageNet), and GoogLeNet. CifarNet is used only with the off-the-shelf strategy, AlexNet with all three, and GoogLeNet only with from-scratch and fine-tuning strategy. For lymph node detection, the off-the-shelf strategy gives the worst results, but CifarNet outperforms AlexNet. Full training and fine-tuning lead to the best results, with fine-tuning being most beneficial for GoogLeNet. For interstitial lung disease classification, AlexNet achieves similar performance with all three strategies, and for GoogLeNet, fine-tuning is the most beneficial.

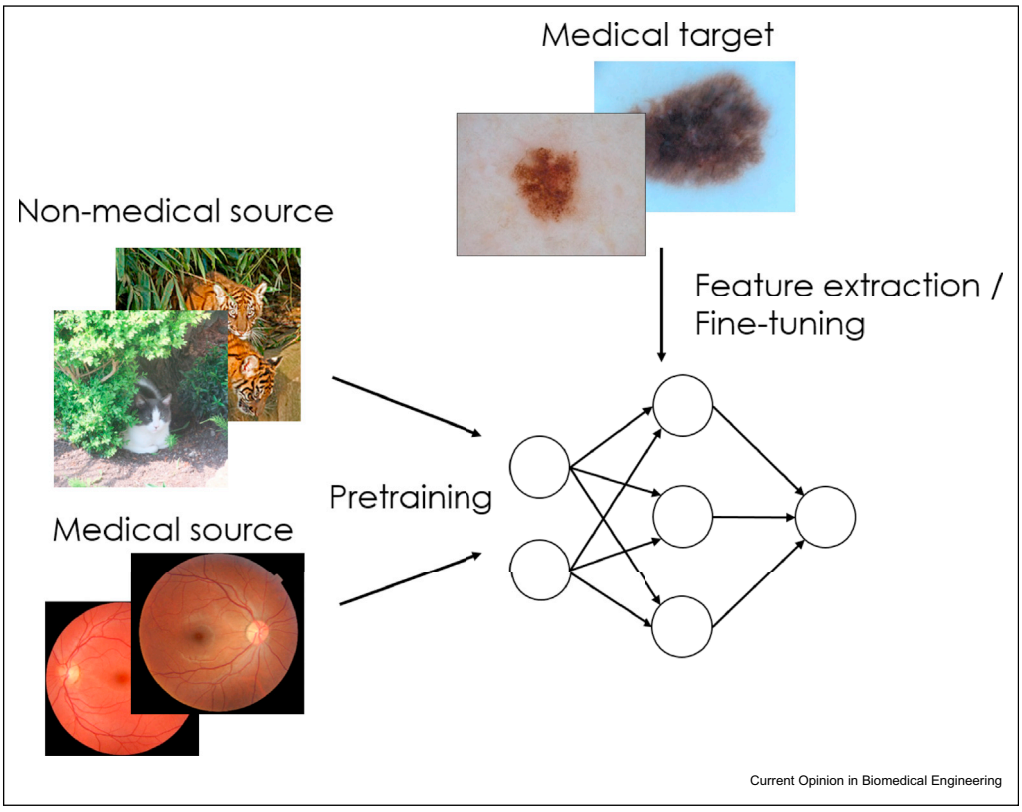
The study by Zhang et al. [38] addresses detection and classification of colorectal polyps in endoscopy images. They pretrain an eight-layer CNN and use the lower layers to extract features from the target data, which are then classified with a support vector machine. ImageNet and Places [39] are used assource data sets. Target data sets include a private endoscopy data set with 2 K images in three classes and a public endoscopy video data set [26] from which 332 images in three classes are extracted. They hypothesize that Places has higher similarity between classes than ImageNet, which would help distinguish small differences in polyps. This indeed leads to higher recognition rates, also while varying other parameters of the classifier.

The study by Cha et al. [5] predicts the response to cancer treatment in the bladder of 82 patients using a five-layer CNN. They compare networks without transfer learning to two other source data sets: 60 K natural images from CIFAR-10 and 160 K bladder regions of interest (ROIs) from 81 patients from a previous study. The experiments show no statistically significant differences in the area under the curve (AUC) values of twofold cross-validation using these strategies.

The study by Christodoulidis et al. [8] addresses classification of interstitial lung disease in patches of CT images. Six public texture source data sets are used for training a seven-layer network on each data set and combining the networks in an ensemble. Individually, the source data sets result in networks with comparable performance, but the performance varies a lot depending on the number of layers transferred. The ensemble outperforms the individual networks. The ensemble also outperforms a network trained on the union of the data sets.

The study by Menegola et al. [25] addresses melanoma classification in skin lesion images. Source data

Figure 1



Transfer learning from nonmedical or medical image data sets. A network is first trained on a source data set. This network can then be used for feature extraction or further training on the medical target data.

consist of ImageNet and Kaggle diabetic retinopathy (DR) [16]. The authors compare off-the-shelf, full-training, and fine-tuning strategies for a visual geometry group (VGG) network. They also investigate 'double transfer': fine-tuning the pretrained ImageNet model on KaggleDR and only then on the target task. Fine-tuning outperforms off-the-shelf features when transferring from both sources. When transferring from ImageNet, off-the-shelf features outperform full training, but when transferring from KaggleDR, off-the-shelf features perform comparably with full training. Double transfer performs worse than transfer from ImageNet alone. This is in contrast to the hypothesis of the authors that KaggleDR will lead to best results because of the visual similarity of the data.

The study by Ribeiro et al. [29] investigates pretraining and fine-tuning with nine different source data sets (natural images, texture images, and endoscopy images) for classification of polyps in endoscopy images. Different from most other articles, they extract data sets of the same number of classes and images from the available types of data for pretraining. The experiments show that texture data sets perform best as source data, but if the size of the source data set is small, it is better to select a larger unrelated source data set.

The study by Shi et al. [33] addresses prediction of occult invasive disease in ductal carcinoma in situ in mammography images of 140 patients. Three public data sets are used as the source data: ImageNet, texture data set DTD [9], and data set of mammography images INbreast [27]. The authors pretrain a 16-layer VGG network, extract off-the-shelf features from the target data using different network layers, and train a logistic regression classifier. They hypothesize that INbreast is most similar to the target data and will lead to the best results (and conversely, the least similar ImageNet will lead to the worst results) and report that the average AUCs are consistent with this hypothesis.

The study by Du et al. [13] addresses classification of 15 K epithelium and stroma ROIs in 158 digital pathology images. ImageNet and Places are used as the source data. They extract off-the-shelf features from different layers of several architectures, where only AlexNet is trained on both sources. Comparing the

AUCs of the AlexNet trained on ImageNet and Places, the layer used to extract the features (lower layers are better) has more influence than which data are used for pretraining.

The study by Mormont et al. [28] focuses on tissue classification. They argue that experiments are often carried out on a single data set; therefore, as target data, eight-tissue classification data sets with 1 K–30 K images and two to ten classes are used. Seven architectures that are all trained on ImageNet are compared. The method consists of extracting features off-the-shelf or after fine-tuning and training a supervised classifier. The results show that fine-tuning usually outperforms the other methods for any network, especially for multiclass data sets. The last layer is never the best for feature extraction, possibly because the features are very specific for natural images. Furthermore, the results do not suggest that larger data sets necessarily lead to better results — the smallest and the largest data sets lead to the best performances equally often.

The study by Lei et al. [21] addresses HEp-2 cell classification in the ICPR 2016 challenge as the target task [23]. Compared models include a ResNet pretrained on ImageNet and a ResNet pretrained on data from the earlier edition of the challenge, ICPR 2012 [15]. The authors hypothesize that pretraining on ICPR 2012 will lead to similar feature representations both in the lower and higher layers and show that the network pretrained on ICPR 2012 data outperforms the ImageNet network.

The study by Wong et al. [36] focuses on two tasks: three-class classification of brain tumors in 3D magnetic resonance images and nine-class classification in 2D cardiac computed tomography angiography (CTA) images. They argue that pretrained ImageNet models are not suitable for medical target tasks because of unnecessary resizing of images, too large number of classes, and the absence of 3D information. The classifier is a modified U-Net which is first trained on a segmentation task on the same data, using either manual segmentations or segmentations generated with a simple thresholding method. In tumor classification, where ImageNet is not tested because of the 3D nature of the images, pretraining both with manual and thresholded segmentations outperforms training a network from scratch. In cardiac image classification, pretraining with manual segmentations gives the best results. Pretraining on ImageNet outperforms pretraining on thresholded segmentations. Pretraining on ImageNet also outperforms training from scratch but only for low training sizes.

Public source data sets

A list of publicly available source data sets used in articles comparing multiple sources, but focusing on medical target tasks, is presented in Table 1. ImageNet is a popular choice, although some articles use other object recognition data sets such as CIFAR-10. Several articles use texture data sets, of which a variety is available. Only a few medical source data sets are listed, often because a private medical data set is used.

Table 1

Overview of public data sets used as source data. In several cases, the sample size is larger in cases when networks are trained on patches extracted from images, rather than entire images.

Data	Type	Images	Classes	Used in
ImageNet [30]	Object recognition	1.2 M	1 K	[35] [34] [25] [38] [13] [28] [21] [36]
STL-10 [11]	Object recognition	100 K	10	[31]
Places [38]	Scene recognition	2.5M	205	[38] [13]
DTD [9]	Texture classification	5.6 K	47	[29] [33] [8]
ALOT [4]	Texture classification	28 K	250	[29] [8]
KTH-TIPS [12]	Texture classification	810	10	[29] [8]
CIFAR-10 [18]	Object classification	60 K	10	[34] [5]
FMD [32]	Texture classification	1 K	10	[8]
KTB [19]	Texture classification	4.5 K	27	[8]
UIUC [20]	Texture classification	1 K	25	[8]
CALTECH-101 [14]	Object recognition	9 K	101	[29]
COREL-1000	Scene recognition	1 K	10	[29]
KaggleDR [16]	Retinal image classification	35 K	2	[25]
INbreast [27]	Breast lesion classification	410	2	[33]
ICPR 2012 [15]	HEp-2 cell classification	1.5 K	6	[21]

Discussion

We have summarized several articles that use medical or nonmedical data as source data and apply the classifier on medical target data. A limitation is that as such articles are difficult to discover — other than 'transfer learning', which returns over 2 K results when combined with 'medical imaging' on Google Scholar, we have not been able to find keywords that identify when different source of data sets have been used. We encourage readers to notify us of any articles that were not included but investigate this phenomenon.

Summary of results

The results of the comparisons point in different directions. We now group the articles into 'nonmedical is best', 'medical wins', 'no difference', and 'inconclusive'.

Out of twelve articles mentioned in Section Comparisons of source data sets, three articles conclude that natural images outperform medical images. Menegola et al [25] and Schlegl et al. [31] find natural images (STL-10 and ImageNet respectively) more effective than medical [29]. have most success with texture images, compared to natural or medical images. We consider these texture images as nonmedical.

Five articles can be seen as voting for medical images. Three articles [21,33,36] have clear conclusions that better results are achieved with medical source data. Next to this, there are two 'minor' votes for medical images. Tajbakhsh et al [35] only use ImageNet as source data, but they use different target data sets, and conclude that ImageNet is worse for data sets that are less similar to medical images. This could be seen as a vote for medical data sets because these would be most similar to the target data. A similar intuition holds for [38] who use ImageNet and Places as the source data and find that Places is better because of its higher similarity to medical data.

In two articles, there are no clear differences between source data sets. Cha et al. [5] find no significant differences between using different (medical and nonmedical) sources. Du et al. [13] do not demonstrate consistent differences between ImageNet and Places as source data.

From three of the articles, we cannot reason whether nonmedical or medical is better because of the source choices, but they do provide other relevant insights. Shin et al. [34] use two natural data sets as sources and show that a larger natural image data set is not necessarily better. The same holds for Christodoulidis et al. [8] who use only texture data sets as sources but show that training on the union of the data sets does not improve the results. Similarly, Mormont et al. [28] show that both smaller and larger medical sources can be successful as source data.

Limitations

It is difficult to compare these results directly because of differences in how transfer learning is implemented. Examples of variation include the subset of the source data that are used, the architecture of the network, how the transfer was implemented, both in terms of strategy (off-the-shelf or fine-tuning), and which layers were used for the transfer. Furthermore, because we could only find 12 relevant comparisons, a difference of several votes for one or the other approach could be due to chance.

Another issue is that the target data sets in medical imaging can be very small, and it is not clear if the results would generalize to another similar data set. Methods are sometimes compared by looking only at a single run of each method, or at an average over multiple runs, but without considering possible variability in such performances. A recent article comparing medical image challenges [24] shows that in such conditions, rankings of algorithms can easily change, for example, if a slightly different metric is used. Most articles we surveyed performed no statistical significance tests — if this was the case, perhaps the conclusions would be different.

Opportunities for further research

There are opportunities in doing more systematic comparisons. One direction is to use more of the available data sets, both from the nonmedical and medical domains. It would be informative to vary the number of images and number of classes in the data, as in the study by Ribeiro et al. [29]. Also of interest would be comparing different tasks, such as segmentation and classification, involving the same images, as in the study by Wong et al. [36].

The number of public medical source data sets is rather low. A strategy that could be helpful to counteract this, but seems underexplored, is unsupervised pretraining. This would allow the use of larger unlabeled medical data sets, which may be only weakly labeled. Many such data sets are already publicly available, for example, on grand-challenge.org. Another way to increase the number of source data sets would be to share pretrained models, which would also allow transfer learning from private data sets, without sharing the data itself. The feasibility of this approach with respect to privacy and intellectual property of the data would need to be investigated first.

Similarity of data sets is often used to hypothesize about which source data will be best, but definitions of similarity differ. For example, Menegola et al. [25] discuss similarity in terms of visual similarities of the images, and Lei et al. [21] discuss similarity in terms of feature representations. In computer vision, other definitions may be used — for example, Azizpour et al. [1] investigate transfer from ImageNet and Places to 15 other data sets and define similarity in terms of the number and variety of the classes. Given a definition of similarity, it remains a question which data sets would be best to use for pretraining. Arguably, the most similar data set to the target data set is the target data set itself, which might not add any additional information. Investigating how to represent data sets in a feature space (one example can be found in the study by Cheplygina et al. [6]) or how to directly define data set similarity is an important point of investigation.

Instead of considering only the similarity of the source data, perhaps the diversity of the source data is also an important factor. Instead of selecting one data set as the source, it might be a good strategy to use an ensemble,

as in the study by Christodoulidis et al. [8]. It is in fact surprising, given that top performing methods in challenges are often ensembles, that this strategy was not investigated in the articles we reviewed. It might be the case that it is not better to use nonmedical or medical data — the answer to the question posed by the title might simply be 'both'.

Concluding remarks

In conclusion, we looked at 12 articles that compare various source and/or target data sets from different domains. A similar number of articles found that nonmedical or medical data were better, with a slight advantage for the medical data. Several articles showed that larger data sets are not necessarily better. As data sets, data set sizes, architectures, and transfer strategies vary between comparisons and results are often based on a few data sets without significance testing, we urge the community to conduct larger systematic comparisons into this important topic.

Conflict of interest

Nothing declared.

References

1. Azizpour H, Sharif Razavian A, Sullivan J, Maki A, Carlsson S: **From generic to specific deep representations for visual recognition**. In *computer vision and pattern recognition workshops (CVPR-W)*; 2015:36–45.
2. Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H: **Chest pathology detection using deep learning with non-medical training**. In *International symposium on biomedical imaging (ISBI)*. IEEE; 2015:294–297.
3. Bartholmai B, Karwoski R, Zavaletta V, Robb R, Holmes DRI: **The lung tissue research consortium: an extensive open database containing histological, clinical, and radiological data to study chronic lung disease**. In *Insight journal: 2006 MICCAI open science workshop*; 2006.
4. Burghouts GJ, Geusebroek J-M: **Material-specific adaptation of color invariant features**. *Pattern Recogn Lett* 2009, **30**:306–313.
5. Cha KH, Hadjiiski LM, Chan H-P, Samala RK, Cohan RH, Caoili EM, Paramagul C, Alva A, Weizer AZ: **Bladder cancer treatment response assessment using deep learning in CT with transfer learning**. In *SPIE medical imaging*. International Society for Optics and Photonics; 2017. 1013404–1013404.
6. Cheplygina V, Moeskops P, Veta M, Bozorg BD, Pluim J: **Exploring the similarity of medical imaging classification problems**. In *Intravascular imaging and computer assisted stenting, and large-scale Annotation of biomedical data and expert label synthesis (MICCAI LABELS)*; 2017. https://link.springer.com/chapter/10.1007/978-3-319-67534-3_7.
7. Cheplygina V, de Bruijne M, Pluim JP: *Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis*. 2018. arXiv preprint arXiv:1804.06353.
8. Christodoulidis S, Anthimopoulos M, Ebner L, Christe A, Mougiakakou S: **Multisource transfer learning with convolutional neural networks for lung pattern analysis**. *IEEE Journal of Biomedical and Health Informatics* 2017, **21**:76–84.
9. Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A: **Describing textures in the wild**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2014:3606–3613.
10. Ciompi F, de Hoop B, van Riel SJ, Chung K, Scholten ET, Oudkerk M, de Jong PA, Prokop M, van Ginneken B: **Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box**. *Med Image Anal* 2015, **26**:195–202.
11. Coates A, Ng A, Lee H: **An analysis of single-layer networks in unsupervised feature learning**. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*; 2011:215–223.
12. Dana KJ, Van Ginneken B, Nayar SK, Koenderink JJ: **Reflectance and texture of real-world surfaces**. *ACM Trans Graph* 1999, **18**:1–34.
13. Du Y, Zhang R, Zargari A, Thai TC, Gunderson CC, Moxley KM, Liu H, Zheng B, Qiu Y: **A performance comparison of low-and high-level features learned by deep convolutional neural networks in epithelium and stroma classification**. In *Medical imaging 2018: digital pathology*, vol. 10581. International Society for Optics and Photonics; 2018: 1058116.
14. Fei-Fei L, Fergus R, Perona P: **One-shot learning of object categories**. *IEEE Trans Pattern Anal Mach Intell* 2006, **28**: 594–611.
15. Foggia P, Percannella G, Soda P, Vento M: **Benchmarking HEp-2 cells classification methods**. *IEEE Trans Med Imag* 2013, **32**: 1878–1889.
16. Graham B: *Kaggle diabetic retinopathy detection competition report*. University of Warwick; 2015.
17. Greenspan H, Van Ginneken B, Summers RM: **Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique**. *IEEE Trans Med Imag* 2016, **35**:1153–1159.
18. Krizhevsky A, Hinton G: *Learning multiple layers of features from tiny images*. Technical report. Citeseer; 2009.
19. Kylberg G: *Kylberg texture dataset v. 1.0*. Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University; 2011.
20. Lazebnik S, Schmid C, Ponce J: **A sparse texture representation using local affine regions**. *IEEE Trans Pattern Anal Mach Intell* 2005, **27**:1265–1278.
21. Lei H, Han T, Zhou F, Yu Z, Qin J, Elazab A, Lei B: **A deeply supervised residual network for hep-2 cell classification via cross-modal transfer learning**. *Pattern Recogn* 2018, **79**: 290–302.
22. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JA, Van Ginneken B, Sánchez CI: **A survey on deep learning in medical image analysis**. *Med Image Anal* 2017, **42**:60–88.
23. Lovell BC, Percannella G, Saggese A, Vento M, Wiliem A: **International contest on pattern recognition techniques for indirect immunofluorescence images analysis**. In *Pattern recognition (ICPR), 2016 23rd international conference on*. IEEE; 2016:74–76.
24. Maier-Hein L, Eisenmann M, Reinke A, Onogur S, Stankovic M, Scholz P, Arbel T, Bogunovic H, Bradley AP, Carass A, *et al.*: **Is the winner really the best? a critical analysis of common research practice in biomedical image analysis competitions**. *Nat Commun* 2018, **9**:5217. https://doi.org/10.1038/s41467-018-0761.
25. Menegola A, Fornaciali M, Pires R, Bittencourt FV, Avila S, Valle E: **Knowledge transfer for melanoma screening with deep learning**. In *International sympsium on biomedical imaging (ISBI)*. IEEE; 2017:297–300.
26. Mesejo P, Pizarro D, Abergel A, Rouquette O, Beorchia S, Poincloux L, Bartoli A: **Computer-aided classification of gastrointestinal lesions in regular colonoscopy**. *IEEE Trans Med Imag* 2016, **35**:2051–2063.
27. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS: **Inbreast: toward a full-field digital mammographic database**. *Acad Radiol* 2012, **19**:236–248.
28. Mormont R, Geurts P, Marée R: **Comparison of deep transfer learning strategies for digital pathology**. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition workshops; 2018:2262–2271.

29. Ribeiro E, Häfner M, Wimmer G, Tamaki T, Tischendorf J, Yoshida S, Tanaka S, Uhl A: **Exploring texture transfer learning for colonic polyp classification via convolutional neural networks**. In *International symposium on biomedical imaging (ISBI)*. IEEE; 2017:1044–1048.

30. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, *et al.*: **Imagenet large scale visual recognition challenge**. *Int J Comput Vis* 2015, **115**:211–252.

31. Schlegl T, Ofner J, Langs G: **Unsupervised pre-training across image domains improves lung tissue classification**. In *Medical computer vision: algorithms for big data (MICCAI MCV)*. Springer; 2014:82–93.

32. Sharan L, Rosenholtz R, Adelson E: **Material perception: what can you see in a brief glance?** *J Vis* 2009, **9**. 784–784.

33. Shi B, Hou R, Mazurowski MA, Grimm LJ, Ren Y, Marks JR, King LM, Maley CC, Hwang ES, Lo JY: **Learning better deep features for the prediction of occult invasive disease in ductal carcinoma in situ through transfer learning**. In *Medical imaging 2018: computer-aided diagnosis*, vol. 10575. International Society for Optics and Photonics; 2018. 105752R.

34. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM: **Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning**. *IEEE Trans Med Imag* 2016, **35**:1285–1298.

35. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J: **Convolutional neural networks for medical image analysis: full training or fine tuning?** *IEEE Trans Med Imag* 2016, **35**:1299–1312.

36. Wong KC, Syeda-Mahmood T, Moradi M: **Building medical image classifiers with very limited data using segmentation networks**. *Med Image Anal* 2018, **49**:105–116.

37. Yamashita R, Nishio M, Do RKG, Togashi K: **Convolutional neural networks: an overview and application in radiology**. *Insights Imaging* 2018:1–19.

38. Zhang R, Zheng Y, Mak TWC, Yu R, Wong SH, Lau JY, Poon CC: **Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain**. *IEEE J Biomed Health Inf* 2017, **21**:41–47.

39. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A: **Places: a 10 million image database for scene recognition**. *IEEE Trans Pattern Anal Mach Intell* 2017, **40**(6):1452–1464.