Medical Image Analysis 54 (2019) 280–296

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis

Veronika Cheplygina[b,*], Marleen de Bruijne[a,d], Josien P.W. Pluim[b,c]

[a] Biomedical Imaging Group Rotterdam, Departments Radiology and Medical Informatics, Erasmus Medical Center, Rotterdam, the Netherlands
[b] Medical Image Analysis, Department Biomedical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands
[c] Image Sciences Institute, University Medical Center Utrecht, Utrecht, the Netherlands
[d] The Image Section, Department Computer Science, University of Copenhagen, Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 24 September 2018
Revised 20 December 2018
Accepted 25 March 2019
Available online 29 March 2019

Keywords:
Machine learning
Medical imaging
Computer aided diagnosis
Semi-supervised learning
Weakly-supervised learning
Multiple instance learning
Transfer learning
Multi-task learning

ABSTRACT

Machine learning (ML) algorithms have made a tremendous impact in the field of medical imaging. While medical imaging datasets have been growing in size, a challenge for supervised ML algorithms that is frequently mentioned is the lack of annotated data. As a result, various methods that can learn with less/other types of supervision, have been proposed. We give an overview of semi-supervised, multiple instance, and transfer learning in medical imaging, both in diagnosis or segmentation tasks. We also discuss connections between these learning scenarios, and opportunities for future research. A dataset with the details of the surveyed papers is available via https://figshare.com/articles/Database_of_surveyed_literature_in_Not-so-supervised_a_survey_of_semi-supervised_multi-instance_and_transfer_learning_in_medical_image_analysis_/7479416.

1. Introduction

Machine learning has become very important in medical image analysis. Tasks such as segmentation, where each pixel or voxel in an image is assigned to a different anatomical structure or tissue type, and computer-aided diagnosis where a category label or a continuous value is predicted for an entire image, are now almost exclusively done with machine learning methods.

A frequent problem when applying machine learning methods to medical images, is the lack of labeled data (Litjens et al., 2017; Weese and Lorenz, 2016; de Bruijne, 2016), even when larger sets of unlabeled data may be more widely available. An important reason for this is the sheer difficulty of collecting the labels. Manual labeling of the images is an expensive and/or time-consuming process. Such labels might not be needed in clinical practice, therefore reducing the availability of labeled data only to research studies. Another issue is that, even when labeled data is collected, it is not often available to other researchers.

The lack of labeled data motivates approaches that go beyond traditional supervised learning by incorporating other data and/or labels that might be available. These approaches include semi-supervised learning, multiple instance learning and transfer learning, although many other terms exist to describe these approaches. Papers within one of these learning scenarios seem to be aware of other related literature, and surveys are emerging, such as Quellec et al. (2017). However, it seems that there is little interaction between the scenarios, which is a missed opportunity, since their goals are related.

With this survey, we aim to provide an overview of the learning scenarios, describe their connections, identify gaps in the current approaches, and provide several opportunities for future research. The survey is primarily aimed at researchers in medical image analysis. We have however made an effort for the survey to also be accessible to a broader readership.

1.1. Selection of papers

An initial selection of papers was created by screening the results of Google Scholar searches for terms “semi-supervised learning”, “multiple instance learning” and “transfer learning” for medical imaging papers. These papers were used to identify other relevant publications. In the event of multiple similar papers, only the latest paper was included. Only papers that became available online before 2018 were included. After publishing the preprint

* Corresponding author.
E-mail address: v.cheplygina@tue.nl (V. Cheplygina).

https://doi.org/10.1016/j.media.2019.03.009
1361-8415/

Table 1
Notation used throughout the paper.

Feature space	$\mathcal{X}$
Label space	$\mathcal{Y}$
Classifier	$f : \mathcal{X} \rightarrow \mathcal{Y}$
Instance	$\mathbf{x}_i \in \mathcal{X}$
Instance label	$y_i \in \mathcal{Y}$
Bag	$X_i = \{\mathbf{x}_{ij}, j = 1, \ldots, N_i\}$
Bag label	$Y_i \in \mathcal{Y}$
Domain	$\{\mathcal{X}, p(\mathbf{x})\}$
Domain (MIL)	$\{\mathcal{X}, p(X)\}$
Task	$\{\mathcal{Y}, f(\cdot)\}$
Subscript $_S$	source
Subscript $_T$	target
Source domain	$\mathcal{D}_S$
Target domain	$\mathcal{D}_T$
Source task	$\mathcal{T}_S$
Target task	$\mathcal{T}_T$
Training (source) data	D_S
Test (target) data	D_T
Unlabeled (source) data	U
Labeled (target) data	L

Table 2
Acronyms used throughout the paper.

ML	machine learning
SSL	semi-supervised learning
MIL	multiple instance learning
TL	transfer learning
MTL	multi-task learning
SVM	support vector machine
AD	Alzheimer's disease
MCI	mild cognitive impairment
COPD	chronic obstructive pulmonary disease
CT	computed tomography
DR	diabetic retinopathy
MR	magnetic resonance
US	ultrasound

online, we received more suggestions for relevant papers, and included these if these fit our criteria.

This survey does not cover approaches that rely on interaction with the annotator, such as active learning or crowdsourcing, in detail. We focus on machine learning approaches that can be used even if there is no possibility of acquiring additional labels. We focus on classification tasks within medical image analysis, for diagnosis, detection or segmentation purposes.

It is our intention to provide an overview how different learning scenarios are being used rather than a full summary of all related papers. We emphasize that we focus on the types of learning scenarios and the assumptions that are being made, rather than specific classifiers. As such we do not explicitly address the machine/deep learning distinction, as both types of methods can use similar assumptions. A recent survey focusing on deep learning in medical imaging can be found in Litjens et al. (2017).

2. Overview of techniques

In this section we provide a quick overview of the learning scenarios. We also provide examples of each type of learning scenario, based on the application of classifying emphysema, a sign of chronic obstructive pulmonary disease (COPD) in chest computed tomography (CT) images. For readability, we provide a list of notation and acronyms that will be introduced throughout the paper in Tables 1 and 2.

In supervised learning, we have a training set $D_S = \{(\mathbf{x}_i y_i\}$, where each sample $\mathbf{x}_i \in \mathcal{X}$ is associated with a label $y_i \in \mathcal{Y}$. Here $\mathcal{X}$ is the feature space, such as an m-dimensional space of real numbers $\mathbb{R}^m$, and $\mathcal{Y}$ is the label space, such as the set {0, 1} for a binary classification problem or the set $\mathbb{R}$ for a regression problem. We want to use this data to train a classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ that can provide labels for unlabeled samples from a previously unseen test set D_T. For example, the instances are patches in a chest CT image, and they are labeled as emphysema or as normal tissue. At test time, we want to classify all patches in a previously unseen scan as emphysema or not. This example is illustrated in Fig. 1(a).

In *semi-supervised learning* (SSL), in addition to the training set we have an unlabeled set of data U. We want to use this set to improve the predictions of the classifier on D_T. For example, the supervised problem above can be extended with patches from chest CT images that have not been manually labeled by experts. This scenario is covered in Section 3 and illustrated in Fig. 1(b).

In *multiple instance learning* (MIL), the training set itself consists of *bags* of instances $X_i = \{\mathbf{x}_{ij}, j = 1, \ldots, N_i\}$. The bags are labeled. The instances have unknown labels y_{ij} that are somehow related to the bag label Y_i. An example of such a relationship is "if at least one instance is positive, the bag is also positive". For example, this situation can occur if the radiologist only labeled an entire CT scan as containing emphysema or not, but has not indicated its locations.

In MIL the test set also consists of bags, which are unlabeled. Next the goal can be two-fold: to classify the test bags and/or to classify the test instances. In our example, the bag classifier would predict whether a patient has any emphysema, whereas the instance classifier would in addition localize any emphysema in the image. This scenario is covered in Section 4 and illustrated in Fig. 1(c).

In the scenarios above, we assume that the training and test data are from the same *domain* $\mathcal{D} = (\mathcal{X}, p(\mathbf{x}))$, defined by the feature space and distribution of the samples. However, this is not always the case, creating a *transfer learning* (TL) scenario. In this scenario we assume to have a source dataset D_S where the instances $\mathbf{x}_{S_i} \in \mathcal{X}_S$ and a test or target set D_T where the instances $\mathbf{x}_{T_i} \in \mathcal{X}_T$. For example, this can occur when different scanning protocols are used, leading to different appearance of the patches, and therefore different distributions $p(\mathbf{x}_S)$ and $p(\mathbf{x}_T)$. The goal is to train a classifier using D_S, and possibly using either the unlabeled test data D_T, and/or labeled data from the target domain $L \in \mathcal{D}_T$. This scenario is covered in Section 5 and illustrated in Fig. 1(d). As we discuss later, this scenario is not limited to the case where there are differences in the feature distributions.

In Section 6 we discuss the trends within these learning scenarios, the gaps in the current research, and the opportunities and challenges for future research.

3. Semi-supervised learning

In the semi-supervised learning scenario, there are two sets of samples: labeled samples D_S and unlabeled samples U. The goal is to use the samples in U to improve the classifier f, where f is constructed only using samples in D_S. For example, when classifying emphysema vs normal patches, the scans that have been annotated are used to create a set of labeled patches, while the scans without annotations can be used to create a large unlabeled set of patches. We can distinguish two goals in SSL: predicting labels for future data (inductive SSL) and predicting labels for the already available unlabeled samples (transductive SSL) (Zhu and Goldberg, 2009).

Typically semi-supervised approaches work by making additional assumptions that link properties of the distribution of the input features to properties of the decision function (Chapelle et al., 2006; Zhu and Goldberg, 2009). These include the *smoothness assumption*, i.e. samples close together in feature space are likely to be from the same class, the *cluster assumption*, i.e. samples in a cluster are likely to be from the same class, and the *low density assumption*, i.e. class boundaries are likely to be in areas of the feature space that have lower density than the clusters.

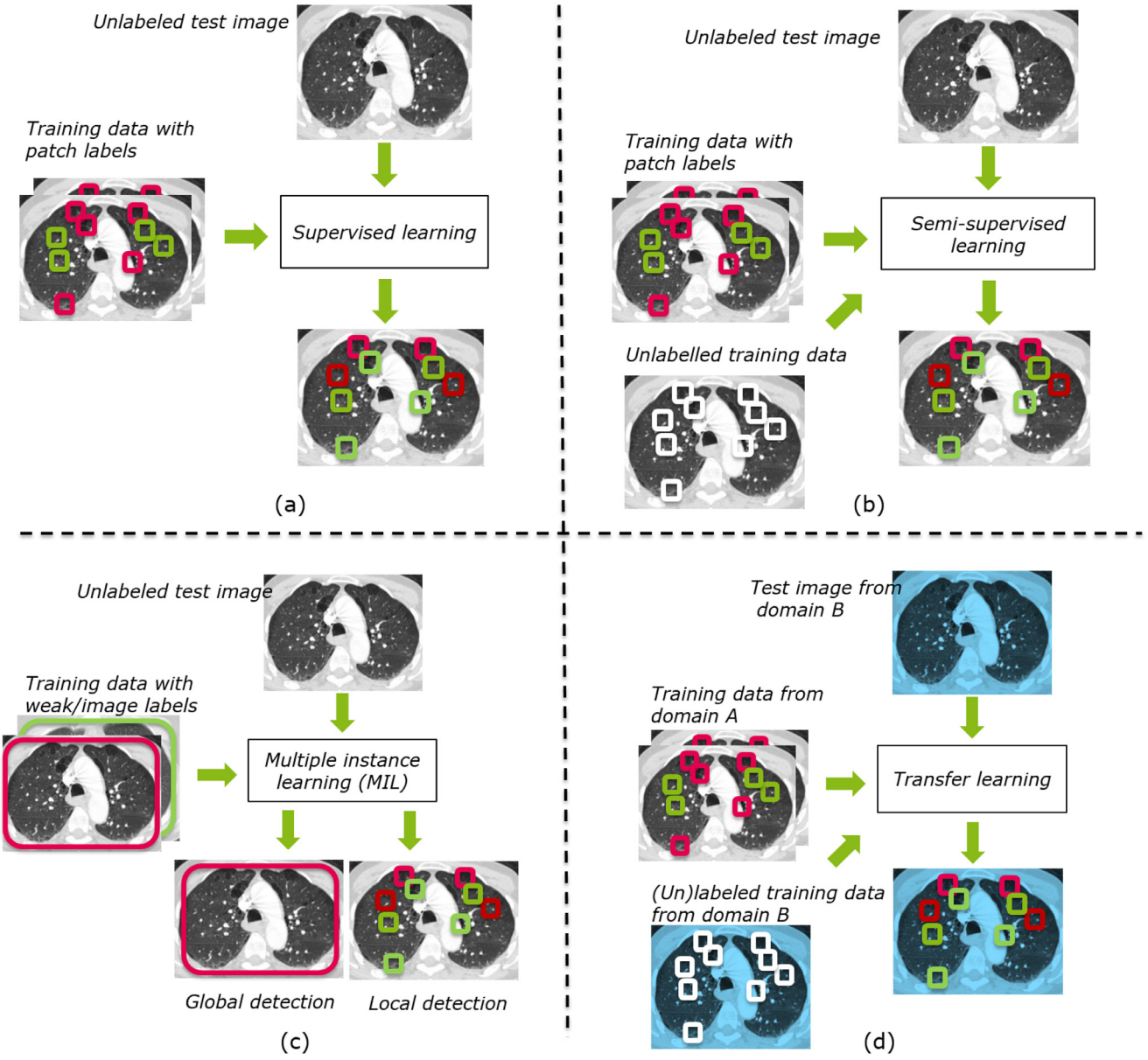

Fig. 1. Learning scenarios, illustrated by a task of classifying healthy (green) vs emphysema (red) tissue in chest CT images. Annotations are made for presentation purposes only and do not necessarily reflect ground truth. (a) Supervised learning, only healthy and abnormal patches are available. (b) Semi-supervised learning (Section 3). In addition to healthy and abnormal patches, unlabeled patches are available. (c) Multiple instance learning (Section 4). Labeled patches are not available, but subject-level labels (whether any abnormal patches are present) are. (d) Transfer learning (Section 5). Labeled patches are available, but for a different domain (here illustrated by different visual characteristics) than in the test set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Many semi-supervised approaches therefore proceed with exploiting such assumptions. A popular method called self-training propagates labels from the labeled to the unlabeled data, and then using the larger, newly labeled set for training. This approach assumes that the method's high confidence predictions are correct, which is likely to be the case with the cluster assumption (Zhu and Goldberg, 2009). Expectation-maximization (Dempster et al., 1977) uses a principle similar to self-training by alternating between assigning soft labels to the unlabeled data given the labeled data and model parameters, and updating the model parameters given all the data. Another related approach is co-training (Blum and Mitchell, 1998), where classifiers are trained with independent sets of features, and the classifiers rely on each other for estimating the confidence of their predictions.

Other popular methods include methods that regularize the classifier based on the unlabeled data, such as graph-based methods and semi-supervised support vector machines (SVMs). Graph-based methods encode similar samples as connected nodes and solve a graph cut problem, therefore assuming low density between classes. Semi-supervised SVMs encourage margins that place unlabeled data outside the margin, also assuming low density separation. An overview of methods and corresponding assumptions can be found in Zhu and Goldberg (2009).

When the additional assumptions do not hold, there is a risk of performing worse than a supervised approach (Cozman and Cohen, 2006; Zhu and Goldberg, 2009). More recent are approaches that do not make additional assumptions about the data, and instead use assumptions already present in the classifier (Loog and Jensen, 2015; Krijthe and Loog, 2017). For example, this can be achieved by linking parameter estimates (such as mean and variance of the samples) based on labeled samples, to those based on all available samples.

Table 3
Overview of semi-supervised learning applications. The last column describes the type of method used, "active" refers to active learning.

Reference	Application	SSL category
Brain		
Song et al. (2009)	tumor segmentation	graph-based
Iglesias et al. (2010)	skull stripping	self-training
Filipovych et al. (2011)	classification of MCI	semi-supervised SVM
Batmanghelich et al. (2011)	classification of AD, MCI	graph-based
Xie et al. (2013)	tissue segmentation	graph-based
Meier et al. (2014)	tumor segmentation	graph-based
Dittrich et al. (2014)	fetal brain segmentation	self-training
Wang et al. (2014)	lesion segmentation	self-training, active
An et al. (2016)	AD classification	graph-based
Baur et al. (2017)	MS lesion segmentation	graph-based
Moradi et al. (2015)	classification of MCI	semi-supervised SVM
Eye		
Adal et al. (2014)	microaneurysm detection	self-training
Mahapatra (2016)	optic disc missing annotation prediction	self-training, graph-based
Breast		
Sun et al. (2016)	mass classification	co-training
Heart		
Zuluaga et al. (2011)	detection of vascular lesions	self-training
Bai et al. (2017)	cardiac segmentation	self-training
Wang et al. (2017)	aneurysm volume estimation	graph-based
Lung		
Prasad et al. (2009)	segmentation of emphysema in CT	self-training/co-training, active
van Rikxoort et al. (2010)	classification of tuberculosis patterns in CT	self-training/co-training
Abdomen		
Tiwari et al. (2010)	classification of cancerous areas in prostate	graph-based
Park et al. (2014)	prostate segmentation	graph-based, active
Borga et al. (2016)	liver segmentation	graph-based
Mahapatra (2016)	predicting missing expert annotations of Crohn's disease	self-training, graph-based
Histology and microscopy		
Singh et al. (2011)	cell type classification in microscopy	self-training
Parag et al. (2014)	cell type segmentation in microscopy	graph-based, active
Xu et al. (2016)	neuron segmentation in microscopy	graph-based
Su et al. (2016)	cell segmentation in microscopy	graph-based, active
Multiple		
Gass et al. (2012)	segmentation in two applications	graph-based
Ciurte et al. (2014)	segmentation in four applications	graph-based
Gu et al. (2017)	segmentation in two applications	self-training
Other		
Huang et al. (2008)	segmentation of nasopharyngeal carcinoma lesion in MR	graph-based

3.1. SSL In medical imaging

SSL is a naturally occurring scenario in medical imaging, both in segmentation and diagnosis tasks. In segmentation methods, an expert might label only a part of the image, leaving many samples unlabeled. In computer-aided diagnosis, there might be ambiguity about the label of a subject, so rather than adding these subjects to the training set or removing them completely, they can still be used without labels to improve the classifier. For example, in classification subjects as having Alzheimer's disease (AD) or normal cognitive function (CN), subjects with mild cognitive impairment (MCI) who may or may not develop AD later, are sometimes considered unlabeled (Filipovych et al., 2011; Batmanghelich et al., 2011).

The papers using SSL are summarized in Table 3. Overall there are two main directions. In the first, papers use a self-training or co-training approach for segmentation purposes. We discuss these in Section 3.2. In the second, papers use the unlabeled data to regularize the classifier via graph-based methods or SVMs. These approaches are used both for segmentation and diagnosis tasks. We discuss these papers in Section 3.3.

3.2. Self-training and co-training

A popular approach to SSL in medical imaging is label propagation via self-training. The general idea is as follows. A classifier is first trained on the labeled samples. The trained classifier then classifies the unlabeled samples. These samples, or a subset of these samples, are then added to the training set. This process is repeated several times.

The surveyed papers differ in how they select the subset of unlabeled samples to add to the labeled data. Several authors choose an active learning approach, where expert interaction is needed to verify some of the labels (Parag et al., 2014; Su et al., 2016). As mentioned in the introduction, we do not in detail address active learning, and only focus on methods that can be used even if no additional labels can be acquired.

Other authors measure the uncertainty or confidence of the classifier based on the output (posterior probability) of the classifier itself, and possibly additional classifiers. Wang et al. (2014) add samples with a confidence above a user-selected threshold to the training set. Additional classifiers can be used as well, in which case the method falls under co-training. For example, for skull

segmentation, Iglesias et al. (2010) use two conventional tools and their own classifier, to classify all the unlabeled pixels. The pixels for which the conventional tools agree, but their own classifier is not confident, are added to the labeled set. A similar strategy is used by van Rikxoort et al. (2010) for classifying tuberculosis patterns in chest CT, but with simple classifiers like k nearest neighbor to estimate agreement.

Self-training is popular for segmentation, for propagating labels between pixels/voxels. It is used in the brain (Iglesias et al., 2010; Meier et al., 2014; Wang et al., 2014; Dittrich et al., 2014), retina (Gu et al., 2017), heart (Bai et al., 2017) and several other applications. Self-training is less common for computer-aided detection or diagnosis applications. In the surveyed papers, van Rikxoort et al. (2010) classify volumes of interest in chest CT and Singh et al. (2011) classify cell nuclei, but in both cases the sample size is in the thousands. This suggests that self-training is more often used for applications with larger datasets, which are common in segmentation but less so in computer-aided diagnosis.

A few works investigate how the sample size affects performance. Iglesias et al. (2010); Bai et al. (2017); Gu et al. (2017) all show that increasing the number of samples increases performance, and that the advantages of semi-supervised methods decrease as more labeled data becomes available.

3.3. Graph-based methods and regularization

Another popular strategy is to use the unlabeled data to better estimate the distribution of the data, and as such, regularize the classifier. Graph-based methods and semi-supervised SVMs fall under this category, but make different assumptions about the data.

Graph-based methods construct a graph with the samples as nodes, and similarities of these samples (defined via a distance measure and/or prior knowledge) as edges. The assumption is that connected samples are likely to have the same label. For example each pixel can be represented as a node, and pixels which are close together in the image, can be connected by edges. The goal is to propagate the labels along the graph. This can be achieved with a graph cut algorithm, which finds a labeling of the pixels such that the outputs for the already labeled training pixels are correct, and the outputs of all pixels are smooth along the graph (i.e. there is no salt and pepper noise). However, finding a labeling means that previously unseen images cannot be labeled without running the procedure again, also referred to as the out-of-sample problem. Graph cuts are often used for segmentation (Mahapatra et al., 2016; Song et al., 2009; Ciurte et al., 2014; Wang et al., 2014; Su et al., 2016), the labels are therefore propagated between pixels or superpixels. This means that for a previously unseen image that needs to be segmented, some labeled pixels would be needed.

Graph-based methods can be also used for atlas-based segmentation (Gass et al., 2012; Borga et al., 2016), but with an important difference. Instead of constructing a graph of pixels as above, atlas-based segmentation methods construct a graph of entire images. For example, for a set of MR scans where some scans have a ground truth segmentation, a graph can be constructed with each node representing a different MR scan. Scans which are similar according to some metric, have an edge between them. An unlabeled image can be segmented by propagating the ground truth segmentations from the labeled to the unlabeled images, and then combining the propagated segmentations.

Manifold regularization uses a similar idea of smoothness along a graph, and is able to label previously unseen data. Here the graph Laplacian encodes the similarity of the nodes and is used as a regularizer, encouraging smoothness along the graph. This method is used both for segmentation (Song et al., 2009; Park et al., 2014; Xu et al., 2016) and computer-aided diagnosis (Tiwari et al., 2010; An et al., 2016; Batmanghelich et al., 2011; Wang et al., 2017).

Semi-supervised SVMs use a different assumption, namely that there is a low density region between the classes. Next to fitting a hyperplane using the labeled training samples, semi-supervised SVMs also try to enforce this assumption, by favoring hyperplanes that place unlabeled samples outside the margin. This approach is used for classification of AD or MCI (Filipovych et al., 2011; Moradi et al., 2015).

4. Multiple instance learning

The multiple-instance learning (MIL) scenario can occur when obtaining ground-truth local annotations (i.e. for pixels or patches) is costly, time-consuming or not possible, but global labels for whole images, such as the overall condition of the patient, are available more readily. However, these labels do not apply to all the pixels or patches inside the image. MIL is an extension of supervised learning that can train classifiers using such weakly labeled data. For example, a classifier trained on images (*bags*), where each bag is labeled as normal or abnormal and consists of unlabeled image patches (*instances*), would be able to label novel images, and/or patches of that image as normal or abnormal.

A sample is a *bag* or set $X_i = \{\mathbf{x}_{ij} | j = 1, \ldots, N_i\} \subset \mathbb{R}^m$ of N_i instances, each instance is thus a m-dimensional feature vector. We are given labeled training bags $\{(X_i, Y_i) | i = 1, \ldots N_S\}$ where Y_i is the label. Originally MIL was formulated as a binary classification problem, but multi-class generalizations have also been proposed. For simplicity here we assume that $Y_i \in \{0, 1\}$.

The *standard assumption* is that there exist hidden instance labels $y_{ij} \in \{0, 1\}$ that relate to the bag labels as follows: a bag is positive if and only if it contains at least one positive instance. Another way to interpret this assumption is that most positive instance of the bag decides the bag label. Earlier approaches to MIL focused on finding the area of feature space with the positive instances, for example by explicitly adapting a rectangle to this area (Dietterich et al., 1997) or defining a density that was maximized when many instances from positive bags, but few instances from negative bags were present (Maron and Ratan, 1998). These approaches are computationally expensive, and we have not found examples of them being used in medical imaging. Later approaches sought to find the instance labels (rather than an area of the feature space), for example with a modified support vector machine (Andrews et al., 2002), that iteratively relabels the instances in order to satisfy the bag label constraints. This approach is relatively popular in medical imaging, sometimes achieving the best performance (Kandemir and Hamprecht, 2015). These are examples of *instance-level* classifiers.

Over the years, several other assumptions have been proposed (Foulds and Frank, 2010). A common assumption is the *collective assumption*, where all instances (rather than only the most positive one) contribute to the bag label. In this case, classification can be approached converting the problem into a supervised learning problem, for example by representing each bag as a single feature vector. Such approaches include representing a bag by minimum and maximum values of each feature (Gärtner et al., 2002) or similarities to instances (Chen et al., 2006) or bags (Cheplygina et al., 2015a) in the training set, and then using a linear classifier. These are examples of *bag-level* classifiers.

Originally, the goal in MIL was to train a bag classifier f_B to label previously unseen bags. Instance-level classifiers do this by inferring an instance classifier f_I, and combining the outputs of the bag's instances, for example by the noisy-or rule, $f_B(X_i) = \max_k \{f_I(\mathbf{x}_{ij})\}$. Bag-level classifiers typically represent each bag as a single feature vector and use supervised classifiers for training f_B directly. Such classifiers are often robust, but usually can not provide instance labels. Following Quellec et al. (2017), we refer to

methods that can provide instance labels as "primarily bag level" and methods that cannot as "exclusively bag level". For an in-depth review of MIL (not limited to medical imaging, see Amores (2013); Herrera et al. (2016); Carbonneau et al. (2018).

Because instance-level and some bag-level classifiers can provide instance labels, the focus of MIL became two-fold: classifying bags and classifying instances. This distinction also exists in medical imaging, as discussed in the next section.

4.1. MIL in medical imaging

MIL is a natural learning scenario for medical image analysis because labels are often not available at the desired granularity. The goal is therefore to exploit weaker bag labels for training. This idea can be used for different types of tasks. We adopt a categorization similar to that of Quellec et al. (2017): global detection, i.e. classifying the image as having a target pattern, local detection, i.e. classifying an image patch as having a target pattern, and false positive reduction, i.e. classifying an candidate lesion as true or false positive. Quellec et al. (2017) also discuss a "miscellaneous categorization" category, however, we find that this is very similar to "global detection".

The contributions are summarized in Table 4. The most common scenario where MIL is used, is global detection - classifying an entire image as having a particular disease or not. We discuss this in Section 4.2. However, instance classification - local detection - is also relevant. These goals are sometimes pursued simultaneously (Section 4.3). If only global detection is addressed, often local detection is relevant, but could not be addressed due to lack of labeled instances. We also briefly discuss local detection only (Section 4.4).

Another application of MIL is false positive reduction. An example is classifying candidate abnormalities, that may have been extracted by a different CAD algorithm, as true positives (abnormalities) or false detections. In this context, the candidate under consideration is the bag, and a different view of the candidate is an instance. These views can, for example, be different overlapping patches of a tumor in a CT scan, or different frames where a polyp has been captured in a video. In other words, the instance has an "is a" relationship to the bag (rather than "is part of" as with the global/location detection tasks). Due to this, the instances can be highly correlated, and no instances are truly negative. Here the goal is to classify the bag, and instance classification is not as relevant as in global/local detection. We discuss this scenario in Section 4.5.

4.2. Global detection

The majority of papers on MIL address global detection. The use of MIL is motivated by the fact that strong labels that would enable using supervised learning for local (and therefore also global) detection are not available. Weak labels are available more readily, but may not apply to the entire scan. As an example, consider detection of emphysema illustrated in Fig. 1. In a single scan, typically both healthy issue and emphysematous tissue would be present. For supervised learning, ideally outlines of emphysema would be required. However, these are not available, and only an assessment of whether emphysema is present is available. Another example is in histopathology imaging, where a diagnosis might be available for an entire tissue slice, but not where the cancerous cells are located.

This approach is suitable for many different applications, the more common ones being detection of diabetic retinopathy in retinal images (Venkatesan et al., 2015; Quellec et al., 2012; Kandemir and Hamprecht, 2015) and detection of cancerous regions in histopathology images (Kandemir and Hamprecht, 2015; Xu et al., 2014; Li et al., 2015). Although weak labels are available more easily, the datasets can still be quite small, starting at just 58 bags in a dataset of breast cancer in microarray images (Kandemir and Hamprecht, 2015). Others, such as datasets of COPD in chest CT images (Cheplygina et al., 2014) or tuberculosis in chest x-ray (Melendez et al., 2014; 2016) are in the order of a thousand scans. Only recently, very large datasets started appearing, such as the dataset of 100K chest x-rays used in Li et al. (2017b).

Global detection can be achieved both with instance-level methods and bag-level methods. Overall, bag-level methods seem to be more successful due to their ability to not treat instances independently (as instance-level methods would), but instead consider the correlations and structure of the instances. In such cases a MIL method can even outperform a fully supervised method (Kandemir and Hamprecht, 2015; Wang et al., 2015a; Vural et al., 2006; Samsudin and Bradley, 2010), showing that the lack of strong labels is not the only use case for MIL.

In some cases, these scenarios where it is best not to consider instances independently, are not referred to as MIL, but "batch classification" (Vural et al., 2006) or "group-based classification" (Samsudin and Bradley, 2010). An overview of these scenarios and their relationships to MIL can be found in Cheplygina et al. (2015b).

4.3. Global and local detection

Several papers focus both on global and local detection. In our illustrative example, rather than detecting whether a test scan contains emphysema, we might want to also find out where in the lungs it occurs. Another example in detection of tuberculosis (Melendez et al., 2014; 2016) it is important to both classify the image as having tuberculosis, and highlight the tuberculosis lesions in the image. In fact, in all papers where global detection is the focus, a local detection task could be defined. However, these local detection tasks are often not evaluated, since no labels are available for validation, for example Cheplygina et al. (2015a); Kandemir and Hamprecht (2015).

When both tasks can be addressed, this is done with either instance-level or primarily bag-level method, that can provide instance labels. However, solving two tasks with a single classifier introduces a problem, often overlooked in literature - that the best bag classifier is not necessarily the best instance classifier and vice versa. Cheplygina et al. (2015a) demonstrate that the best bag classifier can lead to unstable instance predictions, when trained on bootstrapped versions of the training set. Kandemir and Hamprecht (2015) compare several classifiers on a dataset of Barrett's cancer diagnosis in histopathology image, for which both bag-level and instance-level labels are available. The best bag classifier is an exclusively bag-level method, while the best instance classifier is an instance-level method that performs reasonably well on bags, but does not have the highest performance.

Papers where both global and local labels are available for training, show similar results. Li et al. (2017b) use both a large number of bag labels and a smaller number of instance labels to train a classifier for global and local detection of various chest x-ray abnormalities. The results show that, when instance classification is the goal, adding more labeled bags does not necessarily increase instance-level performance. Shin et al. (2017) use both bag and instance labels for localization and classification of breast masses. They show that bag labels should be given less weight than the instance labels - i.e. using all the labels together does not lead to the best results.

An important aspect of classifiers doing both global and local detection, is their explanation of the global label in terms of local labels, for example, highlighting abnormalities in an image. If the classifier is trained with global labels, it could happen that it only

Table 4
Overview of multiple instance learning applications. The third column refers to the type of problem addressed - global and or local detection or false positive reduction. The fourth columns refers to the type of classifier used - exclusively (excl bag) or primarily (prim bag) bag-level, or instance-level.

Reference	Application	MIL category	Method
Brain			
Tong et al. (2014)	AD classification	global	excl bag
Chen et al. (2015b)	cerebral small vessel disease detection	global	instance
Dubost et al. (2017)	enlarged perivascular space detection	local	instance
Eye			
Venkatesan et al. (2015)	diabetic retinopathy classification	global	excl bag
Quellec et al. (2012)	diabetic retinopathy classification	global, local	instance
Schlegl et al. (2015)	fluid segmentation	local	instance
Manivannan et al. (2016)	retinal nerve fiber layer visibility classification	global, local	instance
Lu et al. (2017)	fluid detection	global	instance
Breast			
Maken et al. (2014)	breast cancer detection	global	multiple
Sanchez de la Rosa et al. (2015)	breast cancer detection	global, local	excl bag
Shin et al. (2017)	mass localization, classification	global, local	instance
Lung			
Dundar et al. (2007)	pulmonary embolism detection	false positive	instance
Bi and Liang (2007)	pulmonary embolism detection	false positive	instance
Liang and Bi (2007)	pulmonary embolism detection	false positive	instance
Cheplygina et al. (2014)	COPD classification	global	multiple
Melendez et al. (2014)	tuberculosis detection	global, local	instance
Stainvas et al. (2014)	lung cancer lesion classification	false positive	instance
Melendez et al. (2016)	tuberculosis detection	global, local	instance
Kim and Hwang (2016)	tuberculosis detection	global, local	instance
Shen et al. (2016)	lung cancer malignancy prediction	global, local	instance
Cheplygina et al. (2018)	COPD classification	global	instance
Li et al. (2017b)	abnormality detection (14 classes)	global, local	instance
Abdomen			
Dundar et al. (2007)	polyp detection	false positive	instance
Wu et al. (2009)	polyp detection	false positive	instance
Lu et al. (2011)	polyp detection, size estimation	false positive	instance
Wang et al. (2012)	polyp detection	false positive	instance
Wang et al. (2015a)	lesion detection	global	prim bag
Wang et al. (2015b)	lesion detection	global	prim bag
Histology/Microscopy			
Dundar et al. (2010)	breast lesion detection	global	instance
Samsudin and Bradley (2010)	pap smear classification	global	multiple
McCann et al. (2012)	colitis detection	global	instance
Zhang et al. (2013)	skin biopsy annotation	global	multiple
Kandemir et al. (2014)	breast cancer detection	global	excl bag
Xu et al. (2014)	colon cancer detection	global, local	instance
Hou et al. (2015)	glioblastoma, low-grade glioma detection	global	instance
Li et al. (2015)	breast cancer detection	global	prim bag
Mercan et al. (2016)	breast cancer detection	global	instance
Kraus et al. (2016)	cell type classification	global, local	instance
Jia et al. (2017)	cancerous region segmentation (colon)	global, local	instance
Tomczak et al. (2017)	breast cancer detection	global	instance
Multiple			
Vural et al. (2006)	abnormality detection in three applications	false positive	instance
Kandemir and Hamprecht (2015)	abnormality detection in two applications	global, local	multiple
Hwang and Kim (2016)	lesion detection in two applications	global, local	instance
Other			
Situ et al. (2010)	dermoscopic feature annotation	global	prim bag
Liu et al. (2010)	cardiac event detection	global	instance
Yan et al. (2016)	bodypart recognition	global	instance

detects the most abnormal part of an image where multiple abnormalities are present. This creates an issue for the interpretability of the method. Currently work on attention mechanisms such as Ilse et al. (2018) is investigating solutions to this problem.

4.4. Local detection only

Given that we have covered global detection, and a combination of global and local detection, it would seem that local detection is the next logical category. Indeed, recently methods that focus only on the local detection have emerged. However, in such cases global detection is still a task that is being optimized for, so these papers can also be seen as falling under the “global and local detection” category.

Why have a “local detection only” category, if the category is technically empty? We decided to retain this section to explicitly address what might be a perceived difference. For example, a method that works by propagating the bag label to all the instances and training a supervised classifier, thus optimizing using the noisy instance labels, could be seen as a local detection only method. But in such a case, global detection is still assumed, both before training (label propagation) and during training and

evaluation (hyperparameter selection or deciding which classifier is best).

Methods focusing on local detection are often referred to as "weakly supervised". This term is sometimes interchangeably used with MIL, but seems to be common when global detection is not addressed. This might create a false impression that MIL and weak supervision (weak referring to only having bag labels) are disjoint, which is not the case.

On the other hand, not all papers that call themselves weakly supervised, fall under the MIL category. For example, Donner et al. (2009) uses "weakly supervised" to refer to "few annotations" that are used to initialize label propagation. In our classification this would be a SSL method. Rajchl et al. (2016) uses "weakly supervised" to describe that the labels are noisy, which is a different variation not covered in this survey.

4.5. False positive reduction with multiple views

A task that also focuses on classification of bags, but uses different assumptions, is known as false positive reduction within MIL (Quellec et al., 2017). Here the bag represents a candidate abnormality, such as tumor or lesion, possibly detected by a different method. The instances represent different views of a single abnormality (Wu et al., 2009; Lu et al., 2011; Wang et al., 2012). For example, Wang et al. (2012) addresses classification of polyp candidates in colonoscopy videos. A single polyp candidate is seen in multiple frames of the video. This way, the collection of video frames showing the same candidate is the bag, and each individual frame showing that candidate is an instance.

The bag label in principle applies to all the instances, since they are versions of the same candidate, which is different from the other MIL scenarios with label ambiguity. However, similar to global detection, a MIL classifier can outperform a supervised classifier, because it benefits from combining information of different instances.

A difference from global detection is that the instance classification task is less relevant. Since the goal is to classify the candidate as a whole, and the assumption is that all instances have the same label, it might be less interesting to find out which instances contributed the most to the candidate's label.

5. Transfer learning

Another popular learning scenario is transfer learning (Pan and Yang, 2010). Here the goal is to learn from related learning problems. One example is due to differences between acquisition of images, such as the use of different scanners or scanning protocols. Another example is related classification tasks for the same data, such as detection of different types of abnormalities.

More formally, in the scenarios covered so far we assumed that the training and test data are from the same *domain* $\mathcal{D} = (\mathcal{X}, p(X))$, defined by the feature space and distribution of the samples. We also assumed they addressed the same *task* $\mathcal{T} = (\mathcal{Y}, f(\cdot))$ defined by the label space and the mapping between the feature and the label space. In transfer learning scenarios, we assume that we are dealing either with different domains $\mathcal{D}_S \neq \mathcal{D}_T$ and/or different tasks $\mathcal{T}_S \neq \mathcal{T}_T$.

For example, such differences can be caused by different marginal distributions $p(\mathbf{x})$, different labeling functions $p(y|\mathbf{x})$, or even different feature and label spaces. In our illustrative example in Fig. 1, the $\mathbf{x}$'s are the feature vectors describing the appearance of lung ROIs, and the y's are the categories the patches belong to. Changes in subject groups, scanners and scanning protocols, can affect the distributions $p(\mathbf{x})$, such as "this dataset has lower intensities", $p(y)$, such as "this dataset has a large proportion of emphysema" and/or $p(y|\mathbf{x})$, such as "in this dataset this appearance corresponds to a different category". One or more of these differences mean that the distribution $p(D_S)$ of the training or source set is different from the distribution $p(D_T)$ of the test or target set.

Transfer learning approaches addressing these scenarios can be grouped by what they transfer. In this survey we focus on instance transfer i.e. assuming that source data can be reweighted to train the target classifier, and feature transfer i.e. encoding knowledge from the source domain into the feature representation for the target domain.

5.1. TL in medical imaging

The papers using TL are summarized in Table 5. We discuss these methods based on whether the tasks or the domains are different, or both.

In the "same domain, different task" scenario (Section 5.2), we are often dealing with multiple tasks for the same set of images, such as detecting multiple types of abnormalities, which are not necessarily mutually exclusive. This can be approached by modeling the detection of each type of abnormality as a separate classification problem (binary for presence/absence, or multi-class if for example severity of abnormality is estimated). Although the label spaces $\mathcal{Y}$ may be the same for each task, the labeling functions f are different, leading to $\mathcal{T}_S \neq \mathcal{T}_T$.

This scenario is often approached with feature transfer - learning features that are relevant for multiple tasks, thus effectively increasing the sample size and/or regularizing the classifier. An in-depth explanation of why this works can be found in Ruder (2017). Rather than training multiple tasks simultaneously, representation learning approaches where an (unsupervised) task such as reconstructing the data, is done first, are also possible.

In the "different domain, same task" scenario (Section 5.3), also referred to as domain adaptation, we are dealing with, for example, data acquired with different scanners. This can cause differences in the distributions of the samples $p(\mathbf{x})$ leading to differences in domains $\mathcal{D}_S \neq \mathcal{D}_T$. It is also possible that there are differences in the labeling functions. This means the source and target tasks are not strictly the same, but this can still be assumed by the method.

This scenario is often addressed with instance transfer. Instance transfer involves, for example, weighting source training samples such that only relevant samples receive high weights, or realigning the source domain with the target domain with the goal of bringing $p(\mathbf{x}_S)$ and $p(\mathbf{x}_T)$ closer together. After this, the union of the weighted instances can be used for training. The alignment approach can be referred to as feature transfer by others, because the features are being adapted. However, we group these approaches together since they are all aimed at decreasing the number of irrelevant samples, and/or increasing the number of relevant samples.

Finally, there is also a "different task, different domain" scenario (Section 5.4). Although according to Pan and Yang (2010) this would fall under "unsupervised transfer learning" and only address clustering, we find that this is also relevant in the supervised case, through feature transfer. In this case, the source task is used to pretrain a network. The network can then be used in two strategies (Litjens et al., 2017): for feature extraction, or as a starting point for further training (fine-tuning) of the target task. Both are currently very popular in medical image analysis.

5.2. Same domain, different tasks

Perhaps the earliest way in which transfer of information was leveraged within medical imaging, is inductive transfer learning, or learning different tasks within the same domain. For example, in lung images, we might be interested in detecting different types

Table 5
Overview of transfer learning applications. The last column refers to the type of transfer approach, i.e. whether it is instance transfer (by weighting or aligning samples) or feature transfer (by pretraining on an auxiliary task in the same or different domain, or multi-task learning).

Reference	Topic	Task	Domain	Transfer type
Brain				
Zhang and Shen (2012)	MCI conversion prediction	different	same	feature, multi-task
Wang et al. (2013)	tissue, lesion segmentation	same	different	instance, weight
van Opbroek et al. (2015a)	tissue, lesion segmentation	same	different	instance, weight
Guerrero et al. (2014)	AD classification	same	different	instance, align
van Opbroek et al. (2015b)	tissue, lesion segmentation	same	different	instance, weight
Cheng et al. (2015)	MCI conversion prediction	different	same	feature, multi-task
Goetz et al. (2016)	tumor segmentation	same	different	instance, weight
Wachinger and Reuter (2016)	AD classification	same	different	instance, weight
Cheplygina et al. (2016a)	tissue segmentation	same	different	instance, weight
Ghafoorian et al. (2017)	lesion segmentation	same	different	feature, pretraining
Kamnitsas et al. (2017)	segmentation of abnormalities	same	different	feature, pretraining
Alex et al. (2017)	lesion segmentation	different	same	feature, pretraining
Hofer et al. (2017)	AD classification	same	different	instance, align
Hon and Khan (2017)	AD classification	different	different	feature, pretraining
Kouw et al. (2017)	tissue segmentation	same, different	instance, align	
Breast				
Huynh and Giger (2016)	tumor detection	different	different	feature, pretraining
Samala et al. (2016)	mass detection	same	different	feature, pretraining
Kisilev et al. (2016)	lesion detection, description in mammography or ultrasound	different	same	feature, multi-task
Huynh et al. (2017)	chemotherapy response prediction	different	different	feature, pretraining
Dhungel et al. (2017)	mass detection, classification	different	same	feature, pretraining
Lung				
Bi et al. (2008)	abnormality classification	different	same	feature, multi-task
Schlegl et al. (2014)	lung tissue classification	different	same/different	feature, pretraining
Bar et al. (2015)	chest pathology detection	different	different	feature, pretraining
Ciompi et al. (2015)	nodule classification	different	different	feature, pretraining
Shen et al. (2016)	lung cancer malignancy prediction	different	same	feature, multi-task
Chen et al. (2017b)	attribute classification in nodules	different	same	feature, multi-task
Hussein et al. (2017)	attribute regression, malignancy prediction	different	same	feature, multi-task
Cheplygina et al. (2018)	COPD classification	same	different	instance, weight
Christodoulidis et al. (2017)	ILD classification	different	different/same	feature, pretraining/multi-task
Abdomen				
Ravishankar et al. (2016)	kidney detection	different	different	feature, pretraining
Nappi et al. (2016)	polyp detection	different	different	feature, pretraining
Sonoyama et al. (2016)	endoscopic image classification	same	different	instance, align
Azizi et al. (2017)	prostate cancer detection	same	different	feature, pretraining
Cha et al. (2017)	bladder cancer treatment response prediction	different	different	feature, pretraining
Chen et al. (2017a)	prostate cancer classification	different	different	feature, pretraining
Meng et al. (2017)	liver fibrosis classification	different	different	feature, pretraining
Li et al. (2017a)	gastrointestinal bleed detection	different	different	feature, pretraining
Ribeiro et al. (2017)	polyp classification in endoscopy	different	same/different	feature, pretraining
Mahmood et al. (2017)	depth estimation in endoscopy	same	different	feature, pretraining
Ross et al. (2017)	surgical instrument segmentation	same	different	feature, pretraining
Zhang et al. (2017)	polyp detection	different	different	feature, pretraining
Histology and microscopy				
Ablavsky et al. (2012)	mitochondria segmentation	same	different	instance, regularization
Becker et al. (2014)	mitochondria segmentation	same	different	instance, align
Kandemir (2015)	tumor detection	same	different	instance, align
Bermúdez-Chacón et al. (2016)	organelle segmentation	same	different	instance, regularization
Gadermayr et al. (2016)	glomeruli detection	same	different	instance, weight
Chang et al. (2017)	tissue classification	different	same	feature, pretraining
Phan et al. (2016)	staining pattern detection	different	different	feature, pretraining
Murthy et al. (2017)	visual attribute classification	different	same	feature, multi-task
Huang et al. (2017)	epithelium stroma classification	same	different	feature, pretraining
Spanhol et al. (2017)	breast cancer classification	different	different	feature, pretraining
Multiple				
Hwang and Kim (2016)	lesion detection, 2 applications	different	same	feature, multi-task
Moeskops et al. (2016)	segmentation, 3 applications	different	different	feature, multi-task
Tajbakhsh et al. (2016)	detection and segmentation, 4 applications	different	different	feature, pretraining
Other				
Bi et al. (2008)	heart segment classification	different	same	feature, multi-task
Ciompi et al. (2010)	plaque classification	same	different	instance, weight
Heimann et al. (2014)	US transducer localization	same	different	instance, weight
Chen et al. (2015a)	US standard plane localization	different	different	feature, pretraining
van Engelen et al. (2015)	carotid plaque component segmentation	same	different	instance, weight
Antony et al. (2016)	osteoarthritis quantification	different	different	feature, pretraining
Conjeti et al. (2016)	tissue classification	same	different	instance, align
Elmahdy et al. (2017)	skin lesion classification	different	different	feature, pretraining
Murphree and Ngufor (2017)	melanoma classification	different	different	feature, pretraining
Liu et al. (2017)	thyroid nodule classification	different	different	feature, pretraining
Menegola et al. (2017)	melanoma classification	different	different	feature, pretraining

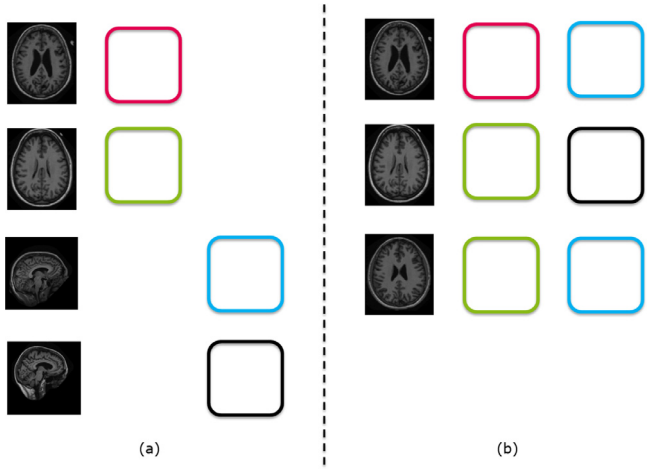

Fig. 2. Two strategies for multi-task learning: (a) tasks with disjoint training set and different output spaces (a red/green classification problem, and a blue/black classification problem) and (b) a multi-label setup where the training set is shared, but multiple labels are used. Brain images from OASIS dataset (Marcus et al., 2007). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of abnormalities. Rather than learning a multi-class task, or learning several binary tasks independently, we could learn these binary tasks jointly. The intuition is that these tasks will share task-independent features, and learning these tasks jointly increases the amount of data, leading to a more robust representation. This scenario includes *multi-task learning* (MTL), where a lot labeled source data is available, and self-taught learning, where no labeled source data is available.

We find that in medical imaging, many works fall under the multi-task learning scenario. There are two strategies to combine different tasks, shown in Fig. 2. In the first strategy, the training sets of each tasks are disjoint and the label spaces of the tasks are different. For example, Bi et al. (2008) describe a probabilistic framework for MTL algorithms and apply it to two applications with different characteristics. The first application is classifying nodules in chest CT, while also using labeled examples of ground glass opacities. Even though the tasks have different label spaces and it is not possible to train on the union of these datasets, learning the tasks jointly increases the effective sample size. Another example is classification of brain MR images into Alzheimer's disease (AD) or cognitively normal (CN). A related task is classifying images with mild cognitive impairment (MCI) subjects can be classified into converters (to AD) and non-converters. Cheng et al. (2015) combine these tasks, even though the training sets and the label spaces are disjoint.

The second strategy is to train two tasks using the same training set, which can also be seen as a multi-label classification problem. For example, the second application in Bi et al. (2008) is classification of multiple heart wall segments per subject. Instead of classifying each segment independently, they simultaneously classify all segments, essentially predicting a vector of labels per subject. This does not increase the sample size, but still benefits the classifier through regularization. The authors also demonstrate that MTL has the largest advantage at low sample sizes, where regularization is most needed. Similarly, Zhang and Shen (2012) jointly learn both the diagnosis (AD, MCI or CN) and two cognitive scores of the subjects. In a further experiment, they predict the change in these labels, i.e. the absolute change in the cognitive scores, and whether the MCI subjects convert to AD or not. They use the same training set, but where each subject has multiple labels.

The second strategy appears to be more common. Other applications where multiple labels are predicted include classification of lung diseases (Li et al., 2017b), and classification of visual attributes of images, such as attributes of lung nodules (Chen et al., 2017b; Hussein et al., 2017) or skin lesions (Murthy et al., 2017).

Finally, there are a couple of examples of self-taught learning, where there are labels for only one of the tasks. This happens in scenarios where one dataset needs to address multiple tasks, for example localization of abnormalities and their classification (Hwang and Kim, 2016), or description (Kisilev et al., 2016). There are then two optimization problems being solved using the same labels. Note that while Pan and Yang (2010) call these works "self-taught learning", in practice other names may be used, such as "self-transfer learning" (Hwang and Kim, 2016) or multi-task learning (Kisilev et al., 2016). There is a relationship between these works and MIL, which we will explore in the discussion.

In the examples above, multi-task learning is done by sharing the weights or parameters for the model, but using different outputs depending on the task. For example, in deep learning, this could be achieved by sharing the hidden layers, but using a different set of output layers. The label space for each of the tasks is therefore the same, as if that task was learned individually. An exception is Moeskops et al. (2016), where multiple tasks - tissue segmentation in MR, pectoral muscle segmentation in MR and coronary artery segmentation in CT - are learned in a joint label space, like a multi-class problem. While in principle this means that confusion between tasks could occur, for example a voxel of brain tissue could be classified as a voxel of the coronary artery, the results show that most errors happen within the same task.

Another way to use different tasks within the same domain, is by learning the tasks sequentially, rather than in parallel, as in multi-task learning. For example, Dhungel et al. (2017) first train a regression model to predict handcrafted features that are known to be related to the target labels. This model is then used for initialization of the target model. Although the handcrafted features are used as labels, they are not provided by experts, so this type of pretraining can be considered to be unsupervised.

There are other ways to add such unsupervised tasks to improve the target (supervised) task. An approach that is gaining popularity is finding a representation from which (part of) the data can be reconstructed. For example, Ross et al. (2017) first decolorize their training images, then use recolorization as an additional task to learn a good representation.

Related to the idea of adding unsupervised tasks is adversarial machine learning (Biggio, 2010), with generative adversarial networks (GANs) (Goodfellow et al., 2014) as a popular technique. GANs work by having an interplay between two networks - a generator, that generates samples based on the training set distribution, and a discriminator, that classifies such samples as either real or generated. By competing with each other, the networks learn a good representation of the data. In a sense, this is an example of learning from multiple tasks on the same data.

One example where GANs can be applied, is detection of abnormalities. For example, Schlegl et al. (2017) learn the distribution of healthy images of the retina. At test time, the GAN aims to reconstruct a healthy version of the unseen image. Since anomalities cannot be reconstructed, comparing the test image and the reconstruction identifies the locations where abnomalities are present. By learning the healthy data distribution, GANs can be applied to image denoising, image reconstruction, and image synthesis, for example, synthesizing CT from MR images. Recently it has been a very active topic in medical image analysis, and since the first version of this preprint, two surveys on GANs in medical imaging have appeared (Kazeminia et al., 2018; Wolterink et al., 2018).

5.3. Different domains, same task

Other early efforts in transfer learning in medical imaging focus on the scenario where the classification task is the same, but the domains are different, for example due to the use of data from different hospitals. Due to the differences in data distributions, it may not be optimal to simply train a classifier on one domain, and then test it on the other domain, or to train a classifier on the union of all available labeled data.

Changes in distributions can occur due to several reasons. For example, van Opbroek et al. (2015b,a); Kouw et al. (2017) address segmentation of MR data from different scanners, which alters the appearance of the images, and different populations, which changes the distribution of classes in the data. Ciompi et al. (2010); Conjeti et al. (2016) address differences between in vitro and in vivo ultrasound, where the absence/presence of blood flow causes a distribution shift. Bermúdez-Chacón et al. (2016) focus on segmentation of cells in microscopy images of different parts of the brain, which results in heterogeneous appearances.

The methods that address these distributions are mainly instance-transfer methods. One strategy is to change the source distribution by weighting the instances for training, such that the source distribution matches the target distribution as closely as possible. This is possible via importance weighting, where each instance is assigned a weight based on probability of belonging to the target domain. This strategy is optimal only if the marginal distributions are different but the labeling functions are the same, but in practice can also be helpful with different labeling functions (van Opbroek et al., 2015b; Cheplygina et al., 2018). Weights can also be assigned on other characteristics, without explicitly addressing the distributions of the feature vectors. When classifying subjects as having Alzheimer's, Wachinger and Reuter (2016) perform weighting based on patient characteristics such as age, while these factors are not used by the classifier.

Another instance-transfer strategy is to align the source and target domains by a transformation of the feature space. Once the domains are aligned, the instances of the source domain can be used for training. Conjeti et al. (2016) use principal component analysis to align in vitro and in vivo ultrasound images in feature space as a preprocessing step, before training a random forest on the source data and adapting it with the (aligned) target data. Guerrero et al. (2014) align subjects from 1.5 T and 3 T scanners by exploiting correspondences between the two domains. A correspondence-free approach to align representations of MR scans from different datasets is used by Hofer et al. (2017), by assuming a Gaussian distribution for each dataset. Kouw et al. (2017) use pairs of similar (same class) and dissimilar (different class) voxels from scans acquired with different scanners to learn an invariant feature representation. Training on the union of the voxels using this representation outperforms training on source or target data only, or the union of the voxels with the original representation.

Another difference between methods is whether they assume the presence of labeled data from the target domain. Unsupervised transfer is addressed in Wang et al. (2013); Heimann et al. (2014); Cheplygina et al. (2018); van Opbroek et al. (2015b) among others. Other works such as Conjeti et al. (2016); Wachinger and Reuter (2016); Goetz et al. (2016); van Opbroek et al. (2015a) focus on supervised transfer, with a small amount of labeled data from the target domain.

5.4. Different task, different domains

With the development of deep learning methods, it has become more common to transfer information between different tasks and different domains. The idea behind this is to find a good feature representation. This is achieved when a lot of source data is available, which can be used to train a deep network. This pretrained network can then be used to extract "off-the-shelf" features from the target dataset, or as a starting point for further training or fine-tuning the network to the target task. Such strategies are often compared to training a network "from scratch", i.e. without using transfer.

The source data can be from a totally different task. Using non-medical images as source data is now common for 2D networks. Probably the first work to do this is Schlegl et al. (2014). For the target task of classifying tissue types in chest CT slices, they used three different source tasks: natural images, other chest CT images, and head CT images. They found that natural images performed comparably or even slightly better than using only lung images. Using brain images was less effective, possibly due to large homogeneous areas present in brain CT, but not in lung CT, which has more texture information. After this work, more results showing transfer from non-medical images appeared, for example Bar et al. (2015); Ciompi et al. (2015).

Transfer from natural images is used often in practice. Common datasets used for transfer are datasets annually released by the Imagenet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015). The datasets have more than a million images and thousand categories of everyday objects. Since this methodology is so popular, we are not able to provide an exhaustive list of papers that apply it, and focus on papers that investigate underlying causes of when transfer is successful or not.

For detecting and classifying colorectal polyps, Zhang et al. (2017) transfer from Imagenet (1.2 million images in 1000 categories) and from Places, a scene recognition dataset of 2.5 million images in 205 categories such as "bathroom" (Zhou et al., 2017). Zhang et al. (2017) 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.

Menegola et al. (2017) compare off-the-shelf features and fine-tuning strategies for the task of melanoma classification, and use two datasets for pretraining: Imagenet and Kaggle Diabetic Retinopathy (KaggleDR) with 35K images.[1] KaggleDR contains retinal images that are in a sense similar to melanoma images, capturing a single object of interest. The authors find that finetuning outperforms the off-the-shelf strategy, and that transfer from Imagenet only is more successful than transfer from KaggleDR, or from the union of the datasets. Although the advantage of Imagenet over KaggleDR could perhaps be explained by the dataset size, the fact that the union of the datasets performs worse, indicates that there are more factors to be considered.

Ribeiro et al. (2017) investigate pretraining and fine-tuning of different source datasets for classification of polyps in endoscopy images. Different from the previous papers, they extract datasets of the same number of classes and images from the available types of data, making it a more fair comparison. They find that texture datasets perform best as source data, outperforming other datasets of endoscopy images. They also note that increasing the number of images and classes does not always improve performance.

These results do not always hold. In a study of predicting response to cancer treatment in the bladder, Cha et al. (2017) compare networks without TL, networks pretrained on natural images, and networks pretrained on bladder ROIs. They find that there are no statistically significant differences between the methods. Papers discussing comparisons between transfer from medical and non-medical images are summarized in Cheplygina (2018).

[1] https://www.kaggle.com/c/diabetic-retinopathy-detection/data.

These results are interesting if we consider that more traditional transfer learning methods focused on increasing the similarity between the source and target data. The results summarized here suggest that there is a trade-off between similarity, size and perhaps diversity of the source data.

6. Discussion

6.1. Trends

We first examine the overall trends in the use of different learning scenarios. Fig. 3 shows how the surveyed papers for each scenario are distributed across different years. Transfer learning is clearly the most popular, although this has only become evident in recent years. A reason for this might be the availability of datasets and tools. For SSL and MIL, a specific type of data/labels need to be available, while for TL, it is possible to use a completely external dataset in addition to the target data, and pretrained models can be easily downloaded.

There are also trends related to the different application areas. In this paper we have used the following categories, inspired by Litjens et al. (2017): brain, retina, chest, breast, heart, abdomen, histology/microscopy and other applications. Fig. 4 shows the distribution of these applications across the learning scenarios. Overall, brain is the most common application, followed by histology/microscopy and the abdomen. Breast, heart and retina, on the other hand, have relatively few papers. Around 10% of the papers address multiple applications. Although we survey different types of papers than Litjens et al. (2017), the distribution across the applications is quite similar. It would be of interest to compare this distribution to for example recent conference proceedings.

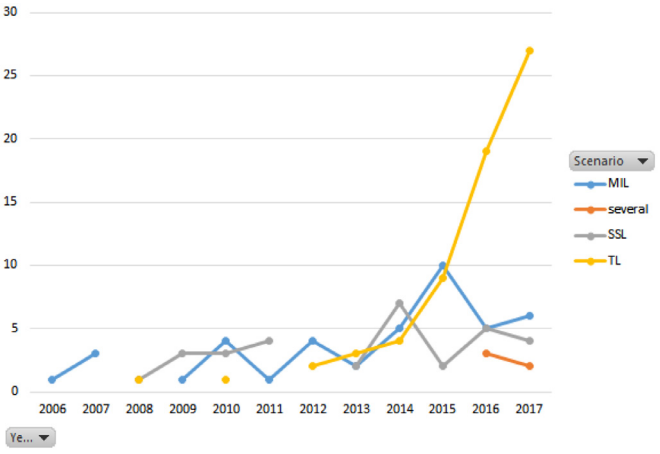

Fig. 3. Number of discussed papers by year, grouped by learning scenario.

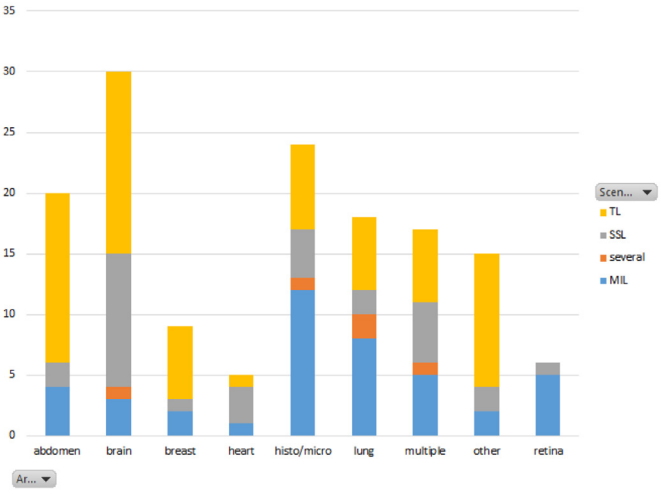

Fig. 4. Distribution of papers across learning scenarios and applications.

The application also influences the popularity of different learning scenarios. For example, MIL is frequently used for histology/microscopy, but is not as common for tasks within the brain. One reason is that in histology/microscopy it is more reasonable to assume that the patches within an image do not have an ordering, and there can be a variable number of patches per image, as is the case in multiple instance learning. However, this is less suitable for the brain, where anatomical correspondences can be determined and are informative, and the MIL scenario is less applicable.

6.2. Related learning scenarios

A gap in the current literature is that relevant learning scenarios are sometimes not considered. However, doing so could further deepen our understanding of the underlying classification problem, possibly leading to a better fit between the problem, assumptions and the method used.

One example of relevant learning scenarios is SSL and MIL. MIL can be seen as a special case of SSL (Zhou and Xu, 2007) if the traditional assumption is used, because the instances in negative bags can be considered labeled, and the instances in positive bags can be considered unlabeled, but with additional constraints that not all instances in a positive bag can be negative. Investigating this relationship by comparing methods with and without such constraints, could help elucidate the importance of the bag structure.

SSL and TL are also related. When dealing with a set of labeled data and a set of unlabeled data for the same task, without knowledge about the domains it could be logical to use a SSL method. On the other hand, if information about domains is available, a TL method would be more appropriate. Comparing the two could help with understanding the differences between datasets - is there perhaps a transfer learning situation where we didn't suspect one before? Since transfer learning has started becoming more popular only recently, it is possible that earlier papers that use multi-center data in SSL, such as Dundar et al. (2007), do not address this issue. Note that in the definitions we used here, SSL and TL are disjoint scenarios, because SSL assumes the training and test data to be from the same learning problem, while in TL there are different but related problems. However, from a more general perspective, a problem could be seen as both SSL and TL, for example when first training on unlabeled data and then transferring the representation to a labeled task.

There are several links between MIL and TL. Using MIL can avoid the need to use a TL method, because MIL labels from the same domain can be acquired more easily. This is illustrated in Melendez et al. (2014), where instance labels are available only for one domain, but bag labels are available for multiple domains. A MIL classifier trained on same-domain bag labels outperforms a fully supervised classifier trained on different-domain instance labels. It would have been valuable to see how a TL approach would compare to the same-domain MIL method, and to the combination of both.

Another link between MIL and TL is in scenarios where two related classification tasks are addressed, such as global detection and local detection. In MIL methods this is usually achieved by training a single classifier, but we can also view this as an example of multi-task or self-transfer learning, where two classifiers are trained with a shared representation. If both weak and strong labels are available in a MIL scenario, using both types of labels can also be seen as *mixed supervision*, which is a fairly recent term in medical imaging papers, see for example Shah et al. (2018).

There are also opportunities in exploring related learning scenarios that are not yet common in medical imaging. Positive and unlabeled learning (Elkan and Noto, 2008) has received quite some attention in the machine learning community. The idea is to learn from only positive and unlabeled examples, which may happen when the expert misses some positives during annotation. The absence of a positive label does not imply that an example is negative, and thus a non-positive example is considered unlabeled. Different from novelty detection, where only negative examples would be used for modeling healthy or normal examples, positive and unlabeled learning aims to still use the abnormalities as well as the unlabeled examples. Although this scenario seems very suitable for medical imaging, the only paper we have found directly addressing this scenario is Zuluaga et al. (2011).

Other possibilities include the "siblings" of MIL, such as batch classification (Vural et al., 2006) and group-based learning (Samsudin and Bradley, 2010). We have grouped these works in the MIL section, as they can be seen as variations of MIL with different assumptions (Cheplygina et al., 2015b). Although from the point of literature search it is counterproductive to use such different names for these scenarios, their similarities and differences could help us better understand the diversity of MIL problems being addressed in the literature.

Note that we did not focus on "weakly-supervised learning" (WSL) in this survey. This is because there does not seem to be an accepted definition of WSL. As we already pointed out in Section 4.4, WSL is sometimes used interchangeably with MIL. In other cases WSL refers to supervision with noisy labels. Finally, WSL can encompass SSL, MIL and learning with noisy labels (Zhou et al., 2017).

6.3. Full potential of available data

The available (labeled) data is not always used to its full potential, possibly due to the constraints of a particular method. For example, papers on MIL may (unnecessarily) convert a regression or multi-class problem into a binary problem, because this is how MIL was traditionally formulated. As a result, different grades or types of a disease can be aggregated into "healthy" and "abnormal". Others may remove more difficult classes. However, this is not necessary helpful for machine learning methods. For example, Menegola et al. (2016) demonstrate that removing one of the two disease classes results in lower performance, possibly because the method has fewer samples in total to learn from.

An opportunity is to use multiple labels when the ground truth is determined by consensus of different experts. Often the individual labels of the experts are combined into consensus labels, which are then used for training. However, as Warfield et al. (2004) and Guan et al. (2017) show, modeling the individual labelers during training can outperform averaging the labelers in advance.

Another opportunity is using clinical variables as additional outputs for the model. These are currently not used very often, but can improve prediction (Zhou et al., 2013). Even age or sex could be included as additional labels to predict. While not interesting prediction tasks by themselves, these could be leveraged via multi-task learning, for example, by using these as auxiliary tasks or "hints" (Ruder, 2017). Clinical reports with more detailed information, such as describing the location of abnormalities, can also provide additional information (Schlegl et al., 2015).

Finally, the data itself can be used for pretraining in an unsupervised manner, for example by reconstructing the data while learning a good representation. This is already being done by a few papers discussed in Section 5.2. However, this approach could be an opportunity for other applications where additional images and/or modalities are available, and that could be used as auxiliary tasks.

6.4. Acquiring additional labels

While the methods in this survey can certainly improve the robustness of classifiers, we feel there is a limit on what can be achieved without additional labels. Active learning methods such as Melendez et al. (2016); Su et al. (2016) aim to minimize the number of labels needed for the same or better performance, by only querying the labels that are most ambiguous or will lead to most improvement for the classifier. Given the same budget for labels, this could potentially lead to better performance overall.

Following the success of crowdsourcing in computer vision, crowdsourcing is also gaining an important place in medical imaging. These methods aim to collect (possibly noisy) labels from the public. When combining multiple annotators, the noise is expected to be reduced. Most studies to date investigated the quality of such labels compared to expert labels (Maier-Hein et al., 2015; Cheplygina et al., 2016b; Mitry et al., 2015). Methods that use the crowdsourced labels inside machine learning methods, are less common, for example Albarqouni et al. (2016). We expect that this will be an important direction for future research.

6.5. Generalization

A main challenge with not-so-supervised learning in medical imaging is that most works are proofs of concept on one or (less frequently) few applications. This makes it difficult to generalize the results and gain insight into whether the method would work in a different problem.

One partial solution would be to vary the characteristics of a single dataset - for example, subsample the training data to create learning curves, change the class priors to investigate the influence of class imbalance, or select or merge different classes. Another partial solution would be to perform ablation experiments, i.e. removing a part of the method's functionality, to understand what factors contribute most to the result.

A related challenge of not generalizing to other applications is publication bias: negative results, and/or results from an existing method may not be published, or published in a less popular venue. Borji (2018) provides an excellent discussion on why this is detrimental to research in computer vision. We feel that this is something that should also be discussed within the medical imaging community.

Challenges such as grand-challenges.org are a great resource for benchmarking algorithms on open datasets. However, these too often address only a single application, with the risk of overfitting to these datasets as a community. We see a promising research direction in platforms where the same methods could be applied to a range of datasets from different medical applications.

7. Conclusion

We have discussed over 140 papers in medical image analysis that focus on classification in a "not-so-supervised" learning scenario, often due to lack of representative annotated data. We focused on semi-supervised, multi-instance and transfer learning, of which transfer learning is the most popular in recent years. While individual papers demonstrate the usefulness of such approaches, there are still many questions on how to best use these methods. We expect future research to benefit from examining the connections between learning scenarios and generalizing the results between applications.

Conflict of Interest

We declare no conflict of interest.

Acknowledgments

We thank the following people for their constructive comments on version 1 of the preprint: Andreas Eklund, Caroline Petitjean, Gwenolé Quellec, Jakub Tomczak, Jesse Krijthe, Marco Loog, Maximilian Ilse, Ragav Venkatesan and Wouter Kouw. We also thank the anonymous reviewers for their constructive comments on version 2 of the preprint. This work was partly funded by the Netherlands Organization for Scientific Research (NWO), grant no. 639.022.010.

References

Ablavsky, V.H., Becker, C.J., Fua, P., 2012. Transfer Learning by Sharing Support Vectors. Technical Report.
Adal, K.M., Sidibé, D., Ali, S., Chaum, E., Karnowski, T.P., Mériaudeau, F., 2014. Automated detection of microaneurysms using scale-adapted blob analysis and semi-supervised learning. Comput. Methods Programs Biomed. 114 (1), 1–10.
Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N., 2016. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans. Med. Imaging 35 (5), 1313–1321.
Alex, V., Vaidhya, K., Thirunavukkarasu, S., Kesavadas, C., Krishnamurthi, G., 2017. Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation. J. Med. Imaging 4 (4), 041311.
Amores, J., 2013. Multiple instance classification: review, taxonomy and comparative study. Artif. Intell. 201, 81–105.
An, L., Adeli, E., Liu, M., Zhang, J., Shen, D., 2016. Semi-supervised hierarchical multimodal feature and sample selection for Alzheimer's disease diagnosis. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 79–87.
Andrews, S., Hofmann, T., Tsochantaridis, I., 2002. Multiple instance learning with generalized support vector machines. In: National Conference on Artificial Intelligence, pp. 943–944.
Antony, J., McGuinness, K., Connor, N.E.O., Moran, K., 2016. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In: International Conference on Pattern Recognition (ICPR).
Azizi, S., Mousavi, P., Yan, P., Tahmasebi, A., Kwak, J.T., Xu, S., Turkbey, B., Choyke, P., Pinto, P., Wood, B., et al., 2017. Transfer learning from rf to b-mode temporal enhanced ultrasound features for prostate cancer detection. Int. J. Comput. Assist. Radiol. Surg. 1–11.
Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D., 2017. Semi-supervised learning for network-based cardiac MR image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 253–260.
Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., Greenspan, H., 2015. Chest pathology detection using deep learning with non-medical training. In: International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 294–297.
Batmanghelich, K.N., Dong, H.Y., Pohl, K.M., Taskar, B., Davatzikos, C., et al., 2011. Disease classification and prediction via semi-supervised dimensionality reduction. In: International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1086–1090.
Baur, C., Albarqouni, S., Navab, N., 2017. Semi-supervised deep learning for fully convolutional networks. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 311–319.
Becker, C., Christoudias, M., Fua, P., Christoudias, C.M., Fua, P., 2014. Domain adaptation for microscopy imaging. IEEE Trans. Med. Imaging 34 (c), 1–14.
Bermúdez-Chacón, R., Becker, C., Salzmann, M., Fua, P., 2016. Scalable unsupervised domain adaptation for electron microscopy. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 326–334.
Bi, J., Liang, J., 2007. Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure. In: Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1–8.
Bi, J., Xiong, T., Yu, S., Dundar, M., Rao, R.B., 2008. An improved multi-task learning approach. Mach. Learn. Knowl. Discov. Databases 117–132.
Biggio, B., 2010. Adversarial Pattern Classification, Ph.D. thesis. University of Cagliari, Cagliari (Italy).
Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: Computational learning theory (COLT). ACM, pp. 92–100.
Borga, M., Andersson, T., Leinhard, O.D., 2016. Semi-supervised learning of anatomical manifolds for atlas-based segmentation of medical images. In: International Conference on Pattern Recognition (ICPR). IEEE, pp. 3146–3149.
Borji, A., 2018. Negative results in computer vision: a perspective. Image Vis. Comput. 69, 1–8.
de Bruijne, M., 2016. Machine learning approaches in medical image analysis: from detection to diagnosis. Med. Image Anal. 33, 94–97.
Carbonneau, M.-A., Cheplygina, V., Granger, E., Gagnon, G., 2018. Multiple instance learning: a survey of problem characteristics and applications. Pattern Recognit. 77, 329–353.
Cha, K.H., Hadjiiski, L.M., Chan, H.-P., Samala, R.K., Cohan, R.H., Caoili, E.M., Paramagul, C., Alva, A., Weizer, A.Z., 2017. Bladder cancer treatment response assessment using deep learning in CT with transfer learning. SPIE Medical Imaging. International Society for Optics and Photonics. 1013404–1013404.
Chang, H., Han, J., Zhong, C., Snijders, A., Mao, J.-H., 2017. Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications. IEEE Trans. Pattern Anal. Mach. Intell.
Chapelle, O., Schölkopf, B., Zien, A., et al., 2006. Semi-Supervised Learning, 2. MIT press Cambridge.
Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P.A., 2015a. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE J. Biomed. Health Inform. 19 (5), 1627–1636.
Chen, L., Tong, T., Ho, C.P., Patel, R., Cohen, D., Dawson, A.C., Halse, O., Geraghty, O., Rinne, P.E.M., White, C.J., Nakornchai, T., Bentley, P., Rueckert, D., 2015b. Identification of cerebral small vessel disease using multiple instance learning. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). Springer International Publishing, pp. 523–530.
Chen, Q., Xu, X., Hu, S., Li, X., Zou, Q., Li, Y., 2017a. A transfer learning approach for classification of clinical significant prostate cancers from mpMRI scans. SPIE Medical Imaging. International Society for Optics and Photonics. 101344F–101344F.
Chen, S., Qin, J., Ji, X., Lei, B., Wang, T., Ni, D., Cheng, J.-Z., 2017b. Automatic scoring of multiple semantic attributes with multi-task feature leverage: a study on pulmonary nodules in CT images. IEEE Trans. Med. Imaging 36 (3), 802–814.
Chen, Y., Bi, J., Wang, J., 2006. MILES: Multiple-instance learning via embedded instance selection. IEEE Trans. Pattern Anal. Mach. Intell. 28 (12), 1931–1947.
Cheng, B., Liu, M., Suk, H.-I., Shen, D., Alzheimer's Disease Neuroimaging Initiative, 2015. Multimodal manifold-regularized transfer learning for MCI conversion prediction. Brain Imaging Behav. 9 (4), 1–14.
Cheplygina, V., 2018. Cats or Cat Scans: Transfer Learning from Natural or Medical Image Source Datasets? arXiv:1810.05444.
Cheplygina, V., van Opbroek, A., Ikram, M.A., Vernooij, M.W., de Bruijne, M., 2016a. Asymmetric similarity-weighted ensembles for image segmentation. In: International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 273–277.
Cheplygina, V., Peña, I.P., Pedersen, J.H., Lynch, D.A., Sørensen, L., de Bruijne, M., 2018. Transfer learning for multi-center classification of chronic obstructive pulmonary disease. IEEE J. Biomed. Health Inform. 22 (5), 1486–1496.
Cheplygina, V., Perez-Rovira, A., Kuo, W., Tiddens, H., de Bruijne, M., 2016b. Early experiences with crowdsourcing airway annotations in chest CT. In: Large-scale Annotation of Biomedical data and Expert Label Synthesis (MICCAI LABELS), pp. 209–218.
Cheplygina, V., Sørensen, L., Tax, D.M.J., de Bruijne, M., Loog, M., 2015a. Label stability in multiple instance learning. In: Medical Imaging Computing and Computer Assisted Intervention (MICCAI), pp. 539–546.
Cheplygina, V., Sørensen, L., Tax, D.M.J., Pedersen, J.H., Loog, M., De Bruijne, M., 2014. Classification of COPD with multiple instance learning. In: International Conference on Pattern Recognition (ICPR), pp. 1508–1513.
Cheplygina, V., Tax, D.M.J., Loog, M., 2015b. On classification with bags, groups and sets. Pattern Recognit. Lett. 59, 11–17.
Christodoulidis, S., Anthimopoulos, M., Ebner, L., Christe, A., Mougiakakou, S., 2017. Multisource transfer learning with convolutional neural networks for lung pattern analysis. IEEE J. Biomed. Health Inform. 21 (1), 76–84.
Ciompi, F., de Hoop, B., van Riel, S.J., Chung, K., Scholten, E.T., Oudkerk, M., de Jong, P.A., Prokop, M., van Ginneken, B., 2015. 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. 26 (1), 195–202.
Ciompi, F., Pujol, O., Gatta, C., Rodríguez-Leor, O., Mauri-Ferré, J., Radeva, P., 2010. Fusing in-vitro and in-vivo intravascular ultrasound data for plaque characterization. Int. J. Cardiovasc. Imaging 26 (7), 763–779.
Ciurte, A., Bresson, X., Cuisenaire, O., Houhou, N., Nedevschi, S., Thiran, J.-P., Cuadra, M.B., 2014. Semi-supervised segmentation of ultrasound images based on patch representation and continuous min cut. PLoS ONE 9 (7), e100972.
Conjeti, S., Katouzian, A., Roy, A.G., Peter, L., Sheet, D., Carlier, S., Laine, A., Navab, N., 2016. Supervised domain adaptation of decision forests: transfer of models trained in vitro for in vivo intravascular ultrasound tissue characterization. Med. Image Anal. 32, 1–17.
Cozman, F., Cohen, I., 2006. Risks of semi-supervised learning. Semi-Supervised Learn. 56–72.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodol.) 1–38.
Dhungel, N., Carneiro, G., Bradley, A.P., 2017. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med. Image Anal. 37, 114–128.
Dietterich, T., Lathrop, R., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. 89 (1–2), 31–71.
Dittrich, E., Raviv, T.R., Kasprian, G., Donner, R., Brugger, P.C., Prayer, D., Langs, G., 2014. A spatio-temporal latent atlas for semi-supervised learning of fetal brain segmentations and morphological age estimation. Med. Image Anal. 18 (1), 9–21.
Donner, R., Wildenauer, H., Bischof, H., Langs, G., 2009. Weakly supervised group–wise model learning based on discrete optimization. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 12, pp. 860–868.
Dubost, F., Bortsova, G., Adams, H., Ikram, A., Niessen, W., Vernooij, M., de Bruijne, M., 2017. GP-Unet: Lesion detection from weak labels with a 3D regression network. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 214–221.
Dundar, M.M., Badve, S., Raykar, V.C., Jain, R.K., Sertel, O., Gurca, M.N., 2010. A multiple instance learning approach toward optimal classification of pathology slides. In: International Conference on Pattern Recognition (ICPR). IEEE, pp. 2732–2735.
Dundar, M.M., Fung, G., Krishnapuram, B., Rao, R.B., 2007. Multiple-instance learning algorithms for computer-aided detection. IEEE Trans. Biomed. Eng. 55 (3), 1015–1021.

Elkan, C., Noto, K., 2008. Learning classifiers from only positive and unlabeled data. In: International Conference on Knowledge Discovery and Data Mining. ACM, pp. 213–220.
Elmahdy, M.S., Abdeldayem, S.S., Yassine, I.A., 2017. Low quality dermal image classification using transfer learning. In: IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, pp. 373–376.
van Engelen, A., van Dijk, A.C., Truijman, M.T.B., van't Klooster, R., van Opbroek, A., van der Lugt, A., Niessen, W.J., Kooi, M.E., de Bruijne, M., 2015. Multi-center MRI carotid plaque component segmentation using feature normalization and transfer learning. IEEE Trans. Med. Imaging 34 (6), 1294–1305.
Filipovych, R., Davatzikos, C., Initiative, A.D.N., et al., 2011. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). Neuroimage 55 (3), 1109–1119.
Foulds, J., Frank, E., 2010. A review of multi-instance learning assumptions. Knowl. Eng. Rev. 25 (1), 1.
Gadermayr, M., Strauch, M., Klinkhammer, B.M., Djudjaj, S., Boor, P., Merhof, D., 2016. Domain Adaptive Classification for Compensating Variability in Histopathological Whole Slide Images. In: International Conference Image Analysis and Recognition. Springer International Publishing, pp. 616–622.
Gärtner, T., Flach, P.A., Kowalczyk, A., Smola, A.J., 2002. Multi-instance kernels. In: International Conference on Machine Learning, pp. 179–186.
Gass, T., Székely, G., Goksel, O., 2012. Semi-supervised segmentation using multiple segmentation hypotheses from a single atlas. In: Medical Computer Vision (MICCAI MCV). Springer, pp. 29–37.
Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R., de Leeuw, F.-E., Tempany, C.M., van Ginneken, B., et al., 2017. Transfer learning for domain adaptation in MRI: application in brain lesion segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 516–524.
Goetz, M., Weber, C., Binczyk, F., Polanska, J., Tarnawski, R., Bobek-Billewicz, B., Koethe, U., Kleesiek, J., Stieltjes, B., Maier-Hein, K.H., 2016. DALSA: Domain adaptation for supervised learning from sparsely annotated MR images. IEEE Trans. Med. Imaging 35 (1), 184–196.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680.
Gu, L., Zheng, Y., Bise, R., Sato, I., Imanishi, N., Aiso, S., 2017. Semi-supervised learning for biomedical image segmentation via forest oriented super pixels (voxels). In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 702–710.
Guan, M. Y., Gulshan, V., Dai, A. M., Hinton, G. E., 2017. Who said What: Modeling IndividualLlabelers improves Classification. arXiv:1703.08774.
Guerrero, R., Ledig, C., Rueckert, D., 2014. Manifold alignment and transfer learning for classification of Alzheimer's disease. In: Machine Learning in Medical Imaging (MICCAI MLMI), 8679. Springer, pp. 77–84.
Heimann, T., Mountney, P., John, M., Ionasec, R., 2014. Real-time ultrasound transducer localization in fluoroscopy images by transfer learning from synthetic training data. Med. Image Anal. 18 (8), 1320–1328.
Herrera, F., Ventura, S., Bello, R., Cornelis, C., Zafra, A., Sánchez-Tarragó, D., Vluymans, S., 2016. Multiple Instance Learning: Foundations and Algorithms. Springer.
Hofer, C., Kwitt, R., Höller, Y., Trinka, E., Uhl, A., 2017. Simple domain adaptation for cross-dataset analyses of brain MRI data. In: International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 441–445.
Hon, M., Khan, N., 2017. Towards Alzheimer's Disease Classification through Transfer Learning. arXiv:1711.11117.
Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., Saltz, J. H., 2015. Efficient multiple instance convolutional neural networks for gigapixel resolution image classification.
Huang, W., Chan, K.L., Gao, Y., Zhou, J., Chong, V., 2008. Semi-supervised nasopharyngeal carcinoma lesion extraction from magnetic resonance images using online spectral clustering with a learned metric. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 51–58.
Huang, Y., Zheng, H., Liu, C., Ding, X., Rohde, G., 2017. Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. IEEE J. Biomed. Health Inform.
Hussein, S., Cao, K., Song, Q., Bagci, U., 2017. Risk Stratification of Lung Nodules using 3D CNN-Based Multi-Task Learning. arXiv:1704.08797.
Huynh, B.Q., Antropova, N., Giger, M.L., 2017. Comparison of breast DCE-MRI contrast time points for predicting response to neoadjuvant chemotherapy using deep convolutional neural network features with transfer learning. SPIE Medical Imaging. International Society for Optics and Photonics. 101340U–101340U
Huynh, B.Q., Giger, M.L., 2016. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J. Med. Imaging 3 (3). 34501–34501
Hwang, S., Kim, H.-E., 2016. Self-Transfer Learning for Fully Weakly Supervised Object Localization 239–246. arXiv:1602.01625.
Iglesias, J.E., Liu, C.-Y., Thompson, P., Tu, Z., 2010. Agreement-based semi-supervised learning for skull stripping. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 13, pp. 147–154.
Ilse, M., Tomczak, J. M., Welling, M., 2018. Attention-Based Deep Multiple Instance Learning. arXiv:1802.04712.
Jia, Z., Huang, X., Eric, I., Chang, C., Xu, Y., 2017. Constrained deep weak supervision for histopathology image segmentation. IEEE Trans. Med. Imaging 36 (11), 2376–2388.
Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: International Conference on Information Processing in Medical Imaging (IPMI). Springer, pp. 597–609.
Kandemir, M., 2015. Asymmetric transfer learning with deep gaussian processes. In: International Conference on Machine Learning, pp. 730–738.
Kandemir, M., Hamprecht, F.A., 2015. Computer-aided diagnosis from weak supervision: a benchmarking study. Comput. Med. Imaging Graph. 42, 44–50.
Kandemir, M., Zhang, C., Hamprecht, F.A., 2014. Empowering multiple instance histopathology cancer diagnosis by cell graphs. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 8674, pp. 228–235.
Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A., 2018. GANs for Medical Image Analysis. arXiv:1809.06222.
Kim, H., Hwang, S., 2016. Scale-Invariant Feature Learning using Deconvolutional Neural Networks for Weakly-Supervised Semantic Segmentation. arxiv preprint. arXiv:1602.04984.
Kisilev, P., Sason, E., Barkan, E., Hashoul, S., 2016. Medical image description using multi-task-loss CNN. In: Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS). Springer, pp. 121–129.
Kouw, W. M., Loog, M., Bartels, L. W., Mendrik, A. M., 2017. MR Acquisition-Invariant Representation Learning. arXiv:1709.07944.
Kraus, O.Z., Ba, J.L., Frey, B.J., 2016. Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics 32 (12), i52–i59.
Krijthe, J.H., Loog, M., 2017. Robust semi-supervised least squares classification by implicit constraints. Pattern Recognit. 63, 115–126.
Li, W., Zhang, J., McKenna, S.J., 2015. Multiple Instance Cancer Detection by Boosting Regularised Trees. Springer International Publishing, pp. 645–652.
Li, X., Zhang, H., Zhang, X., Liu, H., Xie, G., 2017a. Exploring transfer learning for gastrointestinal bleeding detection on small-size imbalanced endoscopy images. In: International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 1994–1997.
Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.-J., Li, F.-F., 2017b. Thoracic Disease Identification and Localization with Limited Supervision. arXiv:1711.06373.
Liang, J., Bi, J., 2007. Computer aided detection of pulmonary embolism with tobogganing and mutiple instance classification in CT pulmonary angiography. In: Information Processing in Medical Imaging (IPMI). Springer, Berlin, Heidelberg, pp. 630–641.
Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.
Liu, Q., Qian, Z., Marvasty, I., Rinehart, S., Voros, S., Metaxas, D.N., 2010. Lesion-specific coronary artery calcium quantification for predicting cardiac event with multiple instance support vector machines. Springer Berlin Heidelberg, pp. 484–492.
Liu, T., Xie, S., Zhang, Y., Yu, J., Niu, L., Sun, W., 2017. Feature selection and thyroid nodule classification using transfer learning. In: International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1096–1099.
Loog, M., Jensen, A.C., 2015. Semi-supervised nearest mean classification through a constrained log-likelihood. IEEE Trans. Neural Netw. Learn. Syst. 26 (5), 995–1006.
Lu, D., Ding, W., Merkur, A., Sarunic, M.V., Beg, M.F., 2017. Multiple instance learning for age-related macular degeneration diagnosis in optical coherence tomography images. In: International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 139–142.
Lu, L., Bi, J., Wolf, M., Salganicoff, M., 2011. Effective 3D object detection and regression using probabilistic segmentation features in CT images. In: Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1049–1056.
Mahapatra, D., 2016. Combining multiple expert annotations using semi-supervised learning and graph cuts for medical image segmentation. Comput. Vis. Image Understanding 151, 114–123.
Mahapatra, D., Vos, F.M., Buhmann, J.M., 2016. Active learning based segmentation of Crohns disease from abdominal MRI. Comput. Methods Programs Biomed. 128, 75–85.
Mahmood, F., Chen, R., Durr, N. J., 2017. Unsupervised reverse domain adaption for synthetic medical images via adversarial training. arXiv:1711.06606.
Maier-Hein, L., Kondermann, D., Roß, T., Mersmann, S., Heim, E., Bodenstedt, S., Kenngott, H.G., Sanchez, A., Wagner, M., Preukschas, A., et al., 2015. Crowdtruth validation: a new paradigm for validating algorithms that rely on image correspondences. Int. J. Comput. Assist. Radiol. Surg. 10 (8), 1201–1212.
Maken, F.A., Gal, Y., Mcclymont, D., Bradley, A.P., 2014. Multiple instance learning for breast cancer magnetic resonance imaging. In: Digital Image Computing: Techniques and Applications (DICTA). IEEE, p. 1.
Manivannan, S., Cobb, C., Burgess, S., Trucco, E., 2016. Sub-category classifiers for multiple-instance learning and its application to retinal nerve fiber layer visibility classification. Medical Image Computing and Computer Assisted Intervention (MICCAI).
Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19 (9), 1498–1507.
Maron, O., Ratan, A.L., 1998. Multiple-instance learning for natural scene classification. In: International Conference on Machine Learning, 15, pp. 341–349.
McCann, M.T., Bhagavatula, R., Fickus, M.C., Ozolek, J.A., Kovacevic, J., 2012. Automated colitis detection from endoscopic biopsies as a tissue screening tool in

diagnostic pathology. In: International Conference on Image Processing (ICIP), 2012, pp. 2809–2812.
Meier, R., Bauer, S., Slotboom, J., Wiest, R., Reyes, M., 2014. Patient-specific semi-supervised learning for postoperative brain tumor segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 17, pp. 714–721.
Melendez, J., van Ginneken, B., Maduskar, P., Philipsen, R.H.H.M., Ayles, H., Snchez, C.I., 2016. On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis. IEEE Trans. Med. Imaging 35 (4), 1013–1024.
Melendez, J., van Ginneken, B., Maduskar, P., Philipsen, R.H.H.M., Reither, K., Breuninger, M., Adetifa, I.M.O., Maane, R., Ayles, H., Sanchez, C.I., 2014. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays. IEEE Trans. Med. Imaging 34 (1), 179–192.
Menegola, A., Fornaciali, M., Pires, R., Avila, S., Valle, E., 2016. Towards automated melanoma screening: Exploring transfer learning schemes. arXiv:1609.01228.
Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F.V., Avila, S., Valle, E., 2017. Knowledge transfer for melanoma screening with deep learning. In: International Sympsium on Biomedical Imaging (ISBI). IEEE, pp. 297–300.
Meng, D., Zhang, L., Cao, G., Cao, W., Zhang, G., Hu, B., 2017. Liver fibrosis classification based on transfer learning and fcnet for ultrasound images. IEEE Access.
Mercan, C., Mercan, E., Aksoy, S., Shapiro, L.G., Weaver, D.L., Elmore, J.G., 2016. Multi-instance multi-label learning for whole slide breast histopathology. In: Gurcan, M.N., Madabhushi, A. (Eds.), Medical Imaging 2016: Digital Pathology. International Society for Optics and Photonics, p. 979108.
Mitry, D., Peto, T., Hayat, S., Blows, P., Morgan, J., Khaw, K.-T., Foster, P.J., 2015. Crowdsourcing as a screening tool to detect clinical features of glaucomatous optic neuropathy from digital photography. PLoS ONE 10 (2), 1–8.
Moeskops, P., Wolterink, J.M., van der Velden, B.H.M., Gilhuijs, K.G.A., Leiner, T., Viergever, M.A., Išgum, I., 2016. Deep learning for multi-task medical image segmentation in multiple modalities. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 478–486.
Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., 2015. Machine learning framework for early MRI-based alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398–412.
Murphree, D. H., Ngufor, C., 2017. Transfer Learning for Melanoma Detection: Participation in ISIC 2017 Skin Lesion Classification Challenge. arXiv:1703.05235.
Murthy, V., Hou, L., Samaras, D., Kurc, T.M., Saltz, J.H., 2017. Center-focusing multi-task CNN with injected features for classification of glioma nuclear images. In: IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 834–841.
Nappi, J.J., Hironaka, T., Regge, D., Yoshida, H., 2016. Deep transfer learning of virtual endoluminal views for the detection of polyps in CT colonography. Int. Soc. Opt. Photon. 97852B–97852B
van Opbroek, A., Ikram, M.A., Vernooij, M.W., De Bruijne, M., 2015a. Transfer learning improves supervised image segmentation across imaging protocols. IEEE Trans. Med. Imaging 34 (5), 1018–1030.
van Opbroek, A., Vernooij, M.W., Ikram, M.A., de Bruijne, M., 2015b. Weighting training images by maximizing distribution similarity for supervised segmentation across scanners. Med. Image Anal. 24 (1), 245–254.
Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.
Parag, T., Plaza, S., Scheffer, L., 2014. Small sample learning of superpixel classifiers for EM segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 389–397.
Park, S.H., Gao, Y., Shi, Y., Shen, D., 2014. Interactive prostate segmentation using atlas-guided semi-supervised learning and adaptive feature selection. Med. Phys. 41 (11). 111715–111715
Phan, H.T.H., Kumar, A., Kim, J., Feng, D., 2016. Transfer learning of a convolutional neural network for HEp-2 cell image classification. In: International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1208–1211.
Prasad, M., Sowmya, A., Wilson, P., 2009. Multi-level classification of emphysema in HRCT lung images. Pattern Anal. Appl. 12 (1), 9–20.
Quellec, G., Cazuguel, G., Cochener, B., Lamard, M., 2017. Multiple-instance learning for medical image and video analysis. IEEE Rev. Biomed. Eng. 10, 213–234.
Quellec, G., Lamard, M., Abràmoff, M.D., Decencière, E., Lay, B., Erginay, A., Cochener, B., Cazuguel, G., 2012. A multiple-instance learning framework for diabetic retinopathy screening. Med. Image Anal. 16 (6), 1228–1240.
Rajchl, M., Lee, M. C., Schrans, F., Davidson, A., Passerat-Palmbach, J., Tarroni, G., Alansary, A., Oktay, O., Kainz, B., Rueckert, D., 2016. Learning under Distributed Weak Supervision. arXiv:1606.01100.
Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P., Babu, N., Vaidya, V., 2016. Understanding the mechanisms of deep transfer learning for medical images. In: Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS). Springer, pp. 188–196.
Ribeiro, E., Häfner, M., Wimmer, G., Tamaki, T., Tischendorf, J., Yoshida, S., Tanaka, S., Uhl, A., 2017. Exploring texture transfer learning for colonic polyp classification via convolutional neural networks. In: International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1044–1048.
van Rikxoort, E., Galperin-Aizenberg, M., Goldin, J., Kockelkorn, T.T.J.P., van Ginneken, B., Brown, M., 2010. Multi-classifier semi-supervised classification of tuberculosis patterns on chest CT scans. In: Pulmonary Image Analysis (MICCAI PIA), pp. 41–48.
Sanchez de la Rosa, R., Lamard, M., Cazuguel, G., Coatrieux, G., Cozic, M., Quellec, G., 2015. Multiple-instance learning for breast cancer detection in mammograms. In: International Conference of the IEEE Engineering in Medicine and Biology Society, 2015, pp. 7055–7058.
Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., Kenngott, H., et al., 2017. Exploiting the Potential of Unlabeled Endoscopic Video Data with Self-Supervised Learning. arXiv:1711.09726.
Ruder, S., 2017. An Overview of Multi-task Learning in Deep Neural Networks. arXiv:1706.05098.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.
Samala, R.K., Chan, H.-P., Hadjiiski, L., Helvie, M.A., Wei, J., Cha, K., 2016. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. Med. Phys. 43 (12), 6654–6666.
Samsudin, N.A., Bradley, A.P., 2010. Nearest neighbour group-based classification. Pattern Recognit. 43 (10), 3458–3467.
Schlegl, T., Ofner, J., Langs, G., 2014. Unsupervised pre-training across image domains improves lung tissue classification. In: Medical Computer Vision: Algorithms for Big Data (MICCAI MCV). Springer, pp. 82–93.
Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 146–157.
Schlegl, T., Waldstein, S.M., Vogl, W.-D., Schmidt-Erfurth, U., Langs, G., 2015. Predicting Semantic Descriptions from Medical Images with Convolutional Neural Networks. In: Information Processing in Medical Imaging (IPMI). Springer International Publishing, pp. 437–448.
Shah, M.P., Merchant, S.N., Awate, S.P., 2018. MS-Net: Mixed-supervision fully-convolutional networks for full-resolution segmentation. In: IMedical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 379–387.
Shen, W., Zhou, M., Yang, F., Dong, D., Yang, C., Zang, Y., Tian, J., 2016. Learning from Experts: Developing Transferable Deep Features for Patient-level Lung Cancer Prediction. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer International Publishing, pp. 124–131.
Shin, S. Y., Lee, S., Yun, I. D., Lee, K. M., 2017. Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. arXiv:1710.03778.
Singh, S., Janoos, F., Pécot, T., Caserta, E., Leone, G., Rittscher, J., Machiraju, R., 2011. Identifying nuclear phenotypes using semi-supervised metric learning. In: Information Processing in Medical Imaging (IPMI). Springer, pp. 398–410.
Situ, N., Yuan, X., Zouridakis, G., 2010. Boosting instance prototypes to detect local dermoscopic features. In: International Conference of the IEEE Engineering in Medicine and Biology Society, 2010, pp. 5561–5564.
Song, Y., Zhang, C., Lee, J., Wang, F., Xiang, S., Zhang, D., 2009. Semi-supervised discriminative classification with application to tumorous tissues segmentation of MR brain images. Pattern Anal. Appl. 12 (2), 99–115.
Sonoyama, S., Tamaki, T., Hirakawa, T., Raytchev, B., Kaneda, K., Koide, T., Yoshida, S., Mieno, H., Tanaka, S., 2016. Transfer Learning for Endoscopic Image Classification. arXiv:1608.06713.
Spanhol, F.A., Oliveira, L.S., Cavalin, P.R., Petitjean, C., Heutte, L., 2017. Deep features for breast cancer histopathological image classification. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 1868–1873.
Stainvas, I., Manevitch, A., Leichter, I., 2014. Cancer detection with multiple radiologists via soft multiple instance logistic regression and l_1 regularization. arXiv:1412.2873.
Su, H., Yin, Z., Huh, S., Kanade, T., Zhu, J., 2016. Interactive cell segmentation based on active and semi-supervised learning. IEEE Trans. Med. Imaging 35 (3), 762–777.
Sun, W., Tseng, T.-L.B., Zhang, J., Qian, W., 2016. Computerized breast cancer analysis system using three stage semi-supervised learning method. Comput. Methods Programs Biomed. 135, 77–88.
Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging 35 (5), 1299–1312.
Tiwari, P., Kurhanewicz, J., Rosen, M., Madabhushi, A., 2010. Semi supervised multi kernel (SeSMiK) graph embedding: identifying aggressive prostate cancer via magnetic resonance imaging and spectroscopy. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 666–673.
Tomczak, J. M., Ilse, M., Welling, M., 2017. Deep Learning with Permutation-invariant Operator for Multi-instance Histopathology Classification. arXiv:1712.00310.
Tong, T., Wolz, R., Gao, Q., Hajnal, J.V., Rueckert, D., 2014. Multiple instance learning for classification of dementia in brain MRI. Med. Image Anal. 16 (Pt 2), 599–606.
Venkatesan, R., Chandakkar, P.S., Li, B., 2015. Simpler non-parametric methods provide as good or better results to multiple-instance learning. In: International Conference on Computer Vision (ICCV), pp. 2605–2613. doi:10.1109/ICCV.2015.299.
Vural, V., Fung, G., Krishnapuram, B., Dy, J., Rao, B., 2006. Batch classification with applications in computer aided diagnosis. In: European Conference on Machine Learning (ECML). Springer, pp. 449–460.
Wachinger, C., Reuter, M., 2016. Domain adaptation for Alzheimer's disease diagnostics. Neuroimage 139, 470–479.
Wang, B., Liu, W., Prastawa, M., Irimia, A., Vespa, P.M., van Horn, J.D., Fletcher, P.T., Gerig, G., 2014. 4D active cut: An interactive tool for pathological anatomy modeling. In: International Symposium on Biomedical Imaging (ISBI), 2014, pp. 529–532.
Wang, B., Prastawa, M., Saha, A., Awate, S.P., Irimia, A., Chambers, M.C., Vespa, P.M., Van Horn, J.D., Pascucci, V., Gerig, G., 2013. Modeling 4D changes in pathological anatomy using domain adaptation: analysis of TBI imaging using a tumor database. Multimodal Brain Image Anal. (MICCAI MBIA) 8159, 31–39.

Wang, L., Li, S., Chen, Y., Lin, J., Liu, C., Zeng, X., Li, S., 2017. Direct aneurysm volume estimation by multi-view semi-supervised manifold learning. In: International Symposium on Biomedical Imaging. IEEE, pp. 1222–1225.

Wang, S., Cong, Y., Fan, H., Yang, Y., Tang, Y., Zhao, H., 2015a. Computer aided endoscope diagnosis via weakly labeled data mining. In: International Conference on Image Processing (ICIP). IEEE, pp. 3072–3076.

Wang, S., Li, D., Petrick, N., Sahiner, B., Linguraru, M.G., Summers, R.M., 2015b. Optimizing area under the ROC curve using semi-supervised learning. Pattern Recognit. 48 (1), 276–287.

Wang, S., McKenna, M.T., Nguyen, T.B., Burns, J.E., Petrick, N., Sahiner, B., Summers, R.M., 2012. Seeing is believing: video classification for computed tomographic colonography using multiple-instance learning. IEEE Trans. Med. Imaging 31 (5), 1141–1153.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.

Weese, J., Lorenz, C., 2016. Four challenges in medical image analysis from an industrial perspective. Med. Image Anal. 33, 44–49.

Wolterink, J. M., Kamnitsas, K., Ledig, C., Išgum, I., 2018. Generative Adversarial Networks and Adversarial Methods in Biomedical Image Analysis. arXiv:1810.10352.

Wu, D., Bi, J., Boyer, K., 2009. A min-max framework of cascaded classifier with multiple instance learning for computer aided diagnosis. In: Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1359–1366.

Xie, Y., Ho, J., Vemuri, B.C., 2013. Multiple atlas construction from a heterogeneous brain MR image collection. IEEE Trans. Med. Imaging 32 (3), 628–635.

Xu, K., Su, H., Zhu, J., Guan, J.-S., Zhang, B., 2016. Neuron segmentation based on CNN with semi-supervised regularization. In: Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 20–28.

Xu, Y., Zhu, J.-Y., Chang, E.I., Lai, M., Tu, Z., 2014. Weakly supervised histopathology cancer image segmentation and classification. Med. Image Anal. 18 (3), 591–604.

Yan, Z., Zhan, Y., Peng, Z., Liao, S., Shinagawa, Y., Zhang, S., Metaxas, D.N., Zhou, X.S., 2016. Multi-instance deep learning: discover discriminative local anatomies for bodypart recognition. IEEE Trans. Med. Imaging 35 (5), 1332–1343.

Zhang, D., Shen, D., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 59 (2), 895–907.

Zhang, G., Yin, J., Li, Z., Su, X., Li, G., Zhang, H., 2013. Automated skin biopsy histopathological image annotation using multi-instance representation and learning. BMC Med. Genom. 6 (Suppl 3). S10–S10

Zhang, R., Zheng, Y., Mak, T.W.C., Yu, R., Wong, S.H., Lau, J.Y., Poon, C.C., 2017. Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain. IEEE J. Biomed. Health Inform. 21 (1), 41–47.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell.

Zhou, J., Liu, J., Narayan, V.A., Ye, J., 2013. Modeling disease progression via multi–task learning. Neuroimage 78, 233–248.

Zhou, Z.-H., Xu, J.-M., 2007. On the relation between multi-instance learning and semi-supervised learning. In: International Conference on Machine learning (ICML), pp. 1167–1174.

Zhu, X., Goldberg, A.B., 2009. Introduction to semi-supervised learning. Synth. Lect. Artif. Intell. Mach. Learn. 3 (1), 1–130.

Zuluaga, M.A., Hush, D., Leyton, E.J.F.D., Hoyos, M.H., Orkisz, M., 2011. Learning from only positive and unlabeled data to detect lesions in vascular CT images. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 9–16.