HHS Public Access

Author manuscript

Med Image Anal. Author manuscript; available in PMC 2021 October 01.

Published in final edited form as:

Med Image Anal. 2020 October ; 65: 101759. doi:10.1016/j.media.2020.101759.

Deep learning with noisy labels: exploring techniques and remedies in medical image analysis

Davood Karimi[a,*], **Haoran Dou**[a], **Simon K. Warfield**[a], **Ali Gholipour**[a]

[a]Department of Radiology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Abstract

Supervised training of deep learning models requires large labeled datasets. There is a growing interest in obtaining such datasets for medical image analysis applications. However, the impact of label noise has not received sufficient attention. Recent studies have shown that label noise can significantly impact the performance of deep learning models in many machine learning and computer vision applications. This is especially concerning for medical applications, where datasets are typically small, labeling requires domain expertise and suffers from high inter- and intra-observer variability, and erroneous predictions may influence decisions that directly impact human health. In this paper, we first review the state-of-the-art in handling label noise in deep learning. Then, we review studies that have dealt with label noise in deep learning for medical image analysis. Our review shows that recent progress on handling label noise in deep learning has gone largely unnoticed by the medical image analysis community. To help achieve a better understanding of the extent of the problem and its potential remedies, we conducted experiments with three medical imaging datasets with different types of label noise, where we investigated several existing strategies and developed new methods to combat the negative effect of label noise. Based on the results of these experiments and our review of the literature, we have made recommendations on methods that can be used to alleviate the effects of different types of label noise on deep models trained for medical image analysis. We hope that this article helps the medical image analysis researchers and developers in choosing and devising new techniques that effectively handle label noise in deep learning.

Graphical Abstract

[*]Corresponding author: Tel.: +1-617-208-9736; fax: +1-617-730-0635;.

Davood Karimi: Conceptualization, Methodology, Investigation, Writing - Original Draft, Writing - Review & Editing

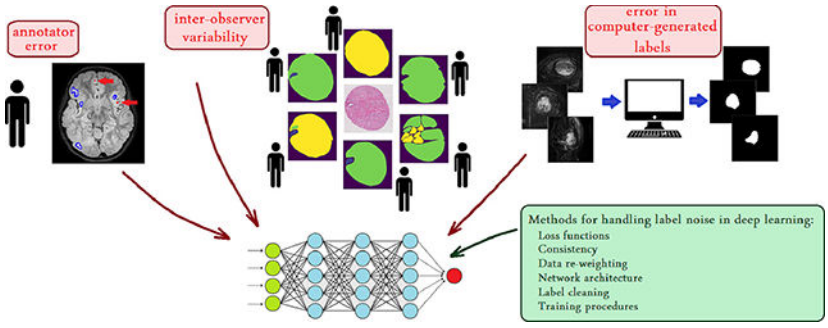
Haoran Dou: Methodology, Software, Investigation

Simon K. Warfield: Supervision, Resources, Data curation, Writing - Original Draft

Ali Gholipour: Conceptualization, Data curation, Supervision, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Project administration, Funding acquisition

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.



Keywords

label noise; deep learning; machine learning; big data; medical image annotation

1. Introduction

1.1. Background

Deep learning has already made an impact on many branches of medicine, in particular medical imaging, and its impact is only expected to grow (Ching et al., 2018; Topol, 2019b). Even though it was first greeted with much skepticism (Wang et al., 2017a), in a few short years it proved itself to be a worthy player in solving many problems in medicine, including problems in disease and patient classification, patient treatment recommendation, outcome prediction, and more (Ching et al., 2018). Many experts believe that deep learning will play an important role in the future of medicine and will be an enabling tool in medical research and practice (Topol, 2019a; Prevedello et al., 2019). With regard to medical image analysis, methods that use deep learning have already achieved impressive, and often unprecedented, performance in many tasks ranging from low-level image processing tasks such as denoising, enhancement, and reconstruction (Wang et al., 2018b), to more high-level image analysis tasks such as segmentation, detection, classification, and registration (Ronneberger et al., 2015; Haskins et al., 2019), and even more challenging tasks such as discovering links between the content of medical images and patient's health and survival (Xu et al., 2019; Mobadersany et al., 2018).

The recent success of deep learning has been attributed to three main factors (LeCun et al., 2015; Sun et al., 2017). First, technical advancements in network architecture design, network parameter initialization, and training methods. Second, increasing availability of more powerful computational hardware, in particular graphical processing units and parallel processing, that allow training of very large models on massive datasets. Last, but not least, increasing availability of very large and growing datasets. However, even though in some applications it has become possible to curate large datasets with reliable labels, in most applications it is very difficult to collect and accurately label datasets large enough to effortlessly train deep learning models. A solution that is becoming more popular is to employ non-expert humans or automated systems with little or no human supervision to label massive datasets (Guo et al., 2016; Deng et al., 2009; Ipeirotis et al., 2010). However,

datasets collected using such methods typically suffer from very high label noise (Wang et al., 2018a; Kuznetsova et al., 2018), thus they have limited applicability in medical imaging.

The challenge of obtaining large datasets with accurate labels is particularly significant in medical imaging. The available data is typically small to begin with, and data access is hampered by such factors as patient privacy and institutional policies. Furthermore, labeling of medical images is very resource-intensive because it depends on domain experts. In some applications, there is also significant inter-observer variability among experts, which will necessitate obtaining consensus labels or labels from multiple experts and proper methods of aggregating those labels (Bridge et al., 2016; Nir et al., 2018). Some studies have been able to employ a large number of experts to annotate large medical image datasets (Gulshan et al., 2016; Esteva et al., 2017). However, such efforts depend on massive financial and logistical resources that are not easy to obtain in many domains. Alternatively, a few studies have successfully used automated mining of medical image databases such as hospital picture archiving and communication systems (PACS) to build large training datasets (Yan et al., 2018; Irvin et al., 2019). However, this method is not always applicable as historical data may not include all the desired labels or images. Moreover, label noise in such datasets is expected to be higher than in expert-labeled datasets. There have also been studies that have used crowd-sourcing methods to obtain labels from non-experts (Gurari et al., 2015; Albarqouni et al., 2016). Even though this method may have potential for some applications, it has a limited scope because in most medical applications non-experts are unable to provide useful labels. Even for relatively simple segmentation tasks, computerized systems have been shown to generate significantly less accurate labels compared with human experts and crowdsourced non-experts (Gurari et al., 2015). In general, lack of large datasets with trustworthy labels is considered to be one of the biggest challenges facing a wider adoption and successful deployment of deep learning methods in medical applications (Langlotz et al., 2019; Ching et al., 2018; Ravì et al., 2016).

1.2. Aims and scope of this paper

Given the outline presented above, it is clear that relatively small datasets with noisy labels are, and will continue to be, a common scenario in training deep learning models in medical image analysis applications. Hence, algorithmic approaches that can effectively handle the label noise are highly desired. In this manuscript, we first review and explain the recent advancements in training deep learning models in the presence of label noise. We review the methods proposed in the general machine learning literature, most of which have not yet been widely employed in medical imaging applications. Then, we review studies that have addressed label noise in deep learning with medical imaging data. Finally, we present the results of our experiments on three medical image datasets with noisy labels, where we investigate the performance of several strategies to deal with label noise, including a number of new methods that we have developed for each application. Based on our results, we make general recommendations to improve deep learning with noisy training labels in medical imaging data.

In the field of medical image analysis, in particular, the notion of label noise is elusive and not easy to define. The term has been used in the literature to refer to different forms of label

imperfections or corruptions. Especially in the era of big data, label noise may manifest itself in various forms. Therefore, at the outset we need to clarify the intended meaning of label noise in this paper and demarcate the scope of this study to the extent possible.

To begin with, it should be clear that we are only interested in label noise, and not data/measurement noise. Specifically, consider a set $\{x_i, y_i\}$ of medical images, x_i, and their corresponding labels, y_i. Although x_i may include measurement noise, that is not the focus of this review. We are only interested in the noise in the label, y_i. Typically, the label y is a discrete variable and can be either an image-wise label, such as in classification problems, or a pixel/voxel-wise label, such as in dense segmentation. Moreover, in this paper we are only concerned with labeled data. Semi-supervised methods are methods that use both labeled and unlabeled training data. Many semi-supervised methods synthesize (noisy) labels for unlabeled data, which are then used for training. Such studies fall within the scope of this study if they use novel or sophisticated methods to handle noisy synthesized labels. Another form of label imperfection that is becoming more common in medical image datasets is when there is only image-level label, and no pixel-level annotations are available (Wang et al., 2017b; Irvin et al., 2019). This type of label is referred to as weak label and is used by methods that are termed weakly supervised learning or multiple-instance learning methods. This type of label imperfection is also beyond the scope of this study. Luckily, there are recent review articles that cover these types of label imperfections. Semi-supervised learning, multiple-instance learning, and transfer learning in medical image analysis have been reviewed in (Cheplygina et al., 2019). Focusing only on medical image segmentation, another recent paper reviewed methods for dealing with scarce and imperfect annotations in general, including weak and sparse annotations (Tajbakhsh et al., 2019).

The organization of this article is as follows. In Section 2 we briefly describe methods for handling label noise in classical (i.e., pre-deep learning) machine learning. In Section 3 we review studies that have dealt with label noise in deep learning. Then, in Section 4 we take a closer look into studies that have trained deep learning models on medical image datasets with noisy labels. Section 5 contains our experimental results with three medical image datasets, where we investigate the impact of label noise and the potential of techniques and remedies for dealing with noisy labels in deep learning. Conclusions are presented in Section 6.

2. Label noise in classical machine learning

Learning from noisy labels has been a long-standing challenge in machine learning (Frénay and Verleysen, 2013; García et al., 2015). Studies have shown that the negative impact of label noise on the performance of machine learning methods can be more significant than that of measurement/feature noise (Zhu and Wu, 2004; Quinlan, 1986). The complexity of label noise distribution varies greatly depending on the application. In general, label noise can be of three different types: class-independent (the simplest case), class-dependent, and class and feature-dependent (potentially much more complicated). Most of the methods that have been proposed to handle noisy labels in classical machine learning fall into one of the following three categories (Frénay and Verleysen, 2013):

1. Methods that focus on model selection or design. Fundamentally, these methods aim at selecting or devising models that are more robust to label noise. This may include selecting the model, the loss function, and the training procedures. It has been known that the impact of label noise depends on the type and design of the classifier model. For example, naive Bayes and random forests are more robust than other common classifiers such as decision trees and support vector machines (Nettleton et al., 2010; Folleco et al., 2008), and that boosting can exacerbate the impact of label noise (Abellán and Masegosa, 2010; McDonald et al., 2003; Long and Servedio, 2010), whereas bagging is a better way of building classifier ensembles in the presence of significant label noise (Dietterich, 2000). Studies have also shown that 0–1 label loss is more robust than smooth alternatives (e.g., exponential loss, log-loss, squared loss, and hinge-loss) (Manwani and Sastry, 2013; Patrini et al., 2016). Other studies have modified standard loss functions to improve their robustness to label noise, for example by making the hinge loss negatively unbounded as proposed in (Van Rooyen et al., 2015). Furthermore, it has been shown that proper reweighting of training samples can improve the robustness of many loss functions to label noise (Liu and Tao, 2015; Natarajan et al., 2013).

2. Methods that aim at reducing the label noise in the training data. A popular approach is to train a classifier using the available training data with noisy labels or a small dataset with clean labels and identify mislabeled data samples based on the predictions of this classifier (Segata et al., 2009). Voting among an ensemble of classifiers has been shown to be an effective method for this purpose (Brodley et al., 1996; Sluban et al., 2010). K-nearest neighbors (KNN)-based analysis of the training data has also been used to remove mislabeled instances (Wilson and Martinez, 1997, 2000). More computationally intensive approaches include those that identify mislabeled instances via their impact on the training process. For example, (Zhang et al., 2009; Malossini et al., 2006) propose to detect mislabeled instances based on their impact on the classification of other instances in a leave-one-out framework. Some methods are similar to outlier-detection techniques. They define some criterion to reflect the classification uncertainty or complexity of a data point and prune those training instances that exceed a certain threshold on that criterion (Gamberger et al., 2000; Sun et al., 2007).

3. Methods that perform classifier training and label noise modeling in a unified framework. Methods in this class can overlap with those of the two aforementioned classes. For instance, some methods learn to denoise labels or to identify and down-weight samples that are more likely to have incorrect labels in parallel with classifier training. Some methods in this category improve standard classifiers such as support vector machines, decision trees, and neural networks by proposing novel training procedures that are more robust to label noise (Khardon and Wachman, 2007; Lin et al., 2004). Alternatively, different forms of probabilistic models have been used to model the label noise and thereby improve various classifiers (Kaster et al., 2010; Kim and Ghahramani, 2006).

3. Deep learning with noisy labels

Deep learning models typically require much more training data than the more traditional machine learning models do. In many applications the training data are labeled by non-experts or even by automated systems. Therefore, the label noise level is usually higher in these datasets compared with the smaller and more carefully prepared datasets used in classical machine learning.

Many recent studies have demonstrated the negative impact of label noise on the performance of deep learning models and have investigated the nature of this impact. It has been shown that, even with regularization, current convolutional neural network (CNN) architectures used for image classification and trained with standard stochastic gradient descent (SGD) algorithms can fit very large training datasets with completely random labels (Zhang et al., 2016). Obviously, the test performance of such a model would be similar to random assignment because the model has only memorized the training data. Given such an enormous representation capacity, it may seem surprising that large deep learning models have achieved record-breaking performance in many real-world applications. The answer to this apparent contradiction, as suggested by (Arpit et al., 2017), is that when deep learning models are trained on typical datasets with mostly correct labels, they do not memorize the data. Instead, at least in the beginning of training, they learn the dominant patterns shared among the data samples. It has been conjectured that this behavior is due to the distributed and hierarchical representation inherent in the design of the state of the art deep learning models and the explicit regularization techniques that are commonly used when training them (Arpit et al., 2017). One study empirically confirmed these ideas by showing that deep CNNs are robust to strong label noise (Rolnick et al., 2017). For example, in hand-written digit classification on the MNIST dataset, if the label accuracy was only 1% higher than random labels, a classification accuracy of 90% was achieved at test time. A similar behavior was observed on more challenging datasets such as CIFAR100 and ImageNet, albeit at much lower label noise levels. This suggests strong learning (as opposed to memorization) tendency of large CNNs. However, somewhat contradictory results have been reported by other studies. For face recognition, for example, it has been found that label noise can have a significant impact on the accuracy of a CNN and that training on a smaller dataset with clean labels is better than training on a much larger dataset with significant label noise (Wang et al., 2018a). The theoretical reasoning and experiments in (Chen et al., 2019b) suggested a quadratic relation between the label noise ratio in the training data and test error.

Although the details of the interplay between memorization and learning mentioned above is not fully understood, experiments in (Arpit et al., 2017) suggest that this trade-off depends on the nature and richness of the data, amount of label noise, model architecture, as well as training procedures including regularization. Ma et al. (2018) show that the local intrinsic dimensionality of the features learned by a deep learning model depends on the label noise. Formal definition of local intrinsic dimensionality is given by Houle (2017). It quantifies the dimensionality of the underlying data manifold. More specifically, given a data point x_i, local intrinsic dimensionality of the data manifold is a measure of the rate of encounter of other data points as the radius of a ball centered at x_i grows. Ma et al. (2018) showed that

when training on data with noisy labels, the local dimensionality of the features initially decreases as the model learns the dominant patterns in the data. As the training proceeds, the model begins to overfit to the data samples with incorrect labels and the dimensionality starts to increase. Drory et al. (2018) establish an analogy between the performance of deep learning models and KNN under label noise. Using this analogy, they empirically show that deep learning models are highly sensitive to label noise that is concentrated, but that they are less sensitive when the label noise is spread across the training data.

The theoretical work on understanding the impact of label noise on the training and generalization of deep neural networks is still ongoing (Martin and Mahoney, 2017). On the practical side, many studies have shown the negative impact of noisy labels on the performance of these models in real-world applications (Yu et al., 2017; Moosavi-Dezfooli et al., 2017; Speth and Hand, 2019). Not surprisingly, therefore, this topic has been the subject of much research in recent years. We review some of these studies below, organizing them under six categories. As this categorization is arbitrary, there is much overlap among the categories and some studies may be argued to belong to more than one category.

Table 1 shows a summary of the methods we have reviewed. For each category of methods, we have shown a set of representative studies along with the applications addressed in the experimental results of the original paper. For each category of methods, we have also suggested some applications in medical image analysis that can benefit from the methods developed in those papers.

3.1. Label cleaning and pre-processing

The methods in this category aim at identifying and either fixing or discarding training data samples that are likely to have incorrect labels. This can be done either prior to training or iteratively in parallel with the training of the main model. Vo et al. (2015) proposed supervised and unsupervised image ranking methods for identifying correctly-labeled images in a large corpus of images with noisy labels. The proposed methods were based on matching each image with a noisy label to a set of representative images with clean labels. This method improved the classification accuracy by 4–6% over the baseline CNN models on three datasets. Veit et al. (2017) trained two CNNs in parallel using a small dataset with correct labels and a large dataset with noisy labels. The two CNNs shared the feature extraction layers. One CNN used the clean dataset to learn to clean the noisy dataset, which was used by the other CNN to learn the main classification task. Experiments showed that this training method was more effective than training on the large noisy dataset followed by fine-tuning on the clean dataset. Ostyakov et al. (2018) trained an ensemble of classifiers on data with noisy labels using cross-validation and used the predictions of the ensemble as soft labels for training the final classifier.

CleanNet, proposed by Lee et al. (2018), extracts a feature vector from a query image with a noisy label and compares it with a feature vector that is representative of its class. The representative feature vector for each class is computed from a small clean dataset. The similarity between these feature vectors is used to decide whether the label is correct. Alternatively, this similarity can be used to assign weights to the training samples, which is the method proposed for image classification by Lee et al. (2018). Han et al. (2019)

improved upon CleanNet in several ways. Most importantly, they removed the need for a clean dataset by estimating the correct labels in an iterative framework. Moreover, they allowed for multiple prototypes (as opposed to only one in CleanNet) to represent each class. Both of these studies reported improvements in image classification accuracy of 1–5% depending on the dataset and noise level.

A number of proposed methods for label denoising are based on classification confidence. Rank Pruning, proposed by Northcutt et al. (2017), identifies data points with confident labels and updates the classifier using only those data points. This method is based on the assumption that data samples for which the predicted probability is close to one are more likely to have correct labels. However, this is not necessarily true. In fact, there is extensive recent work showing that standard deep learning models are not "well calibrated" (Guo et al., 2017; Lakshminarayanan et al., 2017). A classifier is said to have a calibrated prediction confidence if its predicted class probability indicates its likelihood of being correct. For a perfectly-calibrated classifier, $P\left(y_{\text{predicted}} = y_{\text{true}} | \hat{p} = p\right) = p$. It has been shown that deep learning models produce highly over-confident predictions. Many studies in recent years have aimed at improving the calibration of deep learning models (Gal and Ghahramani, 2015; Kendall and Gal, 2017; Pawlowski et al., 2017). In order to reduce the reliance on classifier calibration, the Rank Pruning algorithm, as its name suggests, ranks the data samples based on their predicted probability and removes the data samples that are least confident. In other words, Rank Pruning assumes that the predicted probabilities are accurate in the relative sense needed for ranking. In light of what is known about poor calibration of deep learning models, this might still be a strong assumption. Nonetheless, Rank Pruning was shown empirically to lead to substantial improvements in image classification tasks in the presence of strong label noise. Identification of incorrect labels based on prediction confidence was also shown to be highly effective in extensive experiments on image classification by Ding et al. (2018), improving the classification accuracy on CIFAR-10 by up to 20% in the presence of very strong label nosie. Köhler et al. (2019) proposed an iterative label noise filtering approach based on similar concepts as Rank Pruning. This method estimates prediction uncertainty (using such methods as Deep Ensembles (Lakshminarayanan et al., 2017) or Monte-Carlo dropout (Kendall and Gal, 2017)) during training and relabels data samples that are likely to have incorrect labels.

A different approach, is proposed by Gao et al. (2017). In this approach, termed deep label distribution learning (DLDL), the initial noisy labels are smoothed to obtain a "label distribution", which is a discrete distribution for classification problems. The authors propose methods for obtaining this label distribution from one-hot labels for several applications including multi-class classification and semantic segmentation. For semantic segmentation, for example, a simple kernel smoothing of the segmentation mask is suggested to account for unreliable boundaries. Once this smooth label is obtained, the deep learning model is trained by minimizing the Kullback-Leibler (KL) divergence between the model output and the smooth noisy label. Label smoothing is a well-know trick for improving the test performance of deep learning models (Szegedy et al., 2016; Müller et al., 2019). The DLDL approach was improved by Yi and Wu (2019), where the authors introduced a cross-entropy-based loss term to encourage closeness of estimated labels and

the initial noisy labels and proposed a back-propagation method to iteratively update the initial label distributions as well.

Ratner et al. (2016) used a generative model to model labeling of large datasets used in deep learning and proposed a label denoising method under this scenario. Zhou et al. (2017) proposed a GAN for removing label noise from synthetic data generated to train a CNN. This method was shown to be highly effective in removing label noise and improving the model performance. GANs were used to generate a training dataset with clean labels from an initial dataset with noisy labels by Chiaroni et al. (2019).

3.2. Network architecture

Several studies have proposed adding a "noise layer" to the end of deep learning models. The noise layer proposed by Sukhbaatar et al. (2014) is equivalent to multiplication with the transition matrix between noisy and true labels. The authors developed methods for learning this matrix in parallel with the network weights using error back-propagation. A similar noise layer was proposed by Thekumparampil et al. (2018) for training a generative adversarial network (GAN) under label noise. Sukhbaatar and Fergus (2014) proposed methods for estimating the transition matrix from either a clean or a noisy dataset. Reductions of up to 3.5% in classification error were reported on different datasets. A similar noise layer was proposed by Goldberger and Ben-Reuven (2016), where the authors proposed an EM-type method for optimizing the parameters of the noise layer. Importantly, the authors extended their model to the more general case where the label noise also depends on image features. This more complex case, however, could not be optimized with EM and a back-propagation method was exploited instead. Bekker and Goldberger (2016) used a combination of EM and error back-propagation for end-to-end training with a noise layer. Jindal et al. (2016) suggested that aggressive dropout regularization (with a rate of 90%) can improve the effectiveness of such noise layers.

Focusing on noisy labels obtained from multiple annotators, Tanno et al. (2019) proposed a simple and effective method for estimating the correct labels and annotator confusion matrices in parallel with CNN training. The key observation was that, in order to avoid the ambiguity in simultaneous estimation of true labels and annotator confusion matrices, the traces of the confusion matrices had to be penalized. The entire model including the CNN weights and confusion matrices were learned via SGD. The method was shown to be highly effective in estimating annotator confusion matrices for various annotator types including inaccurate and adversarial ones. Improvements of 8–11% in image classification accuracy were reported compared to the best competing methods.

A number of studies have integrated different forms of probabilistic graphical models into deep neural networks to handle label noise. Xiao et al. (2015) proposed a graphical model with two discrete latent variables y and z, where y was the true label and z was a one-hot vector of size 3 that denoted whether the label noise was zero, class-independent, or class-conditional. Two separate CNNs estimated y and z, and the entire model was optimized in an EM framework. The method required a small dataset with clean labels. The authors showed significant gains compared with baseline CNNs in image classification from large datasets with noisy labels. Vahdat (2017) employed an undirected graphical model to learn the

relationship between correct and noisy labels. The model allowed incorporation of domain-specific sources of information in the form of joint probability distribution of labels and hidden variables. Their method improved the classification accuracy of baseline CNNs by up to 3% on three different datasets. For image classification, Misra et al. (2016) proposed to jointly train two CNNs to disentangle the object presence and relevance in a framework similar to the graphical model-based methods described above. Model parameters and true labels were estimated using SGD. A more elaborate model was proposed by Yao et al. (2018), where an additional latent variable was introduced to model the trustworthiness of the noisy labels.

3.3. Loss functions

A large number of studies keep the model architecture, training data, and training procedures largely intact and only change the loss function (Izadinia et al., 2015). Ghosh et al. (2017) studied the conditions for robustness of a loss function to label noise for training deep learning models. They showed that mean absolute value of error, MAE, (defined as the ℓ_1 norm of the difference between the true and predicted class probability vectors) is tolerant to label noise. This means that, in theory, the optimal classifier can be learned by training with basic error back-propagation. They showed that cross-entropy and mean square error did not possess this property. For a multi-class classification problem, denoting the vector of true and predicted probabilities with $p(y = j|x)$ and $\hat{p}(y = j \mid x)$, respectively, the cross-entropy loss function is defined as $L_{\mathrm{CE}} = \sum_j p(y = j \mid x) \log \hat{p}(y = j \mid x)$. The MAE loss is defined as $L_{\mathrm{MAE}} = \sum_j |p(y = j \mid x) - \hat{p}(y = j \mid x)|$. As opposed to cross-entropy that puts more emphasis on hard examples (desirable for training with clean labels), MAE tends to treat all data points more equally. However, a more recent study argued that because of the stochastic nature of the optimization algorithms used to train deep learning models, training with MAE down-weights difficult samples with correct labels, leading to significantly longer training times and reduced test accuracy (Zhang and Sabuncu, 2018). The authors proposed their own loss functions based on Box-Cox transformation to combine the advantages of MAE and cross-entropy. Similarly, Wang et al. (2019b) analyzed the gradients of cross-entropy and MAE loss functions to show their weaknesses and advantages. They proposed an improved MAE loss function (iMAE) that overcame MAE's poor sample weighting strategy. Specifically, they showed that the ℓ_1 norm of the gradient of L_{MAE} with respect to the logit vector was equal to $4p(y|x)(1 - p(y|x))$, leading to down-weighting of difficult but informative data samples. To fix this shortcoming, they suggested to transform the MAE weights nonlinearly with a new weighting defined as $\exp(T\,p(y|x))(1 - p(y|x))$, where the hyperparameter T was set equal to 8 for training data with noisy labels. In image classification experiments on the CIFAR-10 dataset, compared with cross-entropy and MAE losses, their proposed iMAE loss improved the classification by approximately 1–5% when label noise was low and up to 25% when label noise was very high. In another experiment on person reidentification in video, iMAE improved the mean average precision by 13% compared with cross-entropy.

Thulasidasan et al. (2019) proposed modifying the cross-entropy loss function to enable abstention. Their proposed modification allowed the model to abstain from making a prediction on some data points at the cost of incurring an abstention penalty. They showed

that this policy could improve the classification performance on both random label noise as well as systematic datadependent label noise. Rusiecki (2019) proposed a trimmed cross-entropy loss based on trimmed absolute value criterion. Their central assumption is that, with a well-trained model, data samples with wrong labels result in high loss values. Hence, their proposed loss function simply ignores the training samples with the largest loss values. Note that the central idea in (Rusiecki, 2019) (of down-weighting hard data samples) seems to run against many prevalent techniques in machine learning such as boosting (Freund et al., 1999), hard example mining (Shrivastava et al., 2016), and loss functions such as focal loss (Lin et al., 2017), that steer the training process to focus on hard examples. This is because when the training labels are correct, data points with high loss values constitute the hard examples that the model has not learned yet. Hence, focusing on those examples generally helps improve the model performance. On the other hand, when there is significant label noise, assuming that the model has attained a decent level of accuracy, data points with unusually high loss values are likely to have wrong labels. This idea is not restricted to (Rusiecki, 2019) and it is an idea that is shared by many methods reviewed in this article. This paradigm shift is a good example of the dramatic effect of label noise on the machine learning methodology.

Patrini et al. (2017) proposed two simple ways of improving the robustness of a loss function to label noise for training deep learning models. The proposed correction methods are based on the error confusion matrix T, defined as $T_{i,j} = p(\tilde{y} = e^j \mid y = e^i)$, where $\tilde{y}$ and y are the noisy and true labels, respectively. Assuming T is non-singular, one of the proposed correction strategies is $l_{\text{corr}}(\hat{p}(y \mid x)) = T^{-1} l(\hat{p}(y \mid x))$. This correction is a linear weighting of the loss values for each possible label, where the weights, given by T, are the probability of the true label given the observed label. The authors name this correction method “backward correction” because it is intuitively equivalent to going one step back in the noise process described by the Markov chain represented by T. The alternative approach, named forward correction, is based on correcting the model predictions and only applies to composite proper loss functions (Reid and Williamson (2010)), which include cross-entropy. The corrected loss is defined as $l_{\text{corr}}(h(x)) = l(T^{\mathrm{T}} \psi^{-1}((h(x))))$, where h is the vector of logits, and ψ^{-1} is the inverse of the *link function* for the loss function in consideration, which is the standard softmax for cross-entropy loss. The authors show that both these corrections lead to unbiased loss functions, in the sense that $\forall x E_{\tilde{y} \mid x} l_{\text{corr}} = E_{y \mid x} l$. They also propose a method for estimating T from noisy data and show that their methods lead to performance improvements on a range of computer vision problems and deep learning models. Similar methods have been proposed by Hendrycks et al. (2018), and Boughorbel et al. (2018), where it is suggested to use a small dataset with clean labels to estimate T. Boughorbel et al. (2018) alternate between training on a clean dataset with a standard loss function and training on a larger noisy dataset with the corrected loss function. Mnih and Hinton (2012) proposed a similar loss function based on penalizing the disagreement between the predicted label and the posterior of the true label.

3.4. Data re-weighting

Broadly speaking, these methods aim at down-weighting those training samples that are more likely to have incorrect labels. Ren et al. (2018) proposed to weight the training data using a meta-learning approach. That method required a separate dataset with clean labels, which was used to determine the weights assigned to the training data with noisy labels. Simply put, it optimized the weights on the training samples by minimizing the loss on the clean validation data. The authors showed that this weighting scheme was equivalent to assigning larger weights to training data samples that were similar to the clean validation data in terms of both the learned features and optimization gradient directions. Experiments showed that this method improved upon baseline methods by 0.5% and 3% on CIFAR-10 and CIFAR-100 with only 1000 images with clean labels. More recently, Wang et al. (2019a) proposed to re-weight samples by optimization gradient re-scaling. The underlying idea, again, is to give larger weights to samples that are easier to learn, hence more likely to have correct labels. Pumpout, proposed by Han et al. (2018a), is also based on gradient scaling. The authors propose two methods for identifying data samples that are likely to have incorrect lables. One of their methods is based on the assumption that data samples with incorrect labels are likely to display unusually high loss values. Their second method is based on the value of the backward-corrected loss (Patrini et al., 2017); they suggest that the condition $\mathbf{1}^T T^{-1} l(\hat{p}(y \mid x)) < 0$ indicates data samples with incorrect labels. For training data samples that are suspected of having incorrect labels, the gradients are scaled by $-\gamma$, where $0 < \gamma < 1$. In other words, they perform a scaled gradient *ascent* on the samples with incorrect labels. In several experiments, including image classification with MNIST and CIFAR-10 datasets, they show that their method avoids fitting to incorrect labels and reduces the classification error by up to 40%.

Shen and Sanghavi (2019) proposed a training strategy that can be interpreted as a form of data re-weighting. In each training epoch, they remove a fraction of the data for which the loss is the largest, and update the model parameters to minimize the loss function on the remaining training data. This method assumes that the model gradually converges towards a good classifier such that the mis-labeled training samples exhibit unusually high loss values as training progresses. The authors proved that this simple approach learns the optimal model in the case of generalized linear models. For deep CNNs that are highly nonlinear, they empirically showed the effectiveness of their method on several image classification tasks. As in the case of this method, there is often a close connection between some of the data re-weighting methods and methods based on robust loss functions. Shu et al. (2019) built upon this connection and developed it further by proposing to learn a data re-weighting scheme from data. Instead of assuming a pre-defined weighting scheme, they used a multi-layer perceptron (MLP) model with a single hidden layer to learn a suitable weighting strategy for the task and the dataset at hand. The MLP in this method is trained on a small dataset with clean labels. Experiments on datasets with unbalanced and noisy labels showed that the learned weighting scheme conformed with those proposed in other studies. Specifically, for data with noisy labels the model learned to down-weight samples with large loss functions, the opposite of the form learned for datasets with unbalanced classes. One can argue that this observation empirically justifies the general trend towards down-weighting training samples with large loss values when training with noisy labels.

A common scenario involves labels obtained from multiple sources or annotators with potentially different levels of accuracy. This is a heavily-researched topic in machine learning. A simple approach to tackling this scenario is to use expectation-maximization (EM)-based methods such as (Warfield et al., 2004; Raykar et al., 2010) to estimate the true labels and then proceed to train the deep learning model using the estimated labels. Khetan et al. (2017) proposed an iterative method, whereby model predictions were used to estimate annotator accuracy and then these accuracies were used to train the model with a loss function that properly weighted the label from each annotator. The model was updated via gradient descent, whereas annotator confusion matrices were optimized with an EM method. By contrast, Tanno et al. (2019) estimated the network weights as well as annotator confusion matrices via gradient descent.

3.5. Data and label consistency

It is usually the case that the majority of the training data samples have correct labels. Moreover, there is considerable correlation among data points that belong to the same class (or the features computed from them). These correlations can be exploited to reduce the impact of incorrect labels. A typical example is the work of Lee et al. (2019), where the authors consider the correlation of the features learned by a deep learning model. They suggest that the features learned by various layers of a deep learning model on data samples of the same class should be highly correlated (i.e., clustered). Therefore, they propose training an ensemble of generative models (in the form of linear discriminant classifiers) on the features of the penultimate layer and possibly also other layers of a trained deep learning model. They show significant improvements in classification accuracy on several network architectures, noise levels, and datasets. On CIFAR-10 dataset, they report classification accuracy improvements of 3–20%, with larger improvements for higher label noise levels, compared with a baseline CNN. On more difficult datasets such as CIFAR-100 and SVHN, smaller but still significant improvements of approximately 3–10% are reported. Another example is the work of Zhang et al. (2019), where the authors proposed a method to leverage the multiplicity of data samples with the same (noisy) label in each training batch. All samples with the same label were fed into a light-weight neural network model that assigned a confidence weight to each sample based on the probability of it having the correct label. These weights were used to compute a representative feature vector for that class, which was then used to train the main classification model. Compared with other competing methods, 1–4% higher classification accuracies were reprorted on several datasets. For face identification, Speth and Hand (2019) proposed feature embedding to detect data samples with incorrect labels. Their proposed verification framework used a multi-label Siamese CNN to embed a data point in a lower-dimensional space. The distance of the point to a set of representative points in this lower-dimensional space was used to determine whether the label was incorrect.

Azadi et al. (2015) propose a method that they name “auxiliary image regularization”. Their method requires a small set of auxiliary images with clean labels in addition to the main training dataset with noisy labels. The core idea of auxiliary image regularization is to encourage representation consistency between training images (with noisy labels) and auxiliary images (with known correct labels). For this purpose, their proposed loss function

includes a term based on group sparsity that encourages the features of a training image to be close to those of a small number of auxiliary images. Clearly, the auxiliary images should include good representatives of all expected classes. This method improved the classification accuracy by up to 8% on ImageNet dataset. Chen et al. (2019a) proposed a manifold regularization technique that penalized the KL divergence between the class probability predictions of similar data samples. Because searching for similar samples in high-dimensional data spaces was challenging, they suggested using data augmentation to synthesize similar inputs. They reported 1–3% higher classification accuracy compared with several alternative methods on CIFAR-10 and CIFAR-100. Li et al. (2017a) proposed BundleNet, where multiple images with the same (noisy) labels were stacked together and fed as a single input to the network. Even though the authors do not provide a clear justification of their method and its difference with standard mini-batch training, they show empirically that their method improves the accuracy on image classification with noisy labels. Wang et al. (2018c) used the similarity between images in terms of their deep features in an iterative framework to identify and down-weight training samples that were likely to have incorrect labels. Consistency between predicted labels and data (e.g., images or features) was exploited by Reed et al. (2014). The authors considered the true label as a hidden variable and proposed a model that simultaneously learned the relation between true and noisy labels (i.e., label noise distribution) and an auto-encoder model to reconstruct the data from the hidden variables. They showed improved performance in detection and classification tasks.

3.6. Training procedures

The methods in this category are very diverse. Some of them are based on well-known machine learning methods such as curriculum learning and knowledge distillation, while others focus on modifying the settings of the training pipeline such as learning rate and regularization.

Several methods based on curriculum learning have been proposed to combat label noise. Curriculum learning, first proposed by Bengio et al. (2009), is based on training a model with examples of increasing complexity or difficulty. In the method proposed by Jiang et al. (2017), an LSTM network called Mentor-Net provides a curriculum, in the form of weights on the training samples, to a second network called Student-Net. On CIFAR-100 and ImageNet with various label noise levels, their method improved the classification accuracy by up to 20% and 2%, respectively. Guo et al. (2018) proposed another method based on curriculum learning, named CurriculumNet, for training a model from massive datasets with noisy labels. This method first clusters the training data in some feature space and identifies samples that are more likely to have incorrect labels as those that fall in low-density clusters. The data are then sequentially presented to the main CNN model to be trained. This technique achieved good results on several datasets including ImageNet. The Self-Error-Correcting CNN proposed by Liu et al. (2017) is based on similar ideas; the training begins with noisy labels but as the training proceeds the network is allowed to change a sample's label based on a confidence policy that gives more weight to the network predictions with more training.

Li et al. (2017b) adopted a knowledge distillation approach (Hinton et al., 2015) to train an auxiliary model on a small dataset with clean labels to guide the training of the main model on a large dataset with noisy labels. In brief, their approach amounts to using a pseudo-label, which is a convex combination of the noisy label and the label predicted by the auxiliary model. To reduce the risk of overfitting the auxiliary model on the small clean dataset, the authors introduced a knowledge graph based on the label transition matrix. Reed et al. (2014) also proposed using a convex combination of the noisy labels and labels predicted by the model at its current training stage. They suggested that as the training proceeds, the model becomes more accurate and its predictions can be weighted more strongly, thereby gradually forgetting the original incorrect labels. Zhong et al. (2019) used a similar approach for face identification. They first trained their model on a small dataset with less label noise and then fine-tuned it on data with stronger label noise using an iterative label update strategy similar to that explained above. Their method led to improvements of up to 2% in face recognition accuracy. Following a similar training strategy, Köhler et al. (2019) suggested that there is a point (e.g., a training epoch) when the model learns the true data features and is about to fit to the noisy labels. They proposed two methods, one based on the predictions on a clean dataset and another based on prediction uncertainty measures, to identify that stage in training. The output of the model at that stage can be used to fix the incorrect labels.

A number of studies have proposed methods involving joint training of more than one model. For example, one work suggested simultaneously training two separate but identical networks with random initialization, and only updating the network parameters when the predictions of the two networks differed Malach and Shalev-Shwartz (2017). The idea is that when training with noisy labels, the model starts by learning the patterns in data samples with correct labels. Later in training, the model will struggle to overfit to samples with incorrect labels. The proposed method hopes to reduce the impact of label noise because the decision as to whether or not to update the model is made based on the predictions of the two models and independent of the noisy label. In other words, on data with incorrect labels both models are likely to produce the same prediction, i.e., they will predict the correct label. On easy examples with correct labels, too, both models will make the same (correct) prediction. On hard examples with correct labels, on the other hand, the two models are more likely to disagree. Hence, with the proposed training strategy, the data samples that will be used in later stages of training will shrink to the hard data samples with correct labels. This strategy also improves the computational efficiency since it performs many updates at the start of training but avoids unnecessary updates on easy data samples once the models have sufficiently converged to predict the correct label on those samples. This idea was developed into co-teaching Han et al. (2018b), whereby the two networks identified label-noise-free samples in their mini-batches and shared the update information with the other network. The authors compare their method with several state of the art techniques including Mentor-Net (Jiang et al. (2017). Their method outperformed competing methods in most experiments, while narrowly underperforming in some experiments. Co-teaching was further improved in Yu et al. (2019b), where the authors suggested to focus the training on data samples with lower loss values in order to reduce the risk of training on data with incorrect labels. Along the same lines, Li et al. (2019) proposed a meta-learning objective

that encouraged consistent predictions between a student model trained on noisy labels and a teacher model trained on clean labels. The goal was to train the student model to be tolerant to label noise. Towards this goal, artificial label noise was added on data with correct labels to train the student model. The student model was encouraged to be consistent with the teacher model using a meta-objective in the form of the KL divergence between prediction probabilities. Their method outperformed several competing methods by 1–2% on CIFAR-10 and Clothing1M datasets.

Experiments in Chen et al. (2019b) showed that co-teaching was less effective as the label noise increased. Instead, the authors showed that selecting the data samples with correct labels using cross-validation was more effective. In their proposed approach, the training data was divided into two folds. The model was iteratively trained on one fold and tested on the other. Data samples for which the predicted and noisy labels agreed were assumed to have the correct label and were used in the next training epoch. One study proposed to learn the network parameters by optimizing the joint likelihood of the network parameters and true labels Tanaka et al. (2018). Compared with standard training with cross-entropy loss, this method improved the classification accuracy on CIFAR-10 by 2% with low label noise rate to 17% when label noise rate was very high.

Some studies have suggested modifying the learning rate, batch size, or other settings in the training methodology. For example, for applications where multiple datasets with varying levels of label noise are available, Song et al. (2015) have proposed training strategies in terms of the order of using different datasets during training and proper learning rate adjustments based on the level of label noise in each dataset. Assuming that separate clean and noisy datasets are available, the same study has shown that using different learning rates for training with noisy and clean samples can improve the performance. It has also shown that the optimal ordering of using the two datasets (i.e., whether to train on the noisy dataset or the clean dataset first) depends on the choice of the learning rate. It has also been suggested that when label noise is strong, the effective batch size decreases, and that batch size should be increased with a proper scaling of the learning rate (Rolnick et al., 2017). Sukhbaatar and Fergus (2014) proposed to include samples from a noisy dataset and a clean dataset in each training mini-batch, giving higher weights to the samples with clean labels.

Mixup is a less intuitive but simple and effective method (Zhang et al., 2017). It synthesizes new training data points and labels via a convex combination of pairs of training data points and their labels. More specifically, given two randomly selected training data and label pairs (x_i, y_i) and (x_j, y_j), a new training data point and label are synthesized as $\tilde{x} = \lambda x_i + (1 - \lambda)x_j$ and $\tilde{y} = \lambda y_i + (1 - \lambda)y_j$, where $\lambda \in [0, 1]$ is sampled from a beta distribution. Although mixup is known primarily as a data augmentation and regularization strategy, it has been shown to be remarkably effective for combatting label noise. Compared with basic emprirical risk minimization on CIFAR-10 dataset with different levels of label noise, mixup reduced the classification error by 6.5–12.5%. The authors argue that the reason for this behavior is because interpolation between datapoints makes memorization on noisy labels, as observed in (Zhang et al., 2016), more difficult. In other words, it is easier for the network to learn the linear iterpolation between datapoints with correct labels than to memorize the interploation

when labels are incorrect. The same idea was successfully used in video classification by Ostyakov et al. (2018).

For object boundary segmentation, two studies proposed to improve noisy labels in parallel with model training (Yu et al., 2018; Acuna et al., 2019). This is a task for which large datasets are known to suffer from significant label noise and model performance to be very sensitive to label noise. Both methods consider the true boundary as a latent variable that is estimated in an alternating optimization framework in parallel with model training. One major assumption in Acuna et al. (2019) is the preservation of the length of the boundary during optimization, resulting in a bipartite graph assignment problem. In Yu et al. (2018), a level-set formulation was introduced instead, providing much higher flexibility in terms of the shape and length of the boundary while preserving its topology. Both studies compared their methods with baseline CNNs in terms of F-measure for object edge detection and report impressive improvements. In particular, Acuna et al. (2019) improved upon their baseline CNN by 2–5% on segmentation of different objects. Similarly, Yu et al. (2018) reported improvements of 1–17% compared to a baseline CNN.

4. Deep learning with noisy labels in medical image analysis

In this section, we review studies that have addressed label noise in training deep learning models for medical image analysis. We use the same categorization as in the previous section.

4.1. Label cleaning and pre-processing

For classification of thoracic diseases from chest x-ray scans, Pham et al. (2019) used label smoothing to handle noisy labels. They compared their label smoothing method with simple methods such as ignoring data samples with noisy labels. They found that label smoothing can lead to improvements of up to 0.08 in the area under the receiver operating characteristic curve (AUC).

4.2. Network architectures

The noise layer proposed by Bekker and Goldberger (2016), reviewed above, was used for breast lesion detection in mammograms by Dgani et al. (2018) and slightly improved the detection accuracy.

4.3. Loss functions

To train a network to segment virus particles in transmission electron microscopy images using original annotations that consisted of only the approximate center of each virus, Matuszewski and Sintorn (2018) dilated the annotations with a small and a large structuring element to generate noisy masks for foreground and background, respectively. Consequently, parts of the image in the shape of the union of rings were marked as uncertain regions that were ignored during training. The Dice similarity and intersection-over-union loss functions were modified to ignore those regions. Promising results were reported for both loss functions. Rister et al. (2018) showed that for segmentation of abdominal organs in CT images from noisy training annotations, the intersection-over-union (IOU) loss consistently

outperformed the cross-entropy loss. The mean DSC achieved with the IOU loss was 1–13% higher than the DSC achieved with the cross-entropy loss.

4.4. Data re-weighting

Le et al. (2019) used a data re-weighting method similar to that proposed by Ren et al. (2018) to deal with noisy annotations in pancreatic cancer detection from whole-slide digital pathology images. They trained their model on a large corpus of patches with noisy labels using weights computed from a small set of patches with clean labels. This strategy improved the classification accuracy by 10% compared with training on all patches with clean and noisy labels without re-weighting. For skin lesion classification in dermoscopy images with noisy labels, Xue et al. (2019) used a data re-weighting method that amounted to removing data samples with high loss values in each training batch. This method, which is similar to some of the methods reviewed above such as the method of Shen and Sanghavi (2019), increased the classification accuracy by 2 – 10%, depending on the label noise level.

For segmentation of heart, clavicles, and lung in chest radiographs, Zhu et al. (2019) trained a deep learning model to detect incorrect labels. This model assigned a weight to each sample in a training batch, aiming to down-weight samples with incorrect labels. The main segmentation model was trained in parallel using a loss function that made use of these weights. A pixel-wise weighting was proposed by Mirikharaji et al. (2019) for skin lesion segmentation from highly inaccurate annotations. The method needed a small dataset with correct segmentations alongside the main, larger, dataset with noisy segmentations. For each training image with noisy segmentation, a weight map of the same size was considered to indicate the pixel-wise confidence in the accuracy of the noisy label. These maps were updated in parallel with network parameters with alternating optimization. The authors proposed to optimize the weights on the images in the noisy dataset by reducing the loss on the clean dataset. In essence, the weight on a pixel is increased if that leads to a reduction in the loss on the clean dataset. If increasing the weight on a pixel increases the loss on the clean dataset, that weight is set to zero because the label for that pixel is probably incorrect.

4.5. Data and label consistency

For segmentation of the left atrium in MRI from labeled and unlabeled data, Yu et al. (2019a) proposed training two separate models: a teacher model that produced noisy labels and label uncertainty maps on unlabeled images, and a student model that was trained using the generated noisy labels while taking into account the label uncertainty. The student model was trained to make correct predictions on the clean dataset and to be consistent with the teacher model on noisy labels with uncertainty below a threshold. The teacher model was updated in a moving average scheme involving the weights of the student model.

4.6. Training procedures

For bladder, prostate, and rectum segmentation in MRI, Nie et al. (2018) trained a model on a dataset with clean labels and used it to predict segmentation masks for a separate unlabeled dataset. In parallel, a second model was trained to estimate a confidence map to indicate the regions where the predicted labels were more likely to be correct. The confidence maps were

used to sample the unlabeled dataset for additional training data for the main model. Improvements of approximately 3% in Dice similarity coefficient (DSC) were reported.

Min et al. (2018) employed the ideas proposed by Malach and Shalev-Shwartz (2017) to develop label-noise-robust methods for medical image segmentation. As we reviewed above, the main idea in the method of Malach and Shalev-Shwartz (2017) was to jointly train two separate models and update the models only on the data samples on which the predictions of the two models differed. Instead of considering only the final layer predictions, Min et al. (2018) introduced attention modules at various depths in the networks to use the gradient information at different feature maps to identify and down-weight samples with incorrect labels. They reported promising results for cardiac and glioma segmentation in MRI.

For cystic lesion segmentation in lung CT, Zhang et al. (2018) generated initial noisy segmentations using unsupervised K-means clustering. These segmentations were used to train a CNN. Assuming that the CNN was more accurate than K-means, CNN predictions were used as the training labels for the next epoch. This process was repeated, generating new labels at the end of each training epoch. Experiments showed that the final trained CNN achieved significantly higher segmentation accuracy compared with the K-means method used to generate the initial segmentations. A rather similar method was used for classification of aortic valve malfunctions in MRI by Fries et al. (2019). Using a small dataset of expert-annotated images, simple classifiers based on intensity and shape features were developed. Subsequently, a factor graph-based model was trained to estimate the classification accuracies of these classifiers and to generate pseudo-ground-truth labels on a massive unlabeled dataset. This dataset was then used to train a deep learning classifier. This model significantly outperformed models trained on a small set of expert-labeled images.

5. Experiments

In this section, we present our experiments on three medical image datasets with noisy labels, in which we explored several methods that we implemented, adapted, or developed to analyze and reduce the effect of label noise. Our experiments represent three different machine learning problems, namely, detection, classification, and segmentation. The three datasets associated with these experiments represent three different noise types, namely, label noise due to systematic error by a human annotator, label noise due to inter-observer variability, and error/noise in labels generated by an algorithm (Figure 1). In developing and comparing techniques, our goal was not to achieve the best, state-of-the-art results in each experiment, as that would have required careful design of network architectures, data pre-processing, and training procedures for each problem. Instead, our goal was to show the effects of label noise and the relative effectiveness, merits, and shortcomings of potential methods on common label noise types in medical image datasets.

5.1. Brain lesion detection and segmentation

5.1.1. Data and labels—We used 165 MRI scans from 88 tuberous sclerosis complex (TSC) subjects. Each scan included T1, T2, and FLAIR images. An experienced annotator segmented the lesions in these scans. We then randomly selected 12 scans for accurate annotation and assessment of label noise. Two annotators jointly reviewed these scans in

four separate sessions to find and fix missing or inaccurate annotations. The last reading did not find any missing lesions in any of the 12 scans. Example scans and their annotations are shown in Figure 2. We used these 12 scans and their annotations for evaluation only. We refer to these scans as “the clean dataset”. We used the remaining 153 scans and their imperfect annotations for training. These are referred to as “the noisy dataset”.

In the 12 scans in the clean dataset, 306 lesions were detected in the first reading and 68 lesions in the followup readings, suggesting that approximately 18% of the lesions were missed in the first reading. Annotation error can be modeled as a random variable, in the sense that if the same annotators annotate the same scan a second time (with some time interval) they may not make the same errors. Nonetheless, our analysis shows that smaller or fainter lesions were more likely to be missed. Specifically, Welchs t-tests showed that the lesions that had been missed in the first reading were less dark on the T1 image ($p<0.001$), smaller in size ($p<0.001$), and farther away from the closest lesion ($p=0.004$), compared with lesions that were detected in the first reading. Therefore, in this application the intrinsic limitation of human annotator attention results in systematic errors (noise) in labels.

5.1.2. Methods—For the joint detection and segmentation of lesions in this application, we used a baseline CNN similar to the 3D U-Net (Çiçek et al., 2016). This CNN included four convolutional blocks in each of the contracting and expanding parts. The first convolutional block included 14 feature maps, which increased by a factor of 2 in subsequent convolutional blocks, resulting in the coarsest convolutional block with 112 feature maps. Each convolutional block processed its feature maps with additional convolutional layers with residual connections. All convolutional operations were followed by ReLU activation. The CNN worked on blocks of size 64^3 voxels and it was applied in a sliding-window fashion to process an image. In addition, since this application can be regarded as a detection task, we also used a method based on Faster-RCNN (Ren et al., 2015), where we used a 3D U-Net architecture for the backbone of this method. To train Faster-RCNN, we followed the training methodology of Ren et al. (2015), but made changes to adapt it to 3D images. Based on the distribution of lesion size in our data, we used five different anchor sizes and three different aspect ratios, for a total of 15 anchors in Faster-RCNN. The smallest and largest anchors were $3 \times 3 \times 7$ mm^3 and $45 \times 45 \times 61$ mm^3, respectively. Our evaluation was based on two-fold subject-wise cross-validation, each time training the model on data from approximately half of the subjects and testing on the remaining subjects. Following the latest recommendations in the literature on lesion detection applications (Carass et al., 2017; Commowick et al., 2018; Hashemi et al., 2019), our main evaluation criterion was lesion-count F1 score; but since this is considered a joint segmentation and detection task, we also computed DSC when applicable (i.e., for the 3D U-Net). It is noteworthy that due to the criteria that are used in diagnosis/prognosis and disease modifying treatments, lesion-count measures such as lesion-count F1-score have been considered more appropriate performance measures for lesion detection and segmentation algorithms compared to DSC (Commowick et al., 2018; Hashemi et al., 2019).

The methods developed, implemented, and compared in this task include:

- Faster-RCNN trained on noisy labels.

- Faster-RCNN trained on clean data. Same as the above, but evaluated using two-fold cross-validation on the clean data.
- Faster-RCNN trained with MAE loss (Ghosh et al., 2017).
- 3D U-Net CNN trained on noisy labels with DSC loss.
- 3D U-Net CNN trained on clean data. Same as the above, but evaluated using two-fold cross-validation on the clean data.
- 3D U-Net CNN trained with MAE loss (Ghosh et al., 2017).
- 3D U-Net CNN trained with iMAE loss (Wang et al., 2019b).
- 3D U-Net CNN with data re-weighting. In this method, we ignored data samples with very high loss values. We kept the mean and standard deviation of the losses of the 100 most recent training samples. If the loss for a training sample was higher than 1.5 standard deviations of the mean, the network weights were not updated on that sample. To the best of our knowledge, such a method has not been proposed for brain lesion detection/segmentation prior to this work.
- Iterative label cleaning. This is a novel technique that we have developed for this application. We first trained a random forest classifier to distinguish the true lesions missed by the annotator from the false positive lesions in CNN predictions. This classification was based on six lesion features: mean image intensity in T1, T2, and FLAIR, lesion size, distance to the closest lesion, and mean prediction uncertainty, where uncertainty was computed using the methods of Kendall and Gal (2017). Then, during training of the CNN on the noisy dataset, after each training epoch the random forest classifier was applied on the CNN-detected lesions that were not present in the original noisy labels. Lesions that were classified as true lesions were added to the noisy labels. Hence, this method iteratively improved the noisy labels in parallel with CNN training.

5.1.3. Results—As shown in Table 2, 3D U-Net achieved higher detection accuracy than Faster-RCNN. Since our focus is on label noise, we discuss the results of experiments with each of these two networks independently. For 3D U-Net, both MAE and iMAE loss functions resulted in lower lesion-count F1 score and DSC, compared with the baseline CNN trained with a DSC loss. However, both MAE and iMAE have been proposed as improvements to the cross-entropy. With a cross-entropy loss, our CNN achieved performance similar to iMAE. Interestingly, for Faster-RCNN, compared with the baseline that was trained with the cross-entropy loss, using the MAE loss did improve the lesion-count F1 score by 0.041. This indicates that such loss functions, initially proposed for classification and detection tasks, may be more useful for lesion detection than for lesion segmentation applications. The data re-weighting method resulted in lesion-count F1 score and DSC that were substantially higher than the baseline CNN. Moreover, iterative label cleaning achieved much higher lesion-count F1 score and DSC than the baseline and outperformed the data re-weighting method too. The increase in the lesion-count F1 score shows that iterative label cleaning improves detection of small lesions. The increase in DSC is also interesting and less expected since small lesions account for a small fraction of the

entire lesion volume, which greatly affects the DSC. We attribute the increase in DSC to a better training of the CNN with improved labels. In other words, improving the labels by detecting and adding small lesions helped learning a better CNN that performed better on segmenting larger lesions as well. Comparing the first and the second rows of Table 2 shows that training on the clean dataset achieved results similar to training on the noisy dataset that included an order of magnitude larger number of scans. A similar observation was made for Faster-RCNN, where the lesion-count F1 score increased by 0.012 when trained on the clean dataset. This shows that in this application a small dataset with clean labels can be as good as a large dataset with noisy labels. In creating our clean dataset, we had to limit ourselves to a small number (12) of scans due to limited annotator time. It is likely that the results could further improve with a larger clean dataset.

5.2. Prostate cancer digital pathology classification

5.2.1. Data and labels—We use the data from Gleason2019 challenge. The goal of the challenge is to classify prostate tissue micro-array (TMA) cores as one of the four classes: benign and cancerous with Gleason grades 3, 4, and 5. Data collection and labeling have been described by Nir et al. (2018). In summary, TMA cores have been classified in detail (i.e., pixel-wise) by six pathologists independently. The Cohen's kappa coefficient for the general pathologists on this task is approximately between 0.40 and 0.60 (Allsbrook Jr et al., 2001; Nir et al., 2018), where a value of 0.0 indicates chance agreement and 1.0 indicates perfect agreement. The inter-observer variability also depends on experience (Allsbrook Jr et al., 2001); pathologists who labeled this dataset had different experience levels, ranging from 1 to 27 years. Hence, this is a classification problem and label noise is caused by inter-observer variability due to the subjective nature of grading. An example TMA core and pathologists' annotations are shown in Figure 3.

5.2.2. Methods—We used a MobileNet CNN architecture, which had been shown to be a good choice for this application by Arvaniti et al. (2018); Karimi et al. (2019) and used patches of size 768 × 768 pixels at 40X magnification as suggested by Arvaniti et al. (2018). The main feature of MobileNets is the use of separable convolutional filters, which replace a 3D convolution with a 2D depth-wise convolution (applied separately to each of the input feature maps) followed by a 1D convolution to combine these depth-wise convolutions. Our network had a depth of 7 convolutional blocks. The first block had 16 feature maps. The number of feature maps increased by a factor of 2 in each subsequent block, while reducing their size by a factor of 2 in each dimension. The output of the final convolutional block was flattened and passed through two fully connected layers. All convolutional and fully connected layers were followed by ReLU activations.

An important consideration in this application was how to divide the labels from different pathologists for training and test stages. For most of our experiments, we used the labels from all six pathologists to estimate the ground truth labels on the test data using the Simultaneous Truth and Performance Level Estimation (STAPLE) (Warfield et al., 2004). Our justification here is that, given the high inter-observer variability, this would be our best estimate of the ground truth. For these experiments, we followed a 5-fold cross-validation. Each time, we trained the CNN on 80% of the TMA cores and their labels from the six

pathologists and then evaluated the trained CNN on the STAPLE-estimated labels of the remaining 20% of the cores. However, from the viewpoint of separation between test and train data, this may not be the best approach. Therefore, we performed another set of experiments, where we used the labels from three of the pathologists for training and used STAPLE-estimated ground truth from the other three pathologists on the test set for evaluation. For this set of experiments, too, we followed a 5-fold cross-validation. However, we repeated these experiments twice, each time using labels from three of the pathologists for training. Therefore, each TMA core was tested on twice. We report the average of the two results. Below, we denote the results for this set of experiments with "3–3".

We compared the CNN predictions with the estimated truth by computing the classification accuracy and AUC for 1) distinguishing cancerous (Gleason grades 3–5) from benign tissue, and 2) in separating high-grade (Gleason grades 4 and 5) from low-grade (Gleason grade 3) cancer. In addition, we report the percentage of large classification errors, which we define as when the predicted class is 2 or 3 classes away from the true class, such as when Gleason grade 5 is classified as benign or Gleason grade 3. The compared methods were the following:

- Single pathologist. We used the label provided by one of the pathologists only, ignoring the labels provided by the others. We repeated this for all six pathologists.
- Majority vote. We computed the pixel-wise majority vote and used that for training.
- STAPLE. We used STAPLE to compute a pixel-wise label and used that for training.
- STAPLE + iMAE loss. Similar to the above, but instead of the cross-entropy loss, we used the iMAE loss (Wang et al., 2019b).
- Minimum-loss label. On each training patch, we computed the loss on labels provided by each of the six pathologists and selected the one with the smallest loss for error back-propagation. To the best of our knowledge, this method has not been proposed previously for this application.
- Annotator confusion estimation. We used the method of Tanno et al. (2019), which we reviewed above. This method estimates the labeling patterns of the annotators in parallel with the training of the CNN classification model.

5.2.3. Results—Table 3 summarizes our results. The first row shows the average of results when using the labels from one of the six pathologists. Comparing this row with the second and third rows and the row denoted as "STAPLE (3–3)" shows significant improvements due to using labels from multiple experts. Using the iMAE loss considerably improved the accuracy, especially for classifying cancerous from benign tissue. The *minimum-loss label* method also improved the classification accuracy. The iMAE loss and minimum-loss label method are based on a similar philosophy: to combat label noise, data samples with unusually high loss values should be down-weighted because they are likely to have incorrect labels. While the iMAE loss down-weights the effect of such data samples,

minimum-loss label aims at ignoring incorrect labels by using only the label with the lowest loss for each data sample. The iMAE loss performed better on classifying cancerous vs. benign tissue, whereas the minimum-loss label method performed better than the iMAE loss on classifying high-grade vs. low-grade cancer. This may be because the minimum-loss label method has a more aggressive label denoising policy and label noise (manifested as inter-pathologist disagreement) is known to be higher for high-grade vs. low-grade annotation compared with benign vs. cancerous annotation (Gulshan et al., 2016; Nir et al., 2018). Annotator confusion estimation also significantly improved the accuracy compared with the baseline methods. It can be argued that it is the best among the compared methods, as it achieved the best accuracy on high-grade vs. low-grade classification and close to the best accuracy on cancerous vs. benign classification. It also displayed the lowest rate of large classification errors at 1%. The estimated annotator confusion matrices are shown in Figure 4, which show that the pathologists had a low disagreement for benign vs. cancerous classification but relatively higher disagreement in cancer grading.

Overall, the results when labels from separate pathologists were used for training and test stages, presented in the last three rows of the table, showed similar conclusions. Specifically, using iMAE loss or modeling annotator accuracies led to better results than with cross-entropy loss and much better than when labels from a single expert were used. However, the results were worse than when labels from all six pathologists were used for training and for estimating the truth for the test set, especially for classifying high-grade versus low-grade cancer. We attribute this partly to the high inter-observer variability, which makes the estimated truth more accurate when labels from all six pathologists are used. However, this can also be because using labels from all six pathologists for training and test stages causes some overfitting that is avoided when labels from separate pathologists are used for training and test.

5.3. Fetal brain segmentation in diffusion-weighted MRI

5.3.1. Data and labels—A total of 2562 diffusion weighted (DW) MR images from 65 fetuses (between 12 and 96 images from each fetus) were used in this experiment. One image from each fetus was manually segmented by two experienced annotators. We refer to these as “clean data” and use them for evaluation. For the remaining 2497 images (between 11 and 95 images from each fetus), we generated approximate (i.e., noisy) segmentations using different methods. **Method 1:** these fetuses had reconstructed T2-weighted MR images with accurate brain segmentations, which we could transfer to the DW images via image registration. **Method 2:** we developed an algorithm based on intensity thresholding and morphological operations to synthesize approximate segmentations. This algorithm sometimes generated very inaccurate segmentations, which were detected by computing the DSC between them and the segmentation masks from the T2 image. If this DSC was below a threshold, we replaced the synthesized segmentation with that from the T2 image. This threshold and the parameters of the algorithm can be tuned to generate noisy segmentations with different accuracy levels. **Method 3:** we used a level set method to generate noisy labels. The level set method needs a seed to initialize the segmentation. In one variation of this method, we generated the seed by eroding the segmentation obtained from the T2 image mentioned above (Method 1). This resembles a semi-automatic method, where the level set

method is initialized manually. In another, fully-automatic, variation of this method we used the rough segmentations generated by Method 2, after erosion, to initialize the level set method. After every 50 training epochs, the current CNN predictions were used to initialize the level set method and new training labels were generated. To assess the accuracy of the synthesized segmentations for each method and parameter settings, we applied that method on the 65 images in the clean dataset and computed the DSC between the synthesized and manual segmentations. Figure 5 shows example scans from the clean dataset and several noisy segmentations.

5.3.2. Methods—We trained a CNN, similar to 3D U-Net for experiments in this section. This architecture included four convolutional blocks in each of its contracting and expanding parts. The first block extracted 10 feature maps from the image. The number of feature maps increased by a factor of 2 in subsequent convolutional blocks. Each convolutional block included two standard convolutional layers with a residual connection. Similar to the other networks used in this work, a ReLU activation was used after each convolutional operation. We adopted a five-fold cross-validation strategy for all experiments in this section. The cross-validation was subject-wise, meaning that no scans from the test subjects were used for training. The compared training methods were:

- Baseline CNN.
- Baseline CNN trained with MAE loss.
- Dual CNNs with iterative label update. This is a novel method that we propose for fetal brain segmentation for the first time. We trained two CNNs, with the same architecture as the baseline CNN, but with 0.80 and 1.25 times the number of feature maps as the baseline CNN to encourage diversity. The CNNs were first trained on the initial noisy labels. Subsequently, they were used to predict segmentations on the images with noisy labels. In an iterative framework, first each CNN was trained using the labels predicted by the other CNN or the noisy label, whichever resulted in a lower loss. Then, at the end of each training epoch, each noisy segmentation mask was replaced by the mask predicted by one of the CNNs if any one of them resulted in a lower loss; it was replaced by the average of the two CNN-predicted masks if both resulted in lower losses.

5.3.3. Results—The first row of Table 4 shows the DSC of the synthesized noisy segmentations, computed on the 65 images with manual segmentation. This can be regarded as an estimation of the accuracy of the training labels. The second row shows that strong label noise significantly affects the performance of the baseline CNN; the DSC achieved at test time always trails the DSC of the training labels. This is in disagreement with the results reported for handwritten digit recognition by Rolnick et al. (2017). As we reviewed above, Rolnick et al. (2017) found that given sufficient training data with label accuracy slightly above random noise, classification accuracy of 90% was achieved at test time. This difference is probably because our segmentation problem is more difficult and our training set is much smaller. Nonetheless, it is interesting to note that at the lowest label noise (noise level 1), the test DSC achieved by the baseline CNN (0.889) was higher than that achieved by the same model trained on the clean dataset (0.878), which consisted of approximately 40

times fewer images. For higher noise levels, training with MAE loss improved the classification results compared with the baseline CNN trained with the DSC loss. Dual CNN training with iterative label update performed consistently better than the baseline CNN and also performed much better than MAE loss on noise levels 1–5. For noise level 3–5, DSC achieved with this method was also higher than the DSC of the noisy labels that were used at the start of training.

Table 5 shows more detailed performance measures on three different label noise levels. It shows the mean and standard deviation of the DSC and 95-percentile of the Hausdorff Distance (HD95), as well as the 5-percentile of the DSC (5% DSC) among the 65 test images. The results show that both MAE loss and Dual CNNs with iterative label update reduce the large segmentation errors, quantified with HD95, in the presence of strong label noise. There is also some improvement in 5% DSC, which is a measure of worst-case performance.Worst-case performance of the trained model is affected not only by the model accuracy, but also by the outlier data samples. Although the techniques reviewed in this paper and the methods used in our experiments are hoped to lead to better models that should perform better on average, the link to data outliers and difficult samples is less obvious. The great majority of the studies reviewed in this paper do not address the worst-case performance and data outliers, as those are essentially a different problem than label noise.

6. Discussion and Conclusions

Label noise is unavoidable in many medical image datasets. It can be caused by limited attention or expertise of the human annotator, subjective nature of labeling, or errors in computerized labeling systems. Since deep learning methods are increasingly used in medical image analysis, a proper understanding of the effects of label noise in training data and methods to manage those effects are essential. To help improve this understanding, this paper first presented a review of studies on label noise in machine learning and deep learning, followed by a review of studies on label noise in deep learning for medical image analysis; and second, investigated several existing and new methods and remedies to deal with different types of label noise in three different medical image datasets in detection, segmentation, and classification applications.

Our review of the literature shows that many studies have demonstrated negative effects of label noise in deep learning. Our review also shows that a diverse set of methods have been proposed and successfully applied to handle label noise in deep learning. Most of these methods have been developed for general computer vision and machine learning problems. Moreover, many of these methods have been evaluated on large-scale image classification datasets. Hence, reassessment of their performance for medical image analysis applications is warranted. Given the large variability in data size, label noise, and the nature of tasks that one may encounter in medical image analysis, it is likely that for each application one has to experiment with a number of methods to find the most suitable one. In spite of the need, our review of the literature shows that very few studies have directly addressed the issue of label noise in deep learning for medical image analysis. Therefore, motivated by the need, in a set of experiments reported in Section 5 we investigated and developed several existing

strategies and new methods to reduce the negative impact of label noise in deep learning for different medical image analysis applications. Based on the results of our experiments and the literature, we make general recommendations as follow.

Label cleaning and pre-processing methods can be useful for most medical image analysis applications, but one has to be selective. Some methods in this category rely on prediction confidence for detecting incorrectly-labeled samples (Northcutt et al., 2017), (Ding et al., 2018). These methods can only be effective if the trained model has a well-calibrated prediction. Moreover, some methods in this category rely on matching a data sample or its feature vector with a set of data samples with clean labels (Vo et al., 2015) (Lee et al., 2018). These methods may also have limited applicability in medical image analysis because data samples are larger in size and fewer in number, making the data matching more challenging due to the curse of dimensionality. On the other hand, methods such as that proposed by Veit et al. (2017) could be useful in many detection and classification applications. Interestingly, in our experiments on brain lesion detection in Section 5.1, we achieved our best results by iterative label cleaning, indicating the great potential of these methods.

In the category of studies that suggest changing the network architecture, most methods introduce a noise layer or graphical model to learn the label noise statistics in parallel with model training. These methods are relatively easy to implement and evaluate. Yet, we are aware of only one study that has reported successfully employing such a method in medical image analysis (Dgani et al., 2018). Nonetheless, we demonstrated the potential of these methods with our experiments in Section 5.2, where we obtained our best results with a method in this category involving estimation of the statistics of annotation error. Based on our results and those reported by studies in the machine learning and computer vision literature, we think methods in this category could be highly effective for classification and detection applications. Of particular utility to medical image analysis tasks are methods that enable estimation of labeling error of one or multiple annotators, such as the method of Tanno et al. (2019) that we used in our experiments in Section 5.2.

Noise-robust loss functions have been mainly proposed as substitutes for cross-entropy loss for classification applications. Nonetheless, in addition to our pathology classification experiments (Section 5.2), such loss functions also proved to be useful in our fetal brain segmentation experiments (Section 5.3). In our experiments, we used MAE and iMAE loss functions, which are based on down-weighting data samples that are more likely to be incorrectly-labeled. More aggressive loss functions such as those proposed by Thulasidasan et al. (2019) and Rusiecki (2019) could be more effective under very strong label noise. An advantage of these loss functions is that they are easy to quickly implement and evaluate.

Data re-weighting methods are also typically easy to implement. Several studies have already reported successful application of data re-weighting methods in medical image analysis (Le et al. (2019), Xue et al. (2019), Zhu et al. (2019), Mirikharaji et al. (2019)). In our own experiments, we implemented two variations of data re-weighting and found both of them to be effective. In experiments on lesion detection/segmentation (Section 5.1), we down-weighted data samples with high loss values, whereas in experiments on pathology classification (Section 5.2), where we had multiple labels for each data sample, we down-

weighted high-loss labels. Such methods may be effective in many similar applications in medical image analysis.

Among the six categories of surveyed methods, those based on data and label consistency may be less applicable to medical image analysis tasks. Most of the proposed methods in this category are based on some measure of correlation or similarity between different data samples in the feature space. Due to the large dimensionality of the feature space in deep learning, these methods can suffer from the curse of dimensionality, as suggested by Chen et al. (2019a). This problem can be more serious in medical image analysis due to the relatively large size of medical images and the relatively small number of samples.

Lastly, methods based on novel training procedures encompass a wide range of techniques that could be useful for almost all applications in medical image analysis. Given the diversity of methods in this category from the machine learning and computer vision literature, this seems to be an area with great potential for innovations and flexible application-specific solutions. Previous studies have developed and successfully evaluated such application-specific solutions for various medical image analysis tasks including segmentation (Min et al. (2018); Nie et al. (2018); Zhang et al. (2018)) and classification (Fries et al. (2019)). Our proposed Dual CNNs with iterative label update, presented and tested in Section 5.3, is a successful example of these methods for deep learning with noisy labels.

Deep learning for medical image analysis presents specific challenges that can be different from many computer vision and machine learning applications. These peculiarities may influence the choice of solutions for combating label noise as well. Our experiments in Section 5 revealed some of these challenges. For example, an important characteristic of medical image datasets, in particular those carefully annotated by human experts, is their small size. The data size may have a complicated interplay with label noise. In our experiments on brain lesion segmentation in Section 5.1, a small (n=12) but carefully annotated training dataset resulted in a better model compared with a much larger (n=153) dataset with noisy annotations. By contrast, in our fetal brain segmentation experiment in Section 5.3, more accurate models were trained using many images ($n \approx 2500$) with slightly noisy segmentations than using much fewer (n=65) images with manual segmentations. The interplay between the size and accuracy of the labeled training data also depends on the application. This warrants a reassessment of the optimal ways of obtaining labels from human experts or other means for each application.

The data size may also influence the effectiveness of different strategies for handling label noise. For example, in several studies in computer vision that we reviewed in this paper, down-weighting or completely discarding data samples that were more likely to have incorrect labels proved to be an effective approach. This may be a less effective approach in medical imaging where datasets are relatively small. As shown in Table 2, for brain lesion segmentation we obtained better results by detecting and correcting missing annotations than by ignoring data samples with high loss values. For prostate digital pathology experiments in Section 5.2, where we had access to labels from six pathologists, ignoring high-loss labels proved effective. Nonetheless, on this dataset we achieved better performance by modeling

annotator confusion rather than ignoring high-loss labels. For our fetal brain segmentation, too, we experimented with methods to down-weight or ignore segmentations that were more likely to be incorrect, but we did not achieve good results. Based on our experimental results and observations, it is better to improve the label accuracy or estimate the labeling error using techniques such as those we used in Sections 5.1 and 5.2 rather than to ignore data samples that are likely to have incorrect labels.

Another important consideration in medical image datasets is the subjective nature of annotation and the impact of inter-observer variability. If labels are obtained from a single expert, as in our experiments in Section 5.1, annotations may be systematically biased due to annotation habits or subjective opinion of a single annotator, risking generalizability when compared with the “true label”. The level of inter-observer variability depends significantly on factors such as the application, observer expertise, and attention (Gurari et al., 2015; Lampert et al., 2016; Donovan and Litchfield, 2013; Nagpal et al., 2018). Our experiments in Section 5.2 targeted an application with known high inter-observer variability. Our results suggest that when labels from multiple experts are available, methods that model observer confusion as part of the training process generally perform better than methods that aggregate the labels in a separate step prior to training. Our results also showed significant gains due to using labels from multiple experts.

Results of our experiments with brain lesion segmentation in Section 5.1 and with digital pathology in Section 5.2 share an important lesson. In both of these experiments, we achieved improved performance by modeling annotation error of the human expert(s). In Section 5.1, we observed that the annotator systematically missed smaller, fainter, and more isolated lesions. This is an expected behavior, and similar observations have been reported in previous studies (Robinson et al., 2016; Quekel et al., 1999; Kundel and Revesz, 1976). In our experiments, we exploited CNN prediction uncertainty, which enabled us to devise a novel and effective method to detect and fill in missing annotations in the training labels. Similar methods can be effective in training deep learning models for datasets with incomplete annotations, which are commonplace in medical image analysis. In Section 5.2, on the other hand, we exploited an approach originally proposed for general computer vision applications, and achieved very good performance. This method, which estimated the annotation error of individual experts in parallel with CNN training, proved to be more effective than several other methods including label fusion algorithms.

Our experiment on fetal brain segmentation in DW-MRI in Section 5.3 showed the potential value of computer-generated noisy labels. An interesting observation was that the baseline CNN achieved better results when trained with noisy segmentation masks transferred from the corresponding T2 images than when trained on 65 images that had been manually segmented. There are many situations in medical image analysis where such approximate annotations can be obtained at little or no cost from other images of the same subject, from matched subjects, or from an atlas. Our results demonstrate the potential utility of such annotations. Nonetheless, our results also showed that very inaccurate annotations led to poor training, indicating an important limitation of such labels.

In summary, in our experiments we investigated three common types of label noise in medical image datasets, and the relative effectiveness of several approaches to reduce the negative impact of label noise. The source, statistics, and strength of label noise in medical imaging is diverse; and our study shows that the effects of label noise should be carefully analyzed in training deep learning algorithms. This warrants further investigations and development of robust models and training algorithms.

Acknowledgements

This study was supported in part by the National Institute of Biomedical Imaging and Bioengineering, and the National Institute of Neurological Disorders and Stroke of the National Institutes of Health (NIH) under Award Numbers R01EB018988, R01NS106030, and R01NS079788; and by a Technological Innovations in Neuroscience Award from the McKnight Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the McKnight Foundation.

References

Abellán J, Masegosa AR, 2010 Bagging decision trees on data sets with classification noise, in: International Symposium on Foundations of Information and Knowledge Systems, Springer. pp. 248–265.

Acuna D, Kar A, Fidler S, 2019 Devil is in the edges: Learning semantic boundaries from noisy annotations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11075–11083.

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 transactions on medical imaging 35, 1313–1321. [PubMed: 26891484]

Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI, 2001 Interobserver reproducibility of gleason grading of prostatic carcinoma: general pathologist. Human pathology 32, 81–88. [PubMed: 11172299]

Arpit D, Jastrzebski S, Ballas N, Krueger D, Bengio E, Kanwal MS, Maharaj T, Fischer A, Courville A, Bengio Y, et al., 2017 A closer look at memorization in deep networks, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 233–242.

Arvaniti E, Fricker KS, Moret M, Rupp NJ, Hermanns T, Fankhauser C, Wey N, Wild PJ, Rueschoff JH, Claassen M, 2018 Automated gleason grading of prostate cancer tissue microarrays via deep learning. bioRxiv, 280024.

Azadi S, Feng J, Jegelka S, Darrell T, 2015 Auxiliary image regularization for deep cnns with noisy labels. arXiv preprint arXiv:1511.07069.

Bekker AJ, Goldberger J, 2016 Training deep neural-networks based on unreliable labels, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2682–2686.

Bengio Y, Louradour J, Collobert R, Weston J, 2009 Curriculum learning, in: Proceedings of the 26th annual international conference on machine learning, ACM. pp. 41–48.

Boughorbel S, Jarray F, Venugopal N, Elhadi H, 2018 Alternating loss correction for preterm-birth prediction from ehr data with noisy labels. arXiv preprint arXiv:1811.09782.

Bridge P, Fielding A, Rowntree P, Pullar A, 2016 Intraobserver variability: Should we worry? Journal of medical imaging and radiation sciences 47, 217–220. [PubMed: 31047285]

Brodley CE, Friedl MA, et al., 1996 Identifying and eliminating mislabeled training instances, in: Proceedings of the National Conference on Artificial Intelligence, pp. 799–805.

Carass A, Roy S, Jog A, Cuzzocreo JL, Magrath E, Gherman A, Button J, Nguyen J, Prados F, Sudre CH, et al., 2017 Longitudinal multiple sclerosis lesion segmentation: resource and challenge. NeuroImage 148, 77–102. [PubMed: 28087490]

Chen P, Liao B, Chen G, Zhang S, 2019a A meta approach to defend noisy labels by the manifold regularizer psdr. arXiv preprint arXiv:1906.05509.

Chen P, Liao B, Chen G, Zhang S, 2019b Understanding and utilizing deep neural networks trained with noisy labels. arXiv preprint arXiv:1905.05040.

Cheplygina V, de Bruijne M, Pluim JP, 2019 Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis 54, 280–296. [PubMed: 30959445]

Chiaroni F, Rahal M, Hueber N, Dufaux F, 2019 Hallucinating a cleanly labeled augmented dataset from a noisy labeled dataset using gans.

Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, et al., 2018 Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface 15, 20170387.

Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O, 2016 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer pp. 424–432.

Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, Pop SC, Girard P, Ameli R, Ferré JC, et al., 2018 Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Scientific reports 8, 1–17. [PubMed: 29311619]

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L, 2009 Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee pp. 248–255.

Dgani Y, Greenspan H, Goldberger J, 2018 Training a neural network based on unreliable human annotation of medical images, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 39–42.

Dietterich TG, 2000 An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning 40, 139–157.

Ding Y, Wang L, Fan D, Gong B, 2018 A semi-supervised two-stage approach to learning from noisy labels, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 1215–1224.

Donovan T, Litchfield D, 2013 Looking for cancer: Expertise related differences in searching and decision making. Applied Cognitive Psychology 27, 43–49.

Drory A, Avidan S, Giryes R, 2018 On the resistance of neural nets to label noise. arXiv preprint arXiv:1803.11410.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S, 2017 Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115. [PubMed: 28117445]

Folleco A, Khoshgoftaar TM, Van Hulse J, Bullard L, 2008 Identifying learners robust to low quality data, in: 2008 IEEE International Conference on Information Reuse and Integration, IEEE. pp. 190–195.

Frénay B, Verleysen M, 2013 Classification in the presence of label noise: a survey. IEEE transactions on neural networks and learning systems 25, 845–869.

Freund Y, Schapire R, Abe N, 1999 A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence 14, 1612.

Fries JA, Varma P, Chen VS, Xiao K, Tejeda H, Saha P, Dunnmon J, Chubb H, Maskatia S, Fiterau M, et al., 2019 Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences. BioRxiv, 339630.

Gal Y, Ghahramani Z, 2015 Bayesian convolutional neural networks with bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158.

Gamberger D, Lavrac N, Dzeroski S, 2000 Noise detection and elimination in data preprocessing: experiments in medical domains. Applied Artificial Intelligence 14, 205–223.

Gao BB, Xing C, Xie CW, Wu J, Geng X, 2017 Deep label distribution learning with label ambiguity. IEEE Transactions on Image Processing 26, 2825–2838. [PubMed: 28371776]

García S, Luengo J, Herrera F, 2015 Data preprocessing in data mining. Springer.

Ghosh A, Kumar H, Sastry P, 2017 Robust loss functions under label noise for deep neural networks, in: Thirty-First AAAI Conference on Artificial Intelligence Goldberger,

Goldberger J, Ben-Reuven E, 2016 Training deep neural-networks using a noise adaptation layer.

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, et al., 2016 Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 316, 2402–2410. [PubMed: 27898976]

Guo C, Pleiss G, Sun Y, Weinberger KQ, 2017 On calibration of modern neural networks. arXiv preprint arXiv:1706.04599.

Guo S, Huang W, Zhang H, Zhuang C, Dong D, Scott MR, Huang D, 2018 Curriculumnet: Weakly supervised learning from large-scale web images, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 135–150.

Guo Y, Zhang L, Hu Y, He X, Gao J, 2016 Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, in: European Conference on Computer Vision, Springer. pp. 87–102.

Gurari D, Theriault D, Sameki M, Isenberg B, Pham TA, Purwada A, Solski P, Walker M, Zhang C, Wong JY, et al., 2015 How to collect segmentations for biomedical images? a benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms, in: 2015 IEEE winter conference on applications of computer vision, IEEE. pp. 1169–1176.

Han B, Niu G, Yao J, Yu X, Xu M, Tsang I, Sugiyama M, 2018a Pumpout: A meta approach for robustly training deep neural networks with noisy labels. arXiv preprint arXiv:1809.11008.

Han B, Yao Q, Yu X, Niu G, Xu M, Hu W, Tsang I, Sugiyama M, 2018b Co-teaching: Robust training of deep neural networks with extremely noisy labels, in: Advances in Neural Information Processing Systems, pp. 8527–8537.

Han J, Luo P, Wang X, 2019 Deep self-learning from noisy labels. arXiv preprint arXiv:1908.02160.

Hashemi SR, Salehi SSM, Erdogmus D, Prabhu SP, Warfield SK, Gholipour A, 2019 Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. IEEE Access 7, 1721–1735.

Haskins G, Kruger U, Yan P, 2019 Deep learning in medical image registration: A survey. arXiv preprint arXiv:1903.02026.

Hendrycks D, Mazeika M, Wilson D, Gimpel K, 2018 Using trusted data to train deep networks on labels corrupted by severe noise, in: Advances in Neural Information Processing Systems, pp. 10456–10465.

Hinton G, Vinyals O, Dean J, 2015 Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Houle ME, 2017 Local intrinsic dimensionality i: an extreme-value-theoretic foundation for similarity applications, in: International Conference on Similarity Search and Applications, Springer pp. 64–79.

Ipeirotis PG, Provost F, Wang J, 2010 Quality management on amazon mechanical turk, in: Proceedings of the ACM SIGKDD workshop on human computation, ACM. pp. 64–67.

Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, et al., 2019 Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031.

Izadinia H, Russell BC, Farhadi A, Hoffman MD, Hertzmann A, 2015 Deep classifiers from image tags in the wild, in: Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions, ACM. pp. 13–18.

Jiang L, Zhou Z, Leung T, Li LJ, Fei-Fei L, 2017 Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. arXiv preprint arXiv:1712.05055.

Jindal I, Nokleby M, Chen X, 2016 Learning deep networks from noisy labels with dropout regularization, in: 2016 IEEE 16th International Conference on Data Mining (ICDM), IEEE. pp. 967–972.

Karimi D, Nir G, Fazli L, Black P, Goldenberg L, Salcudean S, 2019 Deep learning-based gleason grading of prostate cancer from histopathology images-role of multiscale decision aggregation and data augmentation. IEEE journal of biomedical and health informatics.

Kaster FO, Menze BH, Weber MA, Hamprecht FA, 2010 Comparative validation of graphical models for learning tumor segmentations from noisy manual annotations, in: International MICCAI Workshop on Medical Computer Vision, Springer. pp. 74–85.

Kendall A, Gal Y, 2017 What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in neural information processing systems, pp. 5574–5584.

Khardon R, Wachman G, 2007 Noise tolerant variants of the perceptron algorithm. Journal of Machine Learning Research 8, 227–248.

Khetan A, Lipton ZC, Anandkumar A, 2017 Learning from noisy singly-labeled data. arXiv preprint arXiv:1712.04577.

Kim HC, Ghahramani Z, 2006 Bayesian gaussian process classification with the em-ep algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 1948–1959. [PubMed: 17108369]

Köhler JM, Autenrieth M, Beluch WH, 2019 Uncertainty based detection and relabeling of noisy image labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 33–37.

Kundel HL, Revesz G, 1976 Lesion conspicuity, structured noise, and film reader error. American Journal of Roentgenology 126, 1233–1238. [PubMed: 179387]

Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, Kamali S, Popov S, Malloci M, Duerig T, et al., 2018 The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982.

Lakshminarayanan B, Pritzel A, Blundell C, 2017 Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems, pp. 6402–6413.

Lampert TA, Stumpf A, Gançarski P, 2016 An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. IEEE Transactions on Image Processing 25, 2557–2572. [PubMed: 27019487]

Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, Flanders AE, Lungren MP, Mendelson DS, Rudie JD, et al., 2019 A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 nih/rsna/acr/the academy workshop. Radiology 291, 781–791. [PubMed: 30990384]

Le H, Samaras D, Kurc T, Gupta R, Shroyer K, Saltz J, 2019 Pancreatic cancer detection in whole slide images using noisy label annotations, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing.

LeCun Y, Bengio Y, Hinton G, 2015 Deep learning. nature 521, 436. [PubMed: 26017442]

Lee K, Yun S, Lee K, Lee H, Li B, Shin J, 2019 Robust inference via generative classifiers for handling noisy labels. arXiv preprint arXiv:1901.11300.

Lee KH, He X, Zhang L, Yang L, 2018 Cleannet: Transfer learning for scalable image classifier training with label noise, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5447–5456.

Li C, Zhang C, Ding K, Li G, Cheng J, Lu H, 2017a Bundlenet: Learning with noisy label via sample correlations. IEEE Access 6, 2367–2377.

Li J, Wong Y, Zhao Q, Kankanhalli MS, 2019 Learning to learn from noisy labeled data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5051–5059.

Li Y, Yang J, Song Y, Cao L, Luo J, Li LJ, 2017b Learning from noisy labels with distillation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1910–1918.

Lin C.f., et al., 2004 Training algorithms for fuzzy support vector machines with noisy data. Pattern recognition letters 25, 1647–1656.

Lin TY, Goyal P, Girshick R, He K, Dollár P, 2017 Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Liu T, Tao D, 2015 Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence 38, 447–461.

Liu X, Li S, Kan M, Shan S, Chen X, 2017 Self-error-correcting convolutional neural network for learning with noisy labels, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE. pp. 111–117.

Long PM, Servedio RA, 2010 Random classification noise defeats all convex potential boosters. Machine learning 78, 287–304.

Ma X, Wang Y, Houle ME, Zhou S, Erfani SM, Xia ST, Wijewickrema S, Bailey J, 2018 Dimensionality-driven learning with noisy labels. arXiv preprint arXiv:1806.02612.

Malach E, Shalev-Shwartz S, 2017 Decoupling" when to update" from" how to update", in: Advances in Neural Information Processing Systems, pp. 960–970.

Malossini A, Blanzieri E, Ng RT, 2006 Detecting potential labeling errors in microarrays by data perturbation. Bioinformatics 22, 2114–2121. [PubMed: 16820424]

Manwani N, Sastry P, 2013 Noise tolerance under risk minimization. IEEE transactions on cybernetics 43, 1146–1151. [PubMed: 23193242]

Martin CH, Mahoney MW, 2017 Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. arXiv preprint arXiv:1710.09553.

Matuszewski DJ, Sintorn IM, 2018 Minimal annotation training for segmentation of microscopy images, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 387–390.

McDonald RA, Hand DJ, Eckley IA, 2003 An empirical comparison of three boosting algorithms on real data sets with artificial class noise, in: International Workshop on Multiple Classifier Systems, Springer. pp. 35–44.

Min S, Chen X, Zha ZJ, Wu F, Zhang Y, 2018 A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. arXiv preprint arXiv:1807.11719.

Mirikharaji Z, Yan Y, Hamarneh G, 2019 Learning to segment skin lesions from noisy annotations. arXiv preprint arXiv:1906.03815.

Misra I, Lawrence Zitnick C, Mitchell M, Girshick R, 2016 Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2930–2939.

Mnih V, Hinton GE, 2012 Learning to label aerial images from noisy data, in: Proceedings of the 29th International conference on machine learning (ICML-12), pp. 567–574.

Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Vega JEV, Brat DJ, Cooper LA, 2018 Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences 115, E2970–E2979.

Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P, 2017 Universal adversarial perturbations, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1765–1773.

Müller R, Kornblith S, Hinton G, 2019 When does label smoothing help? arXiv preprint arXiv:1906.02629.

Nagpal K, Foote D, Liu Y, Wulczyn E, Tan F, Olson N, Smith JL, Mohtashamian A, Wren JH, Corrado GS, et al., 2018. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. arXiv preprint arXiv:1811.06497.

Natarajan N, Dhillon IS, Ravikumar PK, Tewari A, 2013 Learning with noisy labels, in: Advances in neural information processing systems, pp. 1196–1204.

Nettleton DF, Orriols-Puig A, Fornells A, 2010 A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial intelligence review 33, 275–306.

Nie D, Gao Y, Wang L, Shen D, 2018 Asdnet: Attention based semi-supervised deep networks for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 370–378.

Nir G, Hor S, Karimi D, Fazli L, Skinnider BF, Tavassoli P, Turbin D, Villamil CF, Wang G, Wilson RS, Iczkowski KA, Lucia MS, Black PC, Abolmaesumi P, Goldenberg SL, Salcudean SE, 2018 Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. Medical Image Analysis 50, 167–180. URL: http://www.sciencedirect.com/science/article/pii/S1361841518307497, doi: 10.1016/j.media.2018.09.005. [PubMed: 30340027]

Northcutt CG, Wu T, Chuang IL, 2017 Learning with confident examples: Rank pruning for robust classification with noisy labels. arXiv preprint arXiv:1705.01936.

Ostyakov P, Logacheva E, Suvorov R, Aliev V, Sterkin G, Khomenko O, Nikolenko SI, 2018 Label denoising with large ensembles of heterogeneous neural networks, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 0–0.

Patrini G, Nielsen F, Nock R, Carioni M, 2016 Loss factorization, weakly supervised learning and label noise robustness, in: International conference on machine learning, pp. 708–717.

Patrini G, Rozza A, Krishna Menon A, Nock R, Qu L, 2017 Making deep neural networks robust to label noise: A loss correction approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1944–1952.

Pawlowski N, Brock A, Lee MC, Rajchl M, Glocker B, 2017 Implicit weight uncertainty in neural networks. arXiv preprint arXiv:1711.01297.

Pham HH, Le TT, Tran DQ, Ngo DT, Nguyen HQ, 2019 Interpreting chest x-rays via cnns that exploit disease dependencies and uncertainty labels. arXiv preprint arXiv:1911.06475.

Prevedello LM, Halabi SS, Shih G, Wu CC, Kohli MD, Chokshi FH, Erickson BJ, Kalpathy-Cramer J, Andriole KP, Flanders AE, 2019 Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. Radiology: Artificial Intelligence 1, e180031.

Quekel LG, Kessels AG, Goei R, van Engelshoven JM, 1999 Miss rate of lung cancer on the chest radiograph in clinical practice. Chest 115, 720–724. [PubMed: 10084482]

Quinlan JR, 1986 Induction of decision trees. Machine learning 1, 81–106.

Ratner AJ, De Sa CM, Wu S, Selsam D, Ré C, 2016 Data programming: Creating large training sets, quickly, in: Advances in neural information processing systems, pp. 3567–3575. [PubMed: 29872252]

Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ, 2016 Deep learning for health informatics. IEEE journal of biomedical and health informatics 21, 4–21. [PubMed: 28055930]

Raykar VC, Yu S, Zhao LH, Valadez GH, Florin C, Bogoni L, Moy L, 2010 Learning from crowds. Journal of Machine Learning Research 11, 1297–1322.

Reed S, Lee H, Anguelov D, Szegedy C, Erhan D, Rabinovich A, 2014 Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596.

Reid MD, Williamson RC, 2010 Composite binary losses. Journal of Machine Learning Research 11, 2387–2422.

Ren M, Zeng W, Yang B, Urtasun R, 2018 Learning to reweight examples for robust deep learning. arXiv preprint arXiv:1803.09050.

Ren S, He K, Girshick R, Sun J, 2015 Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

Rister B, Yi D, Shivakumar K, Nobashi T, Rubin DL, 2018 Ct organ segmentation using gpu data augmentation, unsupervised labels and iou loss. arXiv preprint arXiv:1811.11226.

Robinson JW, Brennan PC, Mello-Thoms C, Lewis SJ, 2016 The impact of radiology expertise upon the localization of subtle pulmonary lesions, in: Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment, International Society for Optics and Photonics p. 97870K.

Rolnick D, Veit A, Belongie S, Shavit N, 2017 Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694.

Ronneberger O, Fischer P, Brox T, 2015 U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.

Rusiecki A, 2019 Trimmed robust loss function for training deep neural networks with label noise, in: International Conference on Artificial Intelligence and Soft Computing, Springer. pp. 215–222.

Segata N, Blanzieri E, Cunningham P, 2009 A scalable noise reduction technique for large case-based systems, in: International Conference on Case-Based Reasoning, Springer. pp. 328–342.

Shen Y, Sanghavi S, 2019 Learning with bad training data via iterative trimmed loss minimization, in: International Conference on Machine Learning, pp. 5739–5748.

Shrivastava A, Gupta A, Girshick R, 2016 Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 761–769.

Shu J, Xie Q, Yi L, Zhao Q, Zhou S, Xu Z, Meng D, 2019 Meta-weight-net: Learning an explicit mapping for sample weighting. arXiv preprint arXiv:1902.07379.

Sluban B, Gamberger D, Lavra N, 2010 Advances in class noise detection, in: Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence, IOS Press. pp. 1105–1106.

Song S, Chaudhuri K, Sarwate A, 2015 Learning from data with heterogeneous noise using sgd, in: Artificial Intelligence and Statistics, pp. 894–902.

Speth J, Hand EM, 2019 Automated label noise identification for facial attribute recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 25–28.

Sukhbaatar S, Bruna J, Paluri M, Bourdev L, Fergus R, 2014 Training convolutional networks with noisy labels. arXiv preprint arXiv:1406.2080.

Sukhbaatar S, Fergus R, 2014 Learning from noisy labels with deep neural networks. arXiv preprint arXiv:1406.2080 2, 4.

Sun C, Shrivastava A, Singh S, Gupta A, 2017 Revisiting unreasonable effectiveness of data in deep learning era, in: Proceedings of the IEEE international conference on computer vision, pp. 843–852.

Sun J.w., Zhao F.y., Wang C.j., Chen S.f., 2007 Identifying and correcting mislabeled training instances, in: Future generation communication and networking (FGCN 2007), IEEE. pp. 244–250.

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z, 2016 Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

Tajbakhsh N, Jeyaseelan L, Li Q, Chiang J, Wu Z, Ding X, 2019 Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. arXiv preprint arXiv:1908.10454.

Tanaka D, Ikami D, Yamasaki T, Aizawa K, 2018 Joint optimization framework for learning with noisy labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5552–5560.

Tanno R, Saeedi A, Sankaranarayanan S, Alexander DC, Silberman N, 2019 Learning from noisy labels by regularized estimation of annotator confusion. arXiv preprint arXiv:1902.03680.

Thekumparampil KK, Khetan A, Lin Z, Oh S, 2018 Robustness of conditional gans to noisy labels, in: Advances in Neural Information Processing Systems, pp. 10271–10282.

Thulasidasan S, Bhattacharya T, Bilmes J, Chennupati G, Mohd-Yusof J, 2019 Combating label noise in deep learning using abstention. arXiv preprint arXiv:1905.10964.

Topol E, 2019a Deep medicine: how artificial intelligence can make healthcare human again. Hachette UK.

Topol EJ, 2019b High-performance medicine: the convergence of human and artificial intelligence. Nature medicine 25, 44.

Vahdat A, 2017 Toward robustness against label noise in training deep discriminative neural networks, in: Advances in Neural Information Processing Systems, pp. 5596–5605.

Van Rooyen B, Menon A, Williamson RC, 2015 Learning with symmetric label noise: The importance of being unhinged, in: Advances in Neural Information Processing Systems, pp. 10–18.

Veit A, Alldrin N, Chechik G, Krasin I, Gupta A, Belongie S, 2017 Learning from noisy large-scale datasets with minimal supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 839–847.

Vo PD, Ginsca A, Le Borgne H, Popescu A, 2015 Effective training of convolutional networks using noisy web images, in: 2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI), IEEE. pp. 1–6.

Wang F, Chen L, Li C, Huang S, Chen Y, Qian C, Change Loy C, 2018a The devil of face recognition is in the noise, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 765–780.

Wang G, Kalra M, Orton CG, 2017a Machine learning will transform radiology significantly within the next 5 years. Medical physics 44, 2041–2044. [PubMed: 28295412]

Wang G, Ye JC, Mueller K, Fessler JA, 2018b Image reconstruction is a new frontier of machine learning. IEEE transactions on medical imaging 37, 1289–1296. [PubMed: 29870359]

Wang X, Hua Y, Kodirov E, Robertson N, 2019a Emphasis regularisation by gradient rescaling for training deep neural networks with noisy labels. arXiv preprint arXiv:1905.11233.

Wang X, Kodirov E, Hua Y, Robertson NM, 2019b Improving mae against cce under label noise. arXiv preprint arXiv:1903.12141.

Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM, 2017b Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2097–2106.

Wang Y, Liu W, Ma X, Bailey J, Zha H, Song L, Xia ST, 2018c Iterative learning with open-set noisy labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8688–8696.

Warfield SK, Zou KH, Wells WM, 2004 Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging 23, 903–921. [PubMed: 15250643]

Wilson DR, Martinez TR, 1997 Instance pruning techniques, in: ICML, pp. 400–411.

Wilson DR, Martinez TR, 2000 Reduction techniques for instance-based learning algorithms. Machine learning 38, 257–286.

Xiao T, Xia T, Yang Y, Huang C, Wang X, 2015 Learning from massive noisy labeled data for image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2691–2699.

Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, Mak RH, Aerts HJ, 2019 Deep learning predicts lung cancer treatment response from serial medical imaging. Clinical Cancer Research 25, 3266–3275. [PubMed: 31010833]

Xue C, Dou Q, Shi X, Chen H, Heng PA, 2019 Robust learning at noisy labeled medical images: Applied to skin lesion classification. arXiv preprint arXiv:1901.07759.

Yan K, Wang X, Lu L, Summers RM, 2018 Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of Medical Imaging 5, 036501. [PubMed: 30035154]

Yao J, Wang J, Tsang IW, Zhang Y, Sun J, Zhang C, Zhang R, 2018 Deep learning from noisy image labels with quality embedding. IEEE Transactions on Image Processing 28, 1909–1922.

Yi K, Wu J, 2019 Probabilistic end-to-end noise correction for learning with noisy labels. arXiv preprint arXiv:1903.07788.

Yu L, Wang S, Li X, Fu CW, Heng PA, 2019a Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. arXiv preprint arXiv:1907.07034.

Yu X, Han B, Yao J, Niu G, Tsang I, Sugiyama M, 2019b How does disagreement help generalization against label corruption?, in: International Conference on Machine Learning, pp. 7164–7173.

Yu X, Liu T, Gong M, Zhang K, Tao D, 2017 Transfer learning with label noise. arXiv preprint arXiv:1707.09724.

Yu Z, Liu W, Zou Y, Feng C, Ramalingam S, Vijaya Kumar B, Kautz J, 2018 Simultaneous edge alignment and learning, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 388–404.

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O, 2016 Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530.

Zhang C, Wu C, Blanzieri E, Zhou Y, Wang Y, Du W, Liang Y, 2009 Methods for labeling error detection in microarrays based on the effect of data perturbation on the regression model. Bioinformatics 25, 2708–2714. [PubMed: 19661242]

Zhang H, Cisse M, Dauphin YN, Lopez-Paz D, 2017 mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

Zhang L, Gopalakrishnan V, Lu L, Summers RM, Moss J, Yao J, 2018 Self-learning to detect and segment cysts in lung ct images without manual annotation, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 1100–1103.

Zhang W, Wang Y, Qiao Y, 2019 Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7373–7382.

Zhang Z, Sabuncu M, 2018 Generalized cross entropy loss for training deep neural networks with noisy labels, in: Advances in Neural Information Processing Systems, pp. 8778–8788.

Zhong Y, Deng W, Wang M, Hu J, Peng J, Tao X, Huang Y, 2019 Unequal-training for deep face recognition with long-tailed noisy data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7812–7821.

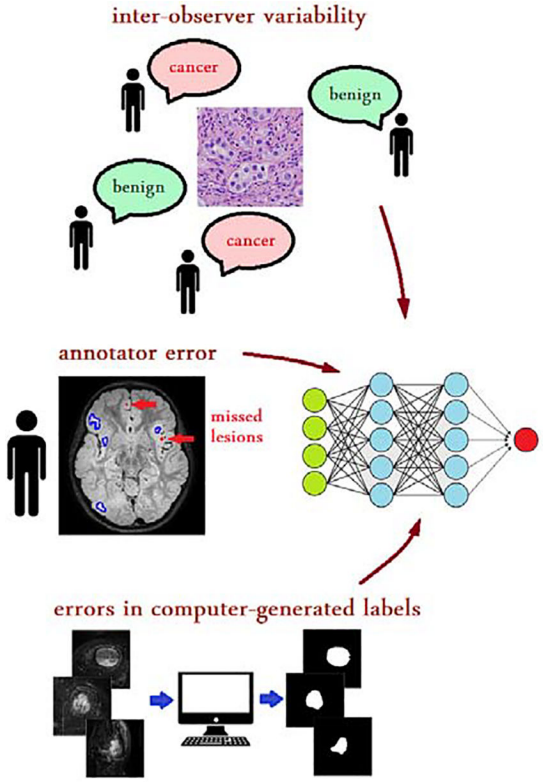
Zhou H, Sun J, Yacoob Y, Jacobs DW, 2017 Label denoising adversarial network (ldan) for inverse lighting of face images. arXiv preprint arXiv:1709.01993

Zhu H, Shi J, Wu J, 2019 Pick-and-learn: Automatic quality evaluation for noisy-labeled image segmentation. arXiv preprint arXiv:1907.11835.

Zhu X, Wu X, 2004 Class noise vs. attribute noise: A quantitative study. Artificial intelligence review 22, 177–210.

Highlights

- Supervised training of deep learning models requires large labeled datasets.
- Label noise can significantly impact the performance of deep learning models.
- We critically review recent progress in handling label noise in deep learning.
- We experimentally study this problem in medical image analysis and drawuseful conclusions.



Label cleaning and pre-processing
- Identifying and removing samples with incorrect labels
- Iterative label cleaning in parallel with training
- Label smoothing
- Label cleaning with generative models

Network architectures
- Estimating and utilizing noise statistics using
  - Noise layers
  - Probabilistic graphical models

Label-noise-robust loss functions
- Mean-absolute-error loss
- Cross entropy with abstention
- "Correcting" standard loss functions using label noise statistics

Data re-weighting
- Meta-learning
- Gradient scaling
- Learned re-weighting
- Learning to weight labels from different annotators

Data and label consistency
- Exploiting data similarity to identify incorrect labels
- Auxiliary image regularization
- Manifold regularization

Training procedures
- Curriculum learning
- Knowledge distillation
- Co-teaching
- Data augmentation (mixup)

Fig. 1.

Label noise is a common feature of medical image datasets. Left: The major sources of label noise include inter-observer variability, human annotator's error, and errors in computer-generated labels. The significance of label noise in such datasets is likely to increase as larger datasets are prepared for deep learning. Right: A quick overview of possible strategies to deal with, or to account for label noise in deep learning.

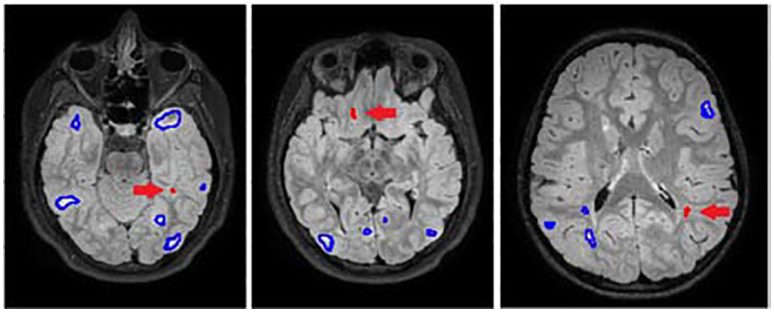


Fig. 2.

The FLAIR images from three TSC subjects and the lesions that were detected (in blue) and missed (in red) by an experienced annotator in the first reading.

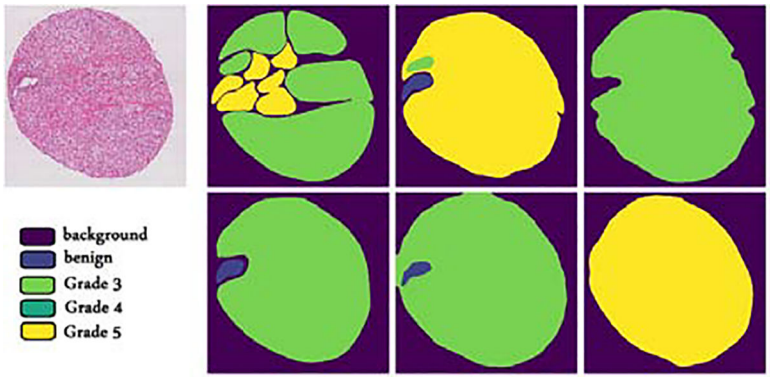


Fig. 3.

An example TMA core image (top left) and annotations from six pathologists (right) with four labels: benign and Gleason cancer grades 3, 4, and 5.

Fig. 4.

Examples of the annotator confusion matrices estimated by the method of Tanno et al. (2019) on the prostate cancer digital pathology data. In each matrix, rows represent for the estimated true label and columns represent the annotator's labels. Classes are in this order: benign and Gleson grades 3–5.

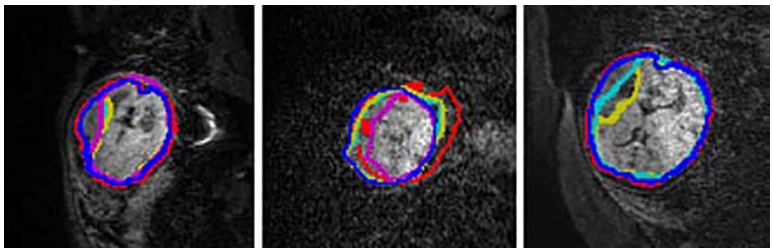


Fig. 5.
Examples of DW-MR fetal brain images along with manual segmentation (blue) and several noisy segmentations (other colors).

Table 1.

Summary of the main categories of methods for learning with noisy labels, representative studies, and potential applications in medical image analysis. The left column indicates the six categories under which we classify the studies reviewed in Sections 2 and 3. The middle column lists several representative studies from the fields of machine learning and computer vision and the applications considered in those studies. The right column suggests potential applications for the methods in each category in medical image analysis. In this column, where applicable, we have cited relevant published studies from the field of medical image analysis and experiments reported in Section 5 of this paper as examples of the application of methods adapted or developed in each category.

Methods category	Representative studies from machine learning and computer vision literature	Potential applications in medical image analysis
Label cleaning and pre-processing	Ostyakov et al. (2018) - image classification Lee et al. (2018) - image classification Northcutt et al. (2017) - image classification Veit et al. (2017) - image classification Gao et al. (2017) - regression, classification, semantic segmentation	most applications, including disease and pathology classification (Pham et al. (2019); experiments in Section 5.2) and lesion detection and segmentation (experiments in Section 5.1)
Network architecture	Sukhbaatar and Fergus (2014) - image classification Vahdat (2017) - image classification Yao et al. (2018) - image classification	lesion detection (Dgani et al. (2018)), pathology classification (experiments in Section 5.2)
Loss functions	Ghosh et al. (2017) - image and text classification Zhang and Sabuncu (2018) - image classification Wang et al. (2019b) - image classification, object detection Rusiecki (2019) - image classification Boughorbel et al. (2018) - electronic health records Hendrycks et al. (2018) - image and text classification	lesion detection (experiments in Section 5.1), pathology classification (experiments in Section 5.2), segmentation (Matuszewski and Sintorn (2018); experiments in Section 5.3)
Data re-weighting	Ren et al. (2018) - image classification Shu et al. (2019) - image classification Khetan et al. (2017) - image classification Tanno et al. (2019) - image classification Shen and Sanghavi (2019) - image classification	lesion detection (Le et al. (2019)) and segmentation (experiments in Section 5.1), lesion classification (Xue et al. (2019); experiments in Section 5.2), segmentation (Zhu et al. (2019); Mirikharaji et al. (2019))
Data and label consistency	Lee et al. (2019) - image classification Zhang et al. (2019) - image classification Speth and Hand (2019) - facial attribute recognition Azadi et al. (2015) - image classification Wang et al. (2018c)- image classification Reed et al. (2014) - image classification, emotion recognition, object detection	lesion detection and classification, segmentation (Yu et al. (2019a))
Training procedures	Zhong et al. (2019) - face recognition Jiang et al. (2017) - image classification Sukhbaatar and Fergus (2014) - image classification Han et al. (2018b) - image classification (Zhang et al., 2017) - image classification Acuna et al. (2019) - boundary segmentation Yu et al. (2018) - boundary segmentation	most applications, including segmentation (experiments in Section 5.3; Min et al. (2018); Nie et al. (2018); Zhang et al. (2018)), lesion detection (experiments in Section 5.1), and classification (Fries et al. (2019))

Table 2.

Performance metrics (DSC and lesion-count F1 score) obtained in the experiment on TSC brain lesion detection using different techniques listed in Section 5.1.2 compared with the baseline models trained with noisy labels (i.e., Faster-RCNN trained on noisy labels, and 3D U-Net CNN) and baseline models trained on clean data (i.e., Faster-RCNN trained on clean data, and 3D U-Net trained on clean data). The best performance metric value (in each column) has been highlighted in bold. The results show that in this application methods based on data re-weighting and iterative label cleaning substantially improved the performance of the CNNs trained with noisy labels. The best results in terms of both the DSC and the lesion-count F1 score were obtained from our 3D U-Net with iterative label cleaning.

Method	DSC	lesion-count F1 score
Faster-RCNN trained on noisy labels	-	0.541
Faster-RCNN trained on clean data	-	0.553
Faster-RCNN trained with MAE loss (Ghosh et al., 2017)	-	0.582
3D U-Net CNN	0.584	0.747
3D U-Net CNN trained on clean data	0.578	0.743
3D U-Net CNN trained with MAE loss	0.541	0.695
3D U-Net CNN trained with iMAE loss	0.485	0.657
3D U-Net CNN with data re-weighting	0.600	0.802
3D U-Net with Iterative label cleaning	**0.605**	**0.819**

Table 3.

Results of the experiment on prostate cancer digital pathology classification using different methods. The highest accuracy in each classification task (column) has been highlighted in bold text.

	Cancerous vs. benign		High-grade vs. low-grade		
Method	**accuracy**	**AUC**	**accuracy**	**AUC**	**Percentage of large classification errors**
Single pathologist	0.80	0.78	0.65	0.61	0.07
Majority vote	0.86	0.87	0.73	0.74	0.03
STAPLE	0.84	0.86	0.73	0.72	0.03
STAPLE + iMAE loss	**0.93**	0.91	0.76	0.79	0.03
Minimum-loss label	0.88	0.88	**0.80**	**0.82**	0.03
Annotator confusion estimation	0.92	**0.93**	**0.80**	**0.82**	**0.01**
STAPLE (3–3)	0.86	0.86	0.69	0.70	0.02
STAPLE + iMAE loss (3–3)	0.90	0.88	0.75	0.78	0.02
Annotator confusion estimation (3–3)	0.90	0.88	0.73	0.76	0.03

Table 4.

Comparison of different methods for fetal brain segmentation in DW-MR images in terms of DSC for different levels of label noise. The highest DSC scores have been highlighted in bold text for each noise level. Our dual CNNs with iterative label update generated the highest DSC scores at small and medium noise levels, whereas the CNN trained with MAE loss generated better results for high noise levels.

	Clean data	**noise level 1 (Method 1)**	**noise level 2 (Method 2)**	**noise level 3 (Method 3)**	**noise level 4 (Method 2)**	**noise level 5 (Method 3)**	**noise level 6 (Method 2)**	**noise level 7 (Method 2)**
Average DSC of the training labels	1.000	0.949	0.924	0.854	0.807	0.790	0.777	0.742
Baseline CNN	0.878	0.889	0.862	0.846	0.755	0.730	0.736	0.724
Baseline CNN trained with MAE loss	-	0.881	0.864	0.840	0.780	0.741	**0.778**	**0.760**
Dual CNNs with iterative label update	-	**0.906**	**0.895**	**0.886**	**0.849**	**0.804**	0.773	0.732

Table 5.

More detailed performance measures for fetal brain segmentation in DW-MRI. According to detailed analysis by three performance measures at different noise levels, our proposed dual CNNs with iterative label update outperformed both the baseline CNN and the baseline CNN trained with the MAE loss.

	noise level 1 (Method 1)			**noise level 3 (Method 3)**			**noise level 5 (Method 3)**		
	DSC	**5% DSC**	**HD95 (mm)**	**DSC**	**5% DSC**	**HD95 (mm)**	**DSC**	**5% DSC**	**HD95 (mm)**
Baseline CNN	0.89 ± 0.06	**0.80**	5.9 ± 2.6	0.85 ± 0.08	0.73	6.8 ± 2.6	0.73 ± 0.10	0.60	8.0 ± 4.9
Baseline CNN trained with MAE loss	0.88 ± 0.07	0.79	**5.6 ± 2.3**	0.84 ± 0.08	0.72	6.2 ± 2.5	0.74 ± 0.10	0.61	8.2 ± 3.6
Dual CNNs with iterative label update	**0.91 ± 0.06**	0.79	5.6 ± 2.4	**0.89 ± 0.08**	**0.75**	**6.0 ± 2.6**	**0.80±0.11**	**0.63**	**7.8 ± 4.0**