Improving Medical Images Classification With Label Noise Using Dual-Uncertainty Estimation

Lie Ju, Xin Wang, Lin Wang, *Graduate Student Member, IEEE*, Dwarikanath Mahapatra, Xin Zhao, Quan Zhou, *Member, IEEE*, Tongliang Liu, *Senior Member, IEEE*, and Zongyuan Ge, *Senior Member, IEEE*

***Abstract*— Deep neural networks are known to be data-driven and label noise can have a marked impact on model performance. Recent studies have shown great robustness to classic image recognition even under a high noisy rate. In medical applications, learning from datasets with label noise is more challenging since medical imaging datasets tend to have instance-dependent noise (IDN) and suffer from high observer variability. In this paper, we systematically discuss the two common types of label noise in medical images - disagreement label noise from inconsistency expert opinions and single-target label noise from biased aggregation of individual annotations. We then propose an uncertainty estimation-based framework to handle these two label noise amid the medical image classification task. We design a dual-uncertainty estimation approach to measure the disagreement label noise and single-target label noise via improved Direct Uncertainty Prediction and Monte-Carlo-Dropout. A boosting-based curriculum training procedure is later introduced for robust learning. We demonstrate the effectiveness of our method by conducting extensive experiments on three different diseases with synthesized and real-world label noise: skin lesions, prostate cancer, and retinal diseases. We also release a large re-engineered database that consists of annotations from more than ten ophthalmologists with an unbiased golden standard dataset for evaluation and benchmarking. The dataset is available at https://mmai.group/peoples/julie/.**

***Index Terms*— Label noise, uncertainty estimation, skin lesions, prostate cancer, retinal diseases.**

I. Introduction

DEEP learning is data-driven, and its success is largely attributed to sufficient human-annotated datasets. However, it is laborious to label massive data and maintain high-quality for the labels. In many applications, labels are acquired from non-experts (e.g., Amazon Mechanical Turk [1]) and sometimes automatically generated from the source information (e.g., downloading from social media with tags [2], extracting labels for X-ray images from associated radiology reports [3], [4]). These processes could introduce potential error or label noise into the model training. Learning from noisy labels is a long-standing challenge and has been well-recognized in the classical image recognition problem. However, learning from noisy labels has not been addressed well in the medical image analysis domain [5], [6].

Although some previous works [7]–[10] have had relative success to alleviate this issue, due to the domain bias and unique challenges existing in the medical imaging domain, it has limited scope for medical applications. First, most explored methods [9], [10] which focus on class-conditional noise (CCN), i.e., only captures the general label flipping patterns between classes for all instances (symmetric noise), or between similar classes (asymmetric). However, label noise in medical images is normally biased from the observer variability. Specifically, not all instances in a category have the equal possibility to be wrongly assigned a label to another similar category, resulting in an instance-dependent noise (IDN) [11]–[15]. Second, medical images can always be labeled by multiple experts and label noise is derived from the disagreement on inconsistency opinions. The existing methods show limited ability to handle this kind of noise. Third, medical images are always small and highly class-imbalanced, most methods tended to select noisy samples out do not work well since *noisy samples* can be easily confused with *hard samples* (minority class but with clean labels) and which are eventually detected as outliers and discarded from training [16], results in a waste of data utilization.

In this paper, we propose a dual-uncertainty-based framework for improving medical imaging classification with label noise. As Fig. 1 shows, the label noise we are trying to tackle includes the following two types: (1) *disagreement:*

Manuscript received November 21, 2021; revised December 28, 2021; accepted January 5, 2022. Date of publication January 7, 2022; date of current version June 1, 2022. The work of Quan Zhou was supported by the National Natural Science Foundation of China under Grant 61876093. The work of Tongliang Liu was supported by the Australian Research Council Project DE-190101473. *(Corresponding author: Zongyuan Ge.)*

Lie Ju and Zongyuan Ge are with the Faculty of Engineering, Monash University, Melbourne, VIC 3800 Australia, also with the Monash Medical AI Group, eResearch Center, Monash University, Melbourne, VIC 3800 Australia, and also with Airdoc, Beijing 100089, China (e-mail: julie334600@gmail.com; zongyuan.ge@monash.edu).

Xin Wang and Xin Zhao are with Airdoc, Beijing 100089, China (e-mail: wangxin@airdoc.com; zhaoxin@airdoc.com).

Lin Wang is with the Monash-Airdoc Joint Research Group, eResearch Center, Monash University, Melbourne, VIC 3800 Australia, and also with the College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, Heilongjiang 150001, China (e-mail: wanglin.mailbox@gmail.com).

Dwarikanath Mahapatra is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: dwarikanath.mahapatra@inceptioniai.org).

Quan Zhou is with the National Engineering Research Center of Communications and Networking, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: quan.zhou@njupt.edu.cn).

Tongliang Liu is with the Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: tongliang.liu@sydney.edu.au).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2022.3141425, provided by the authors.

Digital Object Identifier 10.1109/TMI.2022.3141425

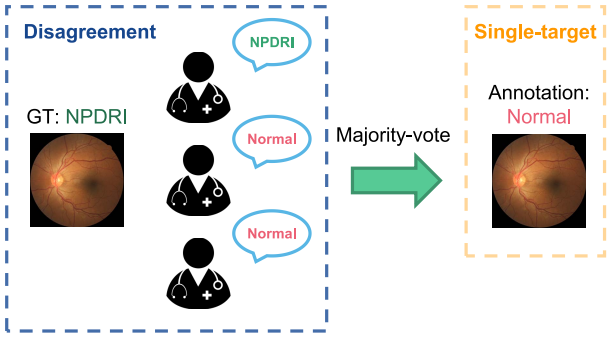


Fig. 1. The illustration of two common types (disagreement and single-target) of label noise in medical images. For instance, the ground truth of the sample is non-proliferative diabetic retinopathy-I (NPDRI) but different opinions from multiple doctors are obtained and result in the *disagreement label noise*. Then, we apply majority-vote but still get the a wrong diagnosis, which is the *single-target label noise*.

Annotations come from multiple doctors with various level of experiences, there exist disagreements and leads to observer variability; (2) *single-target:* Annotations present wrong diagnostic results in nature from sole opinion. Particularly, following the principles of reducing the impact of label noise during model training, we aim to detect those noisy samples explicitly. We first propose the improved Direct Uncertainty Prediction (iDUP) to capture prior knowledge to distinguish the *disagreement* type noise from the noise-free label. Then, we use the uncertainty measurement model such as MC-Dropout [17] to capture the predictive uncertainty for adjudicated *single-target* label using majority-vote. This will assist in further finding out the samples with noise on single-target labels.

Previous works [18], [19] tend to keep the low-loss/high-certainty samples retained for training and would sometimes mistakenly remove noise-free samples coming from minority class. In our approach, we choose to re-weight the samples according to a normalized uncertainty score, which helps preserve the impact of the minority class and the hard samples. Finally, a boosting-like training procedure is proposed to learn from the re-weighted samples in an intuitive and data-driven way through feeding different types of samples (clean and re-weighted samples) we have distinguished via dual-uncertainty estimation.

Our contributions from this work are summarized as follows:

1) We explore and provide metrics for two common types of label noise in the medical images from the perspective of **uncertainty**: *disagreements* from multiple doctors and *single-target* noisy labels.
2) We propose a novel dual-uncertainty-based framework which can significantly improve the robustness of the model identifying samples with label noise through measuring the uncertainty of samples. Furthermore, a boosting training procedure is proposed to handle the class-imbalance in the medical dataset.
3) We conduct extensive experiments on various imaging modalities from three kinds of diseases: skin lesions (ISIC 2019) [21], prostate cancer (Gleason 2019) [22] and retinal diseases (Kaggle DR+) [23].
4) We will release a large annotation-database containing relabelled results from more than ten doctors covering 17 retinal diseases based on Kaggle DR dataset [23]. Moreover, a corresponding unbiased golden standard for evaluation and benchmark will be provided. We hope this dataset would help address the challenges of label noise in the medical image computing and computer-assisted interventions community.

II. Related Work

A. Deep Learning With Noisy Labels

Deep learning with noisy labels has been an active topic in recent years and many works have been proposed to reduce the impact of label noise for practical applications [24]–[26]. Here, we divide the relevant methods into two main groups: sample-based and model-based.

Sample-based methods aim to find out those samples whose labels are likely to be corrupted, and we can have the choice of filtering those noisy samples out or relabelling them. Various methods such as k-nearest neighbor, outlier detection, and Gaussian Mixture Model (GMM) have been widely exploited to select false-labeled samples from noisy training data [2], [9], [18], [27]. Through clustering the training data to clean samples and noisy samples, Huang *et al.* [18] remove the noisy samples and only train from clean data. Yang *et al.* [27] refer to Co-teaching [19] and train two models to improve the robustness under a high noisy level. Some other works do not remove those noisy samples out: Guo *et al.* [2] design a curriculum learning-based training procedure to help model learn the clean samples at an early stage and reduce the over-fitting to noisy data. DivideMix [9] transfers the noisy label challenge to a semi-supervised learning problem and leverage MixMatch [28] technique to more effectively learn from noisy samples. Besides, since noisy labels are corrupted from clean labels, it has the potential to recover some useful information out of it. A typical approach for noisy labels refinement is *label transition matrix estimation*. The transition matrix denotes the transition relationship from clean labels to noisy labels and the clean/noisy class posterior can be inferred using clean/noisy data [29]. There are also some meta-learning techniques [7], which may require some small extra clean data for reference.

Model-based methods target learning to improve the robustness of the model for the noisy data. We review relevant works that consider network architectures, loss functions, and regularization terms. [30]–[34] aim to design loss functions that achieve a slight risk for unseen clean data even when noisy labels exist in the training data. For architectural modifications, a *noise adaptation layer* [35] is proposed to be added to the end of the network, which is equivalent to multiplication with the transition matrix between noisy and true labels. Zhong *et al.* [36] devised a graph convolutional network to correct noisy labels. Regularization terms are usually applied to reduce the over-fitting effect and thus improve the generalization of the model. Recent works propose advanced regularization techniques such as *mixup* [37] and *label smooth-*

ing [38], that can further improve model robustness to the label noise.

B. Medical Image Analysis With Noisy Labels

Unfortunately, not many studies have addressed medical image classification with label noise. Here we review some works from medical imaging classification and explore some relevant potential approaches. Pham *et al.* [39] use label smoothing technique [38] to improve the classification of thoracic diseases from chest x-rays in the CheXpert dataset [4]. Ghesu *et al.* [40] propose uncertainty-driven bootstrapping to filter training samples with the highest predictive uncertainty and improve robustness and accuracy on the ChestX-Ray8 dataset [3]. For skin lesion classification in dermoscopy images, Xue *et al.* [16] proposed an online uncertainty sample mining method and a sample re-weighting strategy to preserve the usefulness of correctly-labeled hard samples. Dgani *et al.* [41] leverage the noise adaptation layer [42] on mammography classification task and outperform standard training methods.

Although some approaches have been considered to address the noisy label issue, we still observe a noticeable gap in applying those methods to medical image classification with label noise. Some of those methods will serve as baselines in this work. Meta-learning-based approaches [7], [20] require a small clean validation dataset on the side to adjust the weights for samples in a mini-batch manner. CurriculumNet [2] leverages density-distance [43] and KMeans [44] to cluster samples to different subsets according to the complexity, then perform training from the subsets respectively. However, the functional ability of CurriculumNet on a small dataset with high-similarity has not been evaluated yet. Co-teaching [19] builds two CNN models, and the samples with the smallest forward-propagation loss in a mini-batch are fed into the other network for the next round of training. However, in the medical dataset, the amount of dataset of each class is usually imbalanced, and those high loss samples are always from minority class and treated as hard samples. Removing those samples would only make the situation worse. Xue *et al.* [16] do not only remove the noisy samples but also re-weight all samples using a probabilistic Local Outlier Factor algorithm (pLOF) [45] to retain the hard samples. DivideMix [9] is proposed to find out potentially noisy samples through GMM and then transfer it into a semi-supervised problem. However, there still lacks direct insights for handling imbalanced data, and the calculated-weights are not available for the "real" unlabeled data. In Table I, we give a summary of those methods. **Sample Selection** denotes the auxiliary technique being used for sample selection; **Hard Sample** denotes whether the method is capable of handling the hard samples appropriately during the sample selection process; **Clean Data** indicates whether the method requires extra clean data for training or model calibration.

TABLE I
THE SUMMARY OF SOME METHODS FOR SOME TYPICAL ISSUES

Method	Sample Selection	Hard Sample	Clean Data
[7], [20]	Meta-Learning	✓	✓
[2]	Subsets Clustering	✓	×
[19]	Loss Ranking	×	×
[18]	Learning Rate Adjustment	×	×
[16]	Outlier Detection	✓	×
[9]	Gaussian Mixture Model	×	×
Ours	Dual-uncertainty Estimation	✓	×

*__Hard Sample__ denotes the ability of the method to handle hard samples.
*__Clean Data__ denotes whether the method requires the clean data.

C. Uncertainty Estimation

There are two types of predictive uncertainty which are mainly studied by previous works: *epistemic uncertainty* and *aleatoric uncertainty* [46]. Monte Carlo Dropout (MC Dropout) [17] is proposed to estimate uncertainty with dropout NNs using approximate Bayesian inference in deep Gaussian processes. Lakshminarayanan *et al.* [47] proposed a Deep Ensemble-based method for the predictive uncertainty estimation. This method can be well implemented on those architectures without Dropout layers.

Recently, uncertainty estimation is widely leveraged in medical imaging tasks. Zhong *et al.* [48] proposed an uncertainty-aware INVASE to quantify predictive confidence of healthcare problems. Wang *et al.* [49] adopted the matting technique into lesions segmentation to handle the uncertain regions in medical scenes. Ghesu *et al.* [50] presented a method that can generate both an image-level probability and a corresponding uncertainty measure using principles of information theory and subjective logic [51]. Ayhan *et al.* [52] proposed an uncertainty estimation method that simulated the scenario when doctors give diagnoses under different environments by doing multiple data augmentations on raw input for the model. In addition to those predictive uncertainty estimation methods, Raghu *et al.* [53] proposed Direct Uncertainty Estimation which can give an unbiased estimation of uncertainty when there is disagreement among the experts.

In this work, we find MC-Dropout is a simple but effective method for uncertainty estimation and show a promising ability to distinguish among clean/noisy samples. Also, the Dropout layer is regarded as a normalization technique that can be inserted into most CNN architectures and no extra model training is required. To handle the disagreement from observer variability, we improve the DUP with the consideration of the number of annotators.

III. METHODOLOGY

A. Problem Definition

Given the i_{th} sample x_i in the dataset $X = [x_1, x_2, \ldots, x_z]$ from the real clinical scenario, we have pairs of the form (sample, annotations from multiple doctors), $(x_i; y_i^1, y_i^2, \ldots, y_i^{n_i})$, where n_i denotes the number of annotations received for the current sample. For example, $n_i = 1$ means the sample is labeled by only one doctor while $n_i > 1$ indicates the sample is labeled by multiple doctors. Label noises may occur due to the inconsistency opinions from doctors or simply wrong annotation.

Hence, under the circumstances where samples with label noise are presented, our goal in this work is to train a classification model h with parameters θ from modified training samples $\hat{X} = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{\hat{z}}]$ based on our proposed

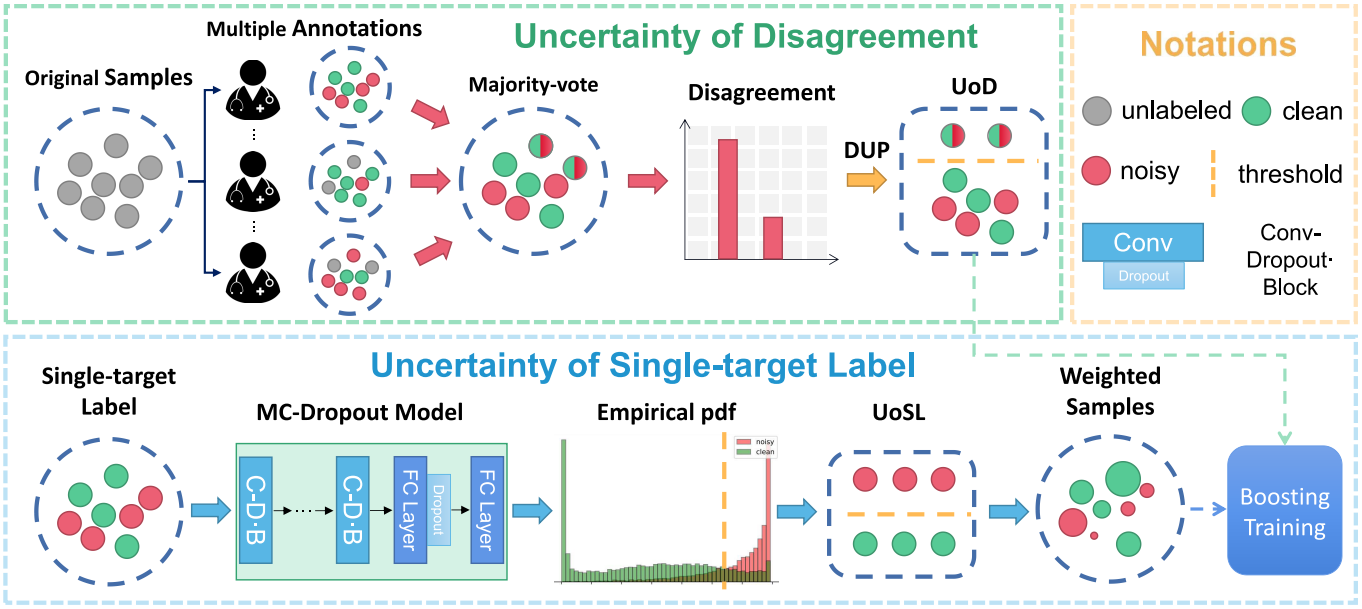


Fig. 2. The overview of our proposed dual-uncertainty estimation framework. For those samples with annotations from multiple doctors, we use improved Direct Uncertainty Prediction to estimate the **UoD** (uncertainty of disagreement), and a threshold is set to filter out those samples with high **UoD**. Then given a sample with an adjudicated single-target label, we leverage the Monte Carlo estimate MC-Dropout model to generate predictive uncertainty scores to estimate the **UoSL** (uncertainty of single-target label) and re-weight those potential noisy samples. Those selected samples will then be fed into a boosting training scheme to train a noisy label tolerant classifier. *C-D.B denotes a convolution block with Dropout layer.

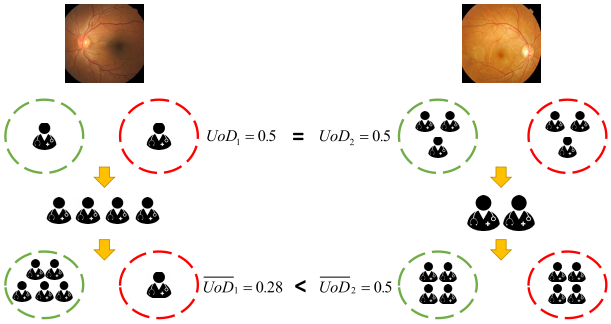


Fig. 3. The limitations of DUP. If samples are labeled by different number of doctors, DUP may can not estimate the real uncertainty of disagreement. In this example, considering a binary classification problem, we use green and red cycle to denote the categories.

selection/weighting criteria and evaluate its performance on an unbiased golden standard test set. We give an overview of our proposed framework in Fig. 2

B. Dual-Uncertainty Estimation

We carry out two kinds of uncertainty as Fig. 1 shows. We define the first one as Uncertainty of Disagreement (UoD) which signifies expert disagreement among multiple doctors. And the second one is defined as Uncertainty of Single-target Label to signify the noise of single adjudicated annotation. We propose to use dual-uncertainty estimation for modeling these two uncertainties respectively.

1) Uncertainty of Disagreement: By applying aggregation methods such as majority voting on a sample with high disagreement among annotators, it tends to generate noisy training labels. However, we may still extract useful information or clues out of these annotations from those disagreements to detect hard samples vs wrong labels since those samples with high disagreement are considered to be easily misdiagnosed by human-experts, also difficult to be learned by DL models. Suppose we can find an uncertainty estimation approach U_1 to map and quantify the annotations $Y_i = [y_i^1, \ldots, y_i^{n_i}]$ for the sample x_i to the UoD, and we can handle those samples with an uncertainty metric and then can be served as a kind of prior knowledge.

To provide a quantified score for the uncertainty of disagreement, Direct Uncertainty Prediction (DUP) [53] is considered to be an unbiased estimate of the true uncertainty which is based on the theory of *Modeling Labeler Bias* [54]. Letting $G = [g_1, g_2, \ldots, g_k]$ denotes the different diseases/grades (e. g,. $k = 5$ represents there are five levels of severity for DR [23]), and $P_i = [p_i^1, p_i^2, \ldots, p_i^k]$ represents the empirical grade distribution, also known as the empirical histogram for the sample x_i is calculated as:

$$p_i^c = \frac{\sum_j \mathbb{1}_{y_i^j = g_c}}{n_i}. \tag{1}$$

By applying DUP, the original UoD is computed as:

$$UoD_i = U_1(x_i) = U(P_i) = 1 - \sum_{j=1}^{k} (p_i^j)^2, \tag{2}$$

Although DUP is proven to be effective to handle the UoD in a dataset with annotations from multiple doctors, it has obvious limitations. For example, given two samples x_1 and x_2 labeled with $Y_1 = [0, 1]$ and $Y_2 = [0, 0, 0, 1, 1, 1]$ from 2 and 6 doctors respectively, we can calculate the empirical

histogram as $P_1 = [0.5, 0.5]$ and $P_2 = [0.5, 0.5]$. For two samples, they obtain the same UoD score: $1 - 0.5^2 - 0.5^2 = 0.5$. However, there is a possibility that the latter should be more uncertain than the former. As Fig. 3 shows, if extra doctors are introduced, the results of UoD may change, and DUP do not have the ability to handle this potential uncertainty.

Here, we propose **improved DUP (iDUP)** to take the variance of the number of annotations into consideration by adding a factor to UoD:

$$iUoD_i = [min(\sum_j \mathbb{1}_{\mathrm{y_i^j = g_c}})]^{\eta} \cdot UoD_i, \tag{3}$$

where η is a hyper-parameter. We use the value of the category with the smallest number (except 0) of votes as a factor.

Since the iDUP is training-free, we can previously set a threshold t_1 to select a fraction of training samples with relatively high uncertainty (UoD $> t_1$) to be eliminated. With pre-calculated UoD from Eq. (3) as the prior knowledge, we can know which samples are likely to be misdiagnosed by human experts since those high-uncertainty samples are most likely to be noisy samples, which are also difficult to be learned or fitted by deep learning models.

2) Uncertainty of Single-Target Label: Estimation of UoD learns an uncertain score from the raw data pairs and has provided a kind of prior knowledge for sample selection. However, we have not tackled the single-target label noise for those samples where the diagnosis from single or the majority-vote result still can be wrong. Previous works [2], [9], [16], [18]–[20] proposed several techniques to reduce the impact of wrongly-labeled samples in trainset by sample selection. However, those methods lack indirect insights on the imbalanced attribute that existed in the medical imaging domain, as we discussed in Sec. II-B. To put the data imbalance attribute into consideration while addressing the single-target noisy label issue, we propose to use uncertainty estimation techniques (e, g., Monte-Carlo Dropout [17]). More specifically, we perform T-times stochastic forward pass on a trained CNN model under random dropout, then we have a pseudo-decision distribution from T "CNN doctors" with various degrees of disagreement. Based on this pseudo-decision distribution by several "CNN doctors", the model performs a Bayesian Approximation [17] to detect potential outliers that have a bad impact on the model's performance.

Formally, for sample x_i fed into the network with T-time forward pass, we obtain a list of T probability vectors for a subject x_i : $\{p^{(1)}{}_i, p^{(2)}{}_i, \ldots, p_i^{(T)}\}$. The uncertainty of a single-target label can be estimated using the mean predictive entropy [46]:

$$UoSL_i = -\sum_{j=1}^{k} m_i^j \cdot log(m_i^j) \tag{4}$$

where $m_i = \frac{1}{T}\sum_{t=1}^{T} p_i^t$ and j corresponds to the j-th class. Mean predictive entropy shows a stronger ability to make representations of variance from MC estimators compared to softmax distributions [46].

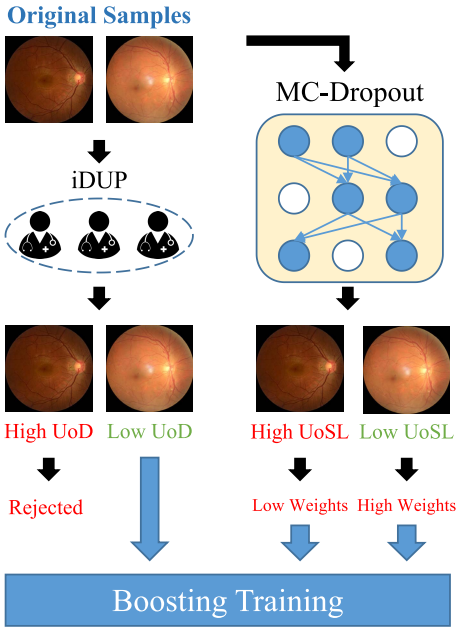


Fig. 4. The illustration of our sample selection strategy. We first use iDUP to inject those samples with high disagreement. Then, a MC-Dropout technique is leveraged for the reuse of injected samples by re-weighting the potential noisy samples with high uncertainty.

C. Boosting Training

1) Sample Selection: As shown in Fig. 4, our sample selection strategy consists of two components to maximize the use of original feature distribution instead of simply removing those high-uncertainty samples.

Given dual-uncertainty score UoD_i and $UoSL_i$ for each sample x_i, we propose to use these uncertainty measurements for sample selection and re-weighting [55] at each training iteration. Firstly, for those samples with multiple annotations, we obtain UoD_i. Specially, $UoD_i = 1$ for those sample with single-target label from only one doctor. A threshold t_{UoD} is set to filter out samples with a high score of UoD_i. Those samples generally show significant controversies from the annotators and are diminishing for the model training. We then reorder the samples for each sample x_i in X according to the results of UoD estimation as follows: $UoD_1 \leq UoD_2 \leq \ldots \leq UoD_z$. Given a threshold t_{UoD} to reject those high-disagreement samples, we have the retained $\hat{n}_i$ out of n_i samples $[\hat{X}, \hat{Y}] = [(\hat{x}_1; \hat{y}_1), (\hat{x}_2; \hat{y}_2), \ldots, (\hat{x}_{\hat{n}_i}; \hat{y}_{\hat{z}})]$.

However, it is noted that the elimination of hard samples from the minority classes is still unavoidable since they are likely to receive a high UoD score. Also, the incomplete data prevent the model from learning abundant feature information from the original distribution. Then, a curriculum training strategy is proposed to reduce the impact of removing hard samples.

2) Curriculum Training: We feed the selected clean samples $\hat{X}$ to the model at an early stage and then we have the network to begin learning all samples with normalized weights in a curriculum-driven way [8]. Overall, the network is jointly

optimized by the following loss functions:

$$L_{FL}(\hat{X}) = -\sum_{i=1}^{\hat{z}}(1 - p_i \cdot q_i)^{\gamma} \cdot log(p_i), \tag{5}$$

$$L_{wCE}(X) = -\sum_{i=1}^{z} w_i \cdot (q_i \cdot \log p_i + (1 - q_i) \cdot \log(1 - p_i)) \tag{6}$$

$$L = \alpha \times L_{FL}(\hat{X}) + \beta \times L_{wCE}(X), \tag{7}$$

where p_i denotes the predictions and q_i denotes the ground-truths. α and β are hyper-parameters as loss weights. We leverage the Focal Loss [56] in Eq. 5 for reducing the impact of imbalance. There are two reasons for using Focal Loss. First, the samples from the minority classes tend to produce a relatively low probability. Focal Loss can help to down-weights the loss assigned to well-classified examples and focus more on those low-probability samples. Second, Focal Loss is introduced as an Online Hard Samples Mining (OHEM) technique. Label noise in medical images is asymmetric (i.e., class-conditional) due to the high similarity between categories which can be easily mis-classified. The model could be quickly overfitted to the wrong information at an early stage and give low-entropy (high confidence) predictions [9]. In this way, Focal Loss can avoid the over-weighting on those noisy samples with the mining of those clean hard samples.

Also, the weights driven from the uncertainty measurement $W = [w_1, \ldots, w_i, \ldots, w_z]$ are assigned to the Cross-Entropy Loss for all samples, where $nUoSL$ is the normalized $UoSL$ mapped to [0, 1]:

$$w_i = \begin{cases} 1 - nUoSL_i, & nUoSL_i > t_{UoSL} \\ 1, & nUoSL_i \leq t_{UoSL} \end{cases} \tag{8}$$

In the early stage of training, we use a small β value and only focus on training retained clean samples $\hat{X}$ selected by iDUP (Eq. (3) & Eq. (5)). Then we increase the weights of the second loss function (Eq. (6)), and the model starts to learn the whole distribution in a re-weighting manner.

The advantages of using the re-weighted training loss can be summarized as follows: first, model learns to generalize on the entire dataset with complete feature distribution, and the risk of discarding of the hard samples in the minority classes can be reduce; second, we do not need to select another threshold value for sample selection so no extra hyper-parameter is introduced.

IV. Experiments

In this section, to demonstrate the complex scenarios in the clinical collection of datasets, we conduct the experiments on three medical imaging datasets, including datasets with synthesized IDN: (1) ISIC 2019 dataset [21] for skin cancer diagnosis; and two real-world datasets: (2) Gleason 2019 [22] for prostate tissue microarray classification; (3) Kaggle+ fundus images database released by this paper, consists of 17 diseases and annotations from multiple ophthalmologists. In the experiments, we explore *single-target* noisy diagnosis scenario for the ISIC 2019 dataset and *disagreement/observer-variability* scenario for the Gleason 2019 and Kaggle+ fundus.

Algorithm 1 The Generation of Instance-Dependent Noise

Input: Training samples and clean associated annotations: $\{X_{train}, Y_{train}\}$; Test samples and clean associated annotations: $\{X_{test}, Y_{test}\}$; CNN $f(\cdot|\theta)$. Noise rate: ρ.

Output: Corrupted annotations for training samples: $\tilde{Y}_{train}$.

Initialize $\boldsymbol{\theta}$ randomly;
while not converged **do**
  Train $f(\cdot|\theta)$ using $\{X_{test}, Y_{test}\}$;
end
$\hat{Y}_{train} = f(X_{train}|\theta)$. (Inference training samples.)
$L_{train} = CE(\hat{Y}_{train}, Y_{test})$; (Compute losses on training samples.)
$L_{ranked} = Rank(L_{train}) = \{l_1, \ldots, l_{N_{train}}\}$; (Rank losses from high to low.)
$I = \{i_1, i_2, \ldots, i_{\rho * N_{train}}\}$; (Select top $\rho * N_{train}$ indexes.)
$\tilde{Y}_{train} = Corrupt(Y_{train}|I)$. (Change the original training annotations to a similar category.)

A. Implementation Details

1) Hyper-Parameters Setting: For experiments evaluated on ISIC dataset, we follow [18], [19] and use Dropout-ResNet-101[1] [57] and a 9-Layer CNN with dropout layer [2] as backbones for a fair comparison. For the other two datasets, we only use the ResNet-101 as the backbone. The batch sizes are set to 16. We only use horizontal flipping as the data augmentation strategy since the overuse of various data augmentation techniques may have a negative impact on the uncertainty measurement [52]. We apply the Adam optimizer for model training, and the initial learning rate is set to 3×10^{-4} and decayed by a factor of 0.5 when there are no more improvements in the performance with the patience of 5. The minimal learning rate is 1×10^{-7}. α is set as 1 and $\beta = (\frac{epoch_i}{epoch_{all}})^2$ with the curriculum training progress. $epoch_{all}$ is set as 5, 5 and 8 for ISIC 2019, Gleason 2019 and Kaggle DR+ respectively. All experiments were run by 5 trials with different random seed on 8 × NVIDIA GTX 1080Ti.

2) Warm-up: For the initial estimation of the predictive uncertainty score, the warm-up procedure is needed for a few epochs by training on all data points using the normal cross-entropy loss. Unlike classic imaging datasets, label noise in medical images is asymmetric (i.e., class-conditional) due to the high similarity between categories. The model could be quickly overfitted to the wrong information at an early stage and give low-entropy predictions. DivideMix [9] proposed to reduce the speed of overfitting to incorrect labels by adding a negative entropy term [38] to the loss function. In Fig. 5, we first conduct experiments on two warm-up strategies and show the comparative results between confidence penalty-based and uncertainty-based technique, as this will

[1] a dropout layer with the probability of 0.3 is inserted after every **Basic Layer**

[2] a dropout layer with the probability of 0.3 is inserted after every convolution layer and the first fully connected layer

reveal models' ability for noisy samples detection. We can see that although the confidence penalty prevents the model from generating low-entropy predictions, it makes the model difficult to learn the correct information from clean samples (see Fig. 5 - (a)) and a good trade-off strategy is needed between noisy samples and clean samples after assessing the uncertainty score.

3) Dataset Split: It is widely accepted that the model tends to learn clean samples in the early stage and gradually overfits to noisy samples when it comes to the end of the training, which results in a performance drop. We follow [9], [58] and report two kinds of results to evaluate the robustness of our proposed framework. For ISIC and Gleason dataset, we did not split the validation dataset. We report both the **best** test accuracy across all epochs and the averaged test accuracy over the **last** 3 epochs. Thus, the performance gap can be monitored. To reduce the bias caused by the way of noise generation for a fair comparison, we split the dataset into train:val:test = 7:1:2 in Kaggle DR+ and report the results on the test dataset with 4-fold cross-validation. These two evaluations are both reasonable for benchmarking the label noise challenge.

TABLE II
THE PERCENTAGE AUC RESULTS ON ISIC 2019

ISIC 2019 (Benign / Malignant)							
		ResNet-101			9 layers CNN		
		10%	20%	40%	10%	20%	40%
Cross-Entropy	B	82.63	80.73	78.27	77.50	76.23	76.12
	L	80.69	78.71	76.66	77.02	76.03	74.23
MixUp	B	81.99	80.83	79.49	76.78	76.05	76.17
	L	**81.59**	77.85	76.86	75.44	75.24	74.45
Label Smooth	B	81.24	80.56	79.65	77.68	77.70	77.00
	L	79.35	78.30	76.62	76.78	76.29	74.48
CurriculumNet	B	82.68	81.12	79.38	78.23	78.05	77.05
	L	79.12	76.23	73.01	72.39	74.30	71.09
Co-teaching	B	80.25	78.80	77.23	75.55	75.26	75.02
	L	79.64	77.61	75.11	70.28	74.56	73.14
DivideMix	B	73.12	71.59	71.44	73.77	74.55	70.46
	L	70.05	70.01	69.25	71.58	70.78	68.41
O2U-Net	B	77.14	76.37	74.04	75.99	75.05	72.14
	L	74.55	74.12	73.97	73.54	72.11	71.56
RW-OUSM	B	80.58	80.51	78.24	75.66	75.55	75.44
	L	79.25	79.37	77.75	73.36	74.85	74.37
Ours	B	**83.26**	**82.90**	**81.01**	**79.42**	**79.00**	**78.43**
	L	79.63	**79.04**	**78.66**	**79.11**	**78.93**	**76.21**

*B denotes the best results over all epochs.
*L denotes the average results of the last three epochs.

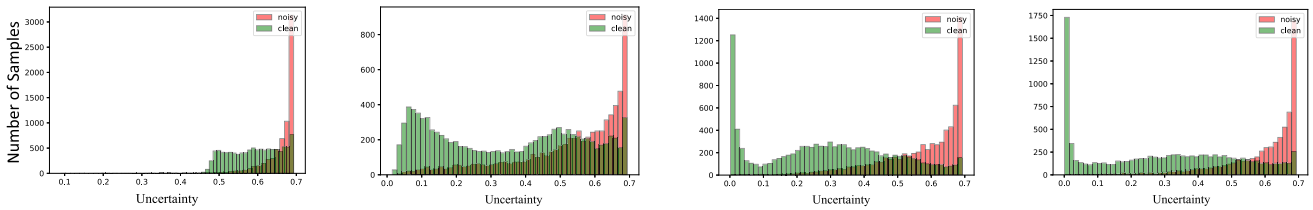
B. ISIC 2019

1) Dataset Statistics and Noise Synthesizing: The ISIC 2019 dataset contains dermoscopic images for skin cancer diagnosis across 8 different categories and each sample also gets a binary diagnosis result of benign or malignant. In this study, 25,331 images (16,971 benign and 8,360 malignant) are available for training. Regarding the synthesizing method, we followed [16] to build the IDN for the ISIC 2019 dataset based on the training loss, which is shown in Algorithm 1. First, we use the test samples to train the CNN, then we evaluate the training samples and get predictions. We calculate and rank the predictions loss, then we select $\rho * N_{train}$ samples with highest prediction loss according to its preseted noise rate ρ. Then we corrupt the training labels by change them to the conflicting category. e.g., 0 to 1 in a binary classification.

2) Comparative Study: We compare our proposed method with baselines [2], [18], [19], [37], [38] discussed in Sec. II-B and re-implemented using the same network backbone architecture for a fair comparative study. Although most methods are designed and evaluated for classic image recognition tasks, here we re-adapted them under well-adjusted hyper-parameters on our target medical dataset for a fair comparison. The comparison baselines are briefly introduced as follow:

- **MixUp** [37] was originally introduced as a regularization term to prevent memorization and sensitivity to adversarial examples. Recently, it has been widely used as an effective component in some robust learning works [9].
- **Label Smoothing [59]** can well help the model prevent from becoming over-confident to those incorrect labels by using soft targets.
- **CurriculumNet [2]** is inspired from Curriculum Learning [8] and learn to decide which samples to learn by measuring the complexity of data.
- **Co-teaching [19]** leverages the mutual information from two parallel networks and selects the high-confidence samples by ranking the losses output from the other network. Only those samples with smaller losses will be kept as clean samples for backpropagation.
- **DivideMix [9]** takes the advantages of semi-supervised learning [28] and divides the samples into labeled and unlabeled data using a Gaussian Mixture Model. Also, a co-guessing stage is used to assign new predicted labels for those potential noisy samples.
- **O2U-Net [18]** switches the model between the under-fitting and over-fitting status by adjusting the learning rate. The variance of losses for each sample will be calculated and ranked. Then, those potential noisy samples will be removed.
- **RW-OUSM [16]** proposed an online uncertainty sample mining method for measuring the probability of the noisy samples. Then, a sample re-weighting strategy is presented to preserve the usefulness of hard samples.

Here, we report the best percentage AUC score across all the epochs and average results over the last three epochs, to evaluate the robustness of our proposed framework, which are denoted by **B** and **L** respectively in Table II. The results show that our proposed framework exceeds all other baselines (in **bold**) with both the best and last epoch under all various noise ratios except 10%. The reason is that training on Cross-Entropy and MixUp methods does not change the original distribution of samples significantly, so they can achieve more stable results under the low noise rate, which is equivalent to being trained on a clean dataset. The performance difference between the best epoch and the last epoch demonstrates the robustness of each method. Without losing the performance of direct training, our proposed framework achieves 1.90% and 1.03% improvement compared to CurriculumNet under the noise rate of 10% and 20%, 0.28% compared to MixUp, 0.68% compared to Label Smoothing and 4.02% compared to CurriculumNet under the 40% label noise.



(a) Confidence Penalty Epoch 5 (b) MC-Dropout Epoch 1 (c) MC-Dropout Epoch 5 (d) MC-Dropout Epoch 15

Fig. 5. The visualization results on ISIC 2019. (a) Warmup using Confidence Penalty. (b) - (d) Warmup using MC-Dropout.

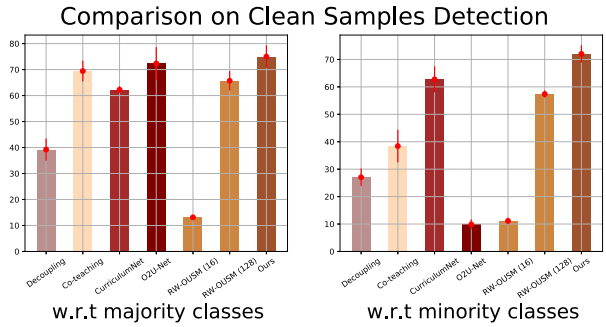


Fig. 6. Comparison of clean samples selection under the 40% noise rate. Most existing methods fail to detect the noisy samples out with respect to the minority classes.

When training on a simpler network (9 layers CNN), our methods still outperform all methods under all noise rate settings. It is found that the performance gap between the best and last epochs is smaller since those CNN architectures with fewer layers show limited ability to fitting the training samples including both clean and noisy samples.

3) The Precision of Sample Selection: The key reason for our approach in obtaining promising results is the sample selection strategy which prevents from wrongly removing the hard samples. To study and verify the superiority of our proposed sample selection strategy, we show the precision of clean samples detection in Fig. 6. We compare several popular sample selection methods for label noise learning such as Decoupling [60] and Co-teaching [19]. From this figure, most methods show good ability to detect noisy samples for majority classes, i.e., negative samples in this case. However, when it comes to minority classes, some of them fail to detect clean samples out since most samples from minority classes can always be regarded as noisy samples due to their high losses. It can be found that RW-OUSM is very sensitive to the batch size, e.g. 13.12% for 16 and 65.69% for 128. More samples in a mini-batch bring better representation for its modeling the label noise by Local Outlier Factor (LOF). For our proposed methods, we emphasize that we do not directly reject those high-uncertainty samples, we reuse them through a re-weighting manner instead. Here, to evaluate the ability of sample selection, we search a threshold to make a distinction between them, e.g., the median of uncertainty values in Fig. 5

In a conclusion, most sample selection-based methods without reuse of removed samples show poor ability when training from imbalanced data, resulting in poor performance. Those samples from minority classes can be regarded as hard samples in the model with a high loss and can be easily rejected by the loss-based sampling strategies. In contrast, those methods such as CurriculumNet that can learn the complete distribution of samples, i.e., train from all samples, can achieve more robust results. Our proposed methods combine the advantages of both strategies, also accompanied with the consideration of experts' disagreement.

C. Gleason 2019

1) Dataset Statistics: Gleason 2019 aims to classify prostate tissue microarray (TMA) cores as one of the four classes: benign and cancerous with Gleason grades 3, 4, and 5. TMA cores have been annotated in detail (i.e., pixel-wise) by six pathologists independently. Because pathologists who labeled this dataset have a different level of experience, ranging from 1 to 27 years, *uncertainty of disagreement* presents among the samples. It is also noted that not all images are labeled by the exact same doctors (i.e., doctor ID 6 only labeled 65 images out of all 244 images). To augment the number of available training and testing samples from the Gleason 2019 dataset, as suggested by [61], a patch-level classification system are normally built to verify the idea. As shown in Fig. 7, for patch generation, each whole-slide image is first resized to 3100 $\times$ 3100 pixels, then small image regions of size 750 $\times$ 750 are sampled from each TMA spot, using a step-size of 375 pixels. For samples selection and annotations, we have the following ties:

- If the areas of background $\geq$ 95% of the whole patch, those patches are removed during the training phase.
- If patches containing more than one annotations from a pathologist, we prefer to choose the grade with large areas as its label.
- If pathologists with different opinions are equal in the majority-vote, we choose the "higher grade". For example, if 3 pathologists say "Grade 4" and 3 pathologists say "Grade 5", and we choose Grade 5.
- For those samples with full agreement, we select some of them as the test set.

With the ties above, we have the processed-data statistics in Table. III. Since there is large observer-variability among six pathologists, we use the different sampling strategies to aggregate their opinions for different benchmark tests.

2) Comparative Study: In this section, we present the results of the comparative study on the discussed previous works as ISIC 2019 dataset. The grading of TMA is a multi-class classification task, we evaluate the model with three metrics: Recall, Precision and F1 Score under a macro manner, i.e., calculate

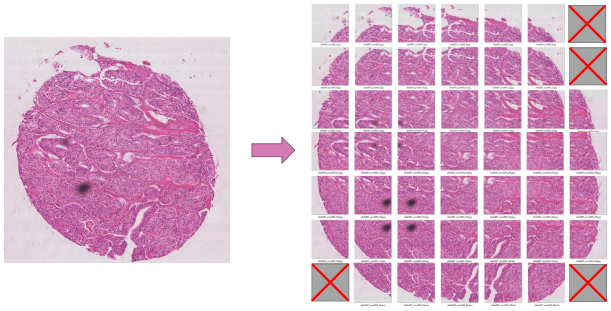


Fig. 7. The illustration of building patch-level TMA samples. For patch creation, original images are resized into 3100 × 3100 pixels, and small image regions of size 750 × 750 were sampled from each TMA spot, using a step of 375 pixels. If the areas of background ≥ 95% of the whole patch, those patches are removed during the training phase.

TABLE III
THE DATA STATISTICS ON GLEASON 2019. **ORIGIN** ARE COUNTED IN THE IMAGE-LEVEL AND OTHERS ARE IN PATCH-LEVEL

Target	Origin	Benign	Gleason 3	Gleason 4	Gleason 5	Sum
Pathologist 1	244	210	1513	4166	617	6506
Pathologist 2	141	614	612	827	369	2422
Pathologist 3	242	776	2165	3190	14	6135
Pathologist 4	244	1521	2183	3430	66	7200
Pathologist 5	246	940	2946	3411	99	7396
Pathologist 6	65	409	652	826	174	2060
Majority-vote	-	1835	3202	5342	255	10634
Test	-	470	781	1161	27	2439

metrics for each category, and find their unweighted mean. Since every image may contain more than one annotation from multiple pathologists, we use the majority-vote results as the aggregated inputs and evaluate the competitive robust learning methods. Besides, we follow [6] and test **Annotator confusion estimation (ACE)** [62] which estimates the labeling patterns of the annotators without direct aggregation. The overall results are shown in Table IV. An interesting phenomenon is that the performance of the last epochs may exceed that of the best epoch in terms of some metrics, e.g., 82.32% for last epochs and 79.39% for best epoch in terms of precision on Label Smooth. Hence, we select the best epochs based on the F1 Score, which is the harmonic mean of the precision and recall.

Similar to the experiments on ISIC 2019 dataset, those methods learn from the whole distribution without removing potential noisy samples obtain promising results. However, Co-teaching and DivideMix failed to maintain the accuracy level in terms of all metrics. There are two reasons. First, compared with ISIC 2019 dataset, Gleason 2019 dataset exhibits a more imbalanced status, e.g., 5342 samples for Gleason 4 but only 255 for Gleason 5. Those samples from minority classes are wrongly removed in an early stage. Second, Co-teaching relies on an estimated noise rate as a hyper-parameter for the threshold selection. However, the actual noise rate of real-world noisy data such as Gleason 2019 can not be observed or calculated since there are no available ground truth annotations for quantification. Our proposed methods achieve best results in terms of recall when

TABLE IV
THE RESULTS OF COMPARATIVE STUDY ON GLEASON 2019 WITH OTHER COMPETITIVE METHODS

Gleason 2019				
	Epoch	Recall	Precision	F1 Score
CE	Best	81.44	82.48	81.88
	Last	80.13	79.94	79.98
MixUp	Best	81.02	80.63	80.82
	Last	78.98	79.61	79.29
Label Smooth	Best	79.27	79.39	79.33
	Last	78.30	80.20	79.23
CurriculumNet	Best	81.48	82.50	81.98
	Last	81.57	79.98	80.76
Co-teaching	Best	79.60	73.37	76.36
	Last	80.51	71.18	75.56
DivideMix	Best	68.33	76.50	70.38
	Last	66.48	75.86	68.53
O2U-Net	Best	79.66	83.31	80.63
	Last	78.55	81.78	80.13
ACE	Best	80.31	81.86	81.08
	Last	80.21	80.98	80.59
Ours ($t_{UoD} = UoD_0$)	Best	**82.34**	84.03	82.50
	Last	**81.27**	**84.12**	**82.01**
Ours ($t_{UoD} = UoD_1$)	Best	82.33	**84.13**	**83.22**
	Last	79.01	79.73	79.37

t_{UoD} is set as the smallest UoD value from all re-ordered UoD values, i.e., $UoD_0 = 0$. That is to say, only those samples that reach full agreement are kept for the next stage of training, which also promises stable training and the performance does not drop so much in the last epoch (from 82.50% to 82.01% in terms of F1 score). By comparison, we choose the second smallest UoD value as the threshold, i.e., $t_{UoD} = UoD_1$. In this case, although more samples used for training are kept and the performance of the best epoch has been slightly improved (from 82.50% to 83.22% in terms of F1 score), the noise rate is unavoidably increased, which leads to a loss of performance (from 83.22% to 79.37% in terms of F1 score). Another interesting finding is that the performance of using cross-entropy (CE) exceeds most other methods but can be more overfitting to noisy samples, e.g. larger performance gap between best epoch and last epochs. It can be concluded that simple CE is better for training from an imbalanced noisy dataset and a suitable early stopping strategy can help train more robust models.

3) The Analysis of Observer Variability: In this section, we further discuss that how the trade-off between the experience of pathologists (i.e., the accuracy of annotation) and the number of available annotations affects the performance of the model. To quantify the observer variability between pathologists, we use two metrics. The first one is quadratic weighted kappa [63], which is widely used for measuring inter-rater agreement on imbalanced data. Here, we use it for quantifying the consistency between every single pathologist and majority-vote. The second one is Fleiss' kappa [64], which is a measure of association that generalizes Cohen's kappa for n ≥ 2 indistinguishable observers. The results are shown in Table V. We find there is a large observer variability between each pathologist with only 0.33 Fleiss' Kappa value. Besides, given an assumption that the pathologists with less experience are likely to mislabel some samples to a

TABLE V
COHEN'S AND FLEISS' KAPPA VALUE OF GLEASON 2019

Name	1	2	3	4	5	6	Fleiss' Kappa
Value	0.52	0.78	0.85	0.78	0.83	0.79	0.33

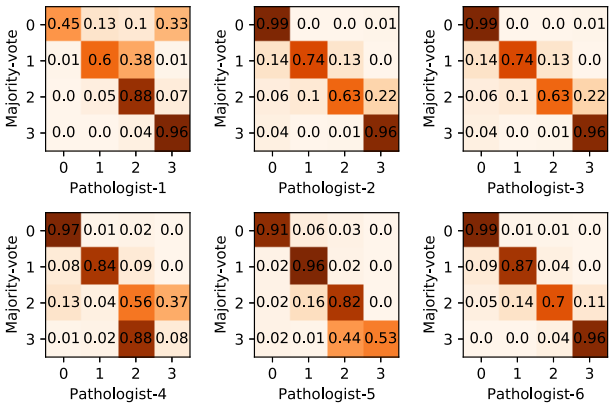


Fig. 8. The illustration of mislabeling between each category. The confusion matrix shows the different preferences of pathologists when giving annotations to the samples.

milder/severer grade, we draw a confusion matrix to show this kind of preference for each pathologist, as Fig. 8 shows. For example, Pathologist 3 would like to wrongly label the most Grade 5 samples to Grade 4, while he/she can give accurate annotations to almost all samples of the other three grades.

We consider the following training settings:

- **SP** denotes Single Pathologist. We use the annotations provided by one of the pathologists only to train models and calculate the average evaluation results of them. It is noted that the number of training samples for each pathologist is inconsistent.
- **SP + Ensemble** denotes that we do not calculate the average accuracy of SP and we ensemble the predictive results to give a joint diagnosis.
- **SP + PL**. It is not fair to train the model for every single pathologist with a different number of annotations. We leverage a simple but effective technique in semi-supervised learning named Pseudo-Labeling [65] to train the model using both labeled and unlabeled data from each pathologist. Similarly, we also perform ensemble predictions (**SP + PL + Ensembles**).
- **SP + SO**. To make use of those unlabeled samples, we assign the majority-vote label from the other five pathologists to the unlabeled data, which is similar to the Second Opinion in the clinical practice [53]. Similarly, we also perform ensemble predictions (**SP + SO + Ensembles**).

First, we present the detailed results trained from single pathologist (SP) annotations in Fig. 9, aiming to demonstrate a trade-off finding between three aspects: (1) the ability of a single pathologist to give a correct diagnosis, i.e., Kappa value in Fig. 9, which denotes the potential noise rate or annotation quality in individual SP; (2) the number of labeled samples for training; (3) the performance of the model, i.e.,

TABLE VI
THE OBSERVER-VARIABILITY ANALYSIS RESULTS ON GLEASON 2019

Gleason 2019	Epoch	Recall	Precision	F1 Score
SP	Best	63.53	68.10	59.03
	Last	61.72	65.37	57.74
SP + PL	Best	75.12	70.13	70.31
	Last	72.33	70.34	69.15
SP + SO	Best	76.05	78.34	77.17
	Last	73.00	75.22	74.09
Majority-vote	Best	81.44	82.48	81.88
	Last	80.13	79.94	79.98
Ours ($t_{UoD} = UoD_0$)	Best	**82.34**	84.03	82.50
	Last	**81.27**	**84.12**	**82.01**
Ours ($t_{UoD} = UoD_1$)	Best	82.33	**84.13**	**83.22**
	Last	79.01	79.73	79.37
SP + Ensemble	Best	69.36	84.17	72.46
	Last	67.01	70.32	65.01
SP + PL + Ensemble	Best	80.02	**85.43**	82.40
	Last	76.71	80.90	78.48
SP + SO + Ensemble	Best	82.43	81.33	81.88
	Last	80.96	80.80	80.87
Ours + Ensemble	Best	**84.07**	85.29	**84.68**
	Last	**82.85**	**82.72**	**82.78**

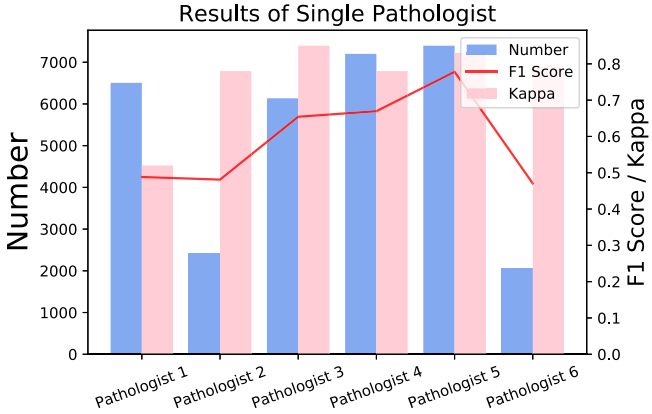


Fig. 9. The performance (F1 Score in red line, right scale) differs significantly with respect to pathologists' annotations (Kappa value in pink bars, right scale) and the number of training data (in blue bars, left scale).

F1 score value in Fig. 9. We can observe that although the amount of training data with annotations from pathologist 1 is three times as many as that of pathologist 6, they achieve similar performance. This may be due to the annotations from pathologist 6 having a higher kappa score, which means less label noise existed and higher annotation quality. A sub-conclusion can be summarized that the quality of annotations contributes more than the number of labeled samples for training.

Table VI gives the detailed quantification results of the comparative study. Since the Ensemble strategy owns more parameters (6 × parameters), which require more resources and longer inference time, for a fair comparative study, we compare methods with different parameters separately.

The semi-supervised technique Pseudo-Labeling [65] leverages the unlabeled samples and has made a significantly improved margin to the performance on SP (from 59.03% to 70.31% in terms of F1 score). The use of SO for unla-

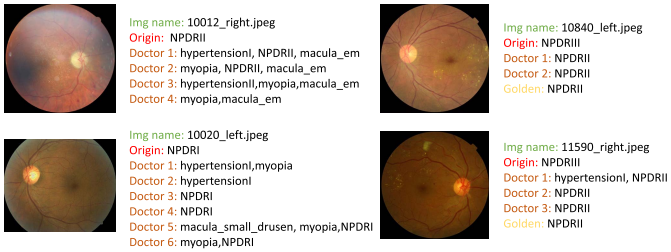


Fig. 10. Some samples of re-labeled datasets. For each sample, there are 2 - 6 ophthalmologists to relabel them. Furthermore, we select some samples with unbiased diagnosis and build a golden standard dataset for evaluation.

beled samples also obtains more promising improvement (from 59.03% to 77.17% in terms of F1 score) since the majority-vote of the other five pathologists can be more reliable. The baselines with "Ensemble" are able to simulate the clinical decision distribution. It is noticed that Ensemble Model (SP + PL + Ensemble and SP + SO + Ensemble, 82.40% and 81.88% on the best F1 Score) outperforms direct training from Majority-vote (81.88% on the best F1 Score) although the number of training samples is the same. We draw the conclusion that the network implicitly learns information from each pathologist during the training phase and has lower uncertainty on the ensemble predictions, where the uncertainty majority-vote is higher when constructing the ground truth for each sample before training. Similarly, we put two results with different selection thresholds here, and our proposed methods are able to reduce this uncertainty level for the single-target (majority-vote) label and thus improve the performance from 81.88% to 83.22% in terms of F1 Score when given $t_{UoD} = UoD_1$. Also, it is found that our methods outperform those ensemble-based strategies. We wonder that if the ensemble strategies can make the model more stable predictions from the models obtained in different epochs, e.g., the smaller performance gap between best and last epochs. To keep the consistency of parameters, six models are trained based on our proposed methods with different random seeds and then we make an ensemble of them. The performance has been further improved both in the best epoch (from 83.22% to 84.68% in terms of F1 score) and last epochs (from 79.37% to 82.78%).

D. Kaggle DR+

1) Data Statistics and Comparison: The original Kaggle Diabetic Retinopathy (DR) dataset [23] consists of 88,702 fundus photographs from 44,351 patients: one photograph per eye. It aims to be used for training a model that is able to classify 5 common levels in DR, which includes Normal, mild Non-proliferative DR (NPDRI), moderate Non-proliferative DR (NPDRII), severe Non-proliferative DR (NPDRIII) and proliferative DR (PDR). However, it is estimated that there are about 30% - 40 % noisy labels in the originally released dataset in two aspects. First, the intra-observer variability among all DR categories, e.g., mild DR is wrongly annotated as moderate DR. Second, only DR grading labels are considered, however, according to to [66], there are more than 10 retinal diseases such as glaucoma, drusen, *et al.* have

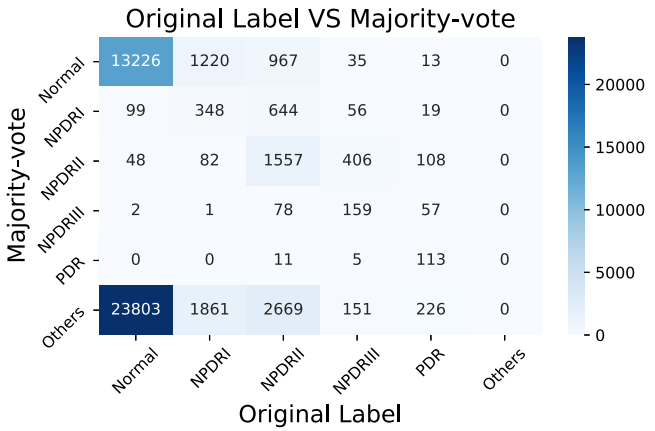


Fig. 11. The illustration of mislabeling in original Kaggle DR dataset compared to Majority-vote aggregated from our re-engineered annotations.

TABLE VII
THE DATA STATISTICS ON SELECTED SAMPLES OF KAGGLE DR+. 0: NORMAL; 1: NPDRI; 2: NPDRII; 3: NPDRIII; 4: PDR; 5: OTHERS

	0	1	2	3	4	5
Original	8001	2978	6866	1142	888	-
Majority-vote	11739	1068	3359	429	190	3036
Golden Standard	500	200	200	129	123	-

TABLE VIII
THE COMPARATIVE RESULTS OF DR CLASSIFICATION

	Kaggle DR+		
	Recall	Precision	F1 Score
Original + CE	46.42 (±2.82)	45.43 (±2.16)	45.42 (±1.76)
MV + CE	48.51 (±2.05)	61.30 (±2.37)	51.05 (±0.86)
MixUp	46.22 (±1.03)	45.01 (±1.67)	45.61 (±1.06)
Label Smooth	49.02 (±0.94)	60.00 (±0.53)	53.95 (±0.57)
CurriculumNet	**50.02** (±3.42)	56.77 (±2.88)	53.18 (±1.62)
Co-teaching	42.17 (±4.20)	50.84 (±1.27)	46.10 (±1.91)
DivideMix	28.62 (±0.26)	26.03 (±0.11)	27.26 (±0.08)
O2U-Net	50.01 (±2.19)	48.65 (±2.64)	49.32 (±1.54)
Ours	49.63 (±1.31)	**64.02** (±3.15)	**55.91** (±0.68)

*All comparative methods are performed on MV.

never been labeled for this dataset by the original annotators. Therefore, we decided to re-engineer this dataset and release the multi-label Kaggle DR+ dataset with a golden standard dataset for an unbiased evaluation, which was relabeled by 10+ ophthalmologists covering 17 different retinal diseases commonly examined during screening.[3] Fig. 10 shows some samples and corresponding annotations of our re-engineered dataset. Some samples may contain several kinds of diseases which exhibit a multi-label distribution. A confusion matrix is presented in Fig. 11. An interesting finding is that there are more than 20,000 samples that contain some other retinal diseases other than DR, but they were wrongly annotated as Normal in the original annotations. Our re-engineered database not only labels some other common or rare retinal-disease categories, but also greatly reduces the noise rate for DR grading task with more second opinions and discussion for those wrongly-diagnosed samples from multiple experienced ophthalmologists.

[3]Please refer to our Appendix for a detailed analysis on original and our released dataset.

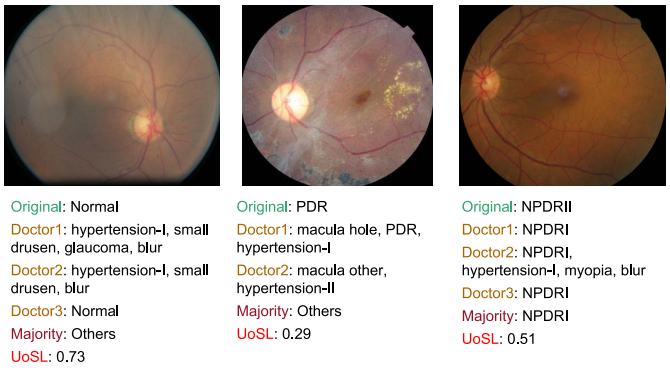


Fig. 12. Examples of applying *UoSL* for outliers detection. **Orinigal** defines the DR label from the original Kaggle DR dataset. **Doctor x** denotes the new label from our Kaggle+ dataset.

TABLE IX
ABLATION STUDY RESULTS

Component	ISIC
Cross-Entropy	78.27
Focal Loss	80.09
Dual-U + FL	78.77
Weighted CE	78.92
Ours	**81.01**

η	Gleason	Kaggle
0 (DUP)	82.53	**55.91**
1/2	80.17	55.02
1	**83.22**	51.09
2	78.26	53.12

TABLE X
THE STUDY OF THE SENSITIVITY OF t_{UoD}

UoD	0.0	0.28	0.44	0.61	0.67
Number of Samples	4401	2339	953	68	5
F1 Score (Best)	82.50	**83.22**	82.01	81.88	81.89
F1 Score (Last)	**82.01**	79.37	79.22	79.02	79.99

To have a better comparison with the original annotations [23], only the samples whose annotated ground truth is included in DR grading are used. The statistics of selected samples are shown in Table VII. We split the train data into the training set: validation set = 8: 2 for early stop and select the checkpoint. Then, an unbiased golden standard dataset (statistics in Table VII) is used for test and performance reports.

2) Quantitative Analysis: Table VIII shows the overall comparative results. It should be noted that direct training from the original-labeled dataset shows a catastrophic performance due to highly imbalanced data distribution. For a fair comparison, we apply an on-the-fly up-sampling strategy for those minority classes. We train two baselines here: train from the original-labeled (denoted by Original + CE) and majority-vote results from our re-engineered dataset respectively (denoted by MV + CE). From Table VIII, we find that the performance has been significantly improved using our newly annotated datasets, e.g., from 45.42% to 51.02% with nearly 6% improvement in terms of F1 score. Then we perform the same benchmark test in ISIC 2019 and Gleason 2019 datasets on our re-engineered dataset (MV) to investigate if there is any further performance gain. We failed to obtain satisfactory results from DivideMix. A phenomenon is that almost all samples are diagnosed into two classes Normal and NPDRIII, resulting in catastrophic performance. This may be due to the scenario that Normal owns most samples and NPDRIII is more easily to be distinguished with other categories. Overall, our proposed methods obtain the best results with 64.02% and 55.91% in terms of precision and F1 score.

3) Out-of-Category Labels Analysis: We also give qualitative results in Fig. 12 to show the ability of our proposed *UoSL* for outliers (noisy and out-of-category label) detection. The first image from the right is wrongly labeled as Normal. The obtained *UoSL* score is a relatively high value of 0.73, which demonstrates our uncertainty-based estimation is sensitive to false-negative samples. We also give two bad cases from our method, which explains the limitations of the dual-uncertainty estimation. The *UoSL* score is 0.29 on the middle image from Fig. 12. The reason behind this may due to the high similarity between PDR (proliferative diabetic retinopathy) and majority-vote results (others, macula and hypertension in this example) in the feature space. Also, we find out that *UoSL* is less generalized to mislabelling between two neighbor classes with small appearance differences from the rightmost example (NPDRI and NPDRII, and $UoSL = 0.51$ in the third image).

E. Ablation Study

1) Components Analysis: To test how each factor makes our model performant, we conduct further ablation studies on various method components. The results of the ablation study indicate that both Focal Loss and Weighted CE can benefit the model compared to CE loss, but Focal Loss with Dual-uncertainty obtain marginal improvements due to the mistaken elimination of some hard samples but with clean labels.

2) Hyper-Parameters Selection: Furthermore, we evaluate the newly-designed iUOD on Gleason 2019 and Kaggle+, which contain annotations from multiple doctors, and the results are shown in Table IX. When $\eta = 0$, the original DUP is applied. We tried several index values for the hyper-parameter η to amplify or reduce the effect of a factor that denotes how much attention should be paid to the uncertainty brought by the number of annotations. When $\eta = 1$, the improved DUP achieves the best results evaluated on Gleason 2019. However, there are no more improvements on selected Kaggle DR+. The selected Kaggle DR+ dataset in our manuscript only contains sole disease label and the inconsistency between ophthalmologists are less compared to Gleason 2019. We can draw a conclusion that improved DUP performs better on the dataset with low quality of annotations and high inconsistency in the number of doctors to give annotations.

Now we discuss the sensitivity of threshold t_{UoD} for high UoD samples rejection. As we presented in Table VI, larger t_{UoD} leads to a smaller loss in accuracy between best and last epochs since most samples kept reaching full agreement by all pathologists. And larger t_{UoD} achieved better results in the best epoch but then the performance dropped more than 3%. Some samples can obtain the same UoD value calculated by Eq. 3 so the values are discontinuous and can be grouped. We give statistics for Gleason 2019 to show the number of

TABLE XI
THE STUDY OF DIFFERENT UNCERTAINTY ESTIMTAION STRATEGIES

Methods	Kaggle DR+
Plain Model	51.05 ($\pm$0.86)
MC-Dropout [17] (T = 4) + W.	55.91 ($\pm$0.25)
MC-Dropout (T = 16) + W.	**56.00** ($\pm$0.68)
MC-Dropout (T = 128) + W.	54.77 ($\pm$1.01)
MC-Dropout (T = 4) + S.	52.33 ($\pm$2.25)
Deep Ensemble [47] (M = 5)	55.80 ($\pm$0.17)
Augmentation [52] (T = 128)	52.17 ($\pm$1.44)

samples for each UoD value in Table X. We also present the performance of our proposed methods under the selection of different t_{UoD} values. From the results, we can see that our proposed methods are insensitive to t_UoD value selection and all settings achieve competitive performance. And a smaller t_UoD is suggested for training a more robust network.

3) Different Uncertainty Estimation Strategies: In this work, we estimate the *uncertainty of single label* of samples by calculating from Eq. (4) using the MC-Dropout. Here, we evaluate the different uncertainty estimation strategies related to our works such as Deep Ensemble [47] and Augmentation [52], as shown in Table XI. Augmentation-based method [52] is evaluated on Diabetic Retinopathy detection task and the range of augmentations is also provided. Hence, we validate the comparative methods on Kaggle DR+ dataset.

First, we investigate the affect of the times of stochastic forward pass (denoted by T) and the selection of augmentation. W. denotes weak augmentations such as random rotation and flip, and S. denotes strong augmentations, which are achieved by RandAugment [67]. From Table XI, it is found that slight larger T value (T = 16) can bring marginal improvements (from 55.91% to 56.00% in terms of F1 score) and too large T values (T = 128) not only requires more inference time but also do harm to the performance of the model (from 56.00% to 54.77% in terms of F1 score).

Then, stronger augmentations are adopted but there is no any improvements. A possible assumption is that stronger augmentations perturbing the raw input or intermediate features of the model may make the model difficult to converge, resulting in the difficulty to make a distinction between clean samples from noise samples in the warm-up phase. We also evaluate the Deep Ensemble method with training five models as suggested by [47] and obtain quite close results. However, it requires more parameters and training time for training ensemble models. Augmentations-based method tends to estimate the uncertainty regardless of the architectural design choices or regularization methods. We use the same augmentation settings and hyper-parameters suggested by [52]. However, only 1.12% improvement is gained. From this study, we can draw a conclusion that MC-Dropout is a simple but effective method for uncertainty estimation out of some existing scenarios such as robust learning.

V. CONCLUSION

In this paper, we defined and explored two unique types of label noise: disagreement and single-target label noise, existed in the medical image classification task and proposed a Dual-uncertainty estimation framework which can significantly improve model's tolerance to the label noise. We conducted extensive experiments on three different disease classification tasks (dermatology, ophthalmology and pathology) and showed the superiority of our proposed framework to handle the uncertainty of annotations in medical images. However, our work has some potential limitations since some typical factors in the medical application such as long-tailed and open-label recognition are still needed to be taken into consideration in the presence of label noise. We will release an annotation-database which relabeled a public fundus dataset with annotations covering 17 retinal diseases from more than 10 ophthalmologists. We hope this dataset will help address the challenges of label noise in the medical image analysis.

REFERENCES

[1] Amazon. (2012). *Amazon Mechanical Turk*. [Online]. Available: https://www.mturk.com/

[2] S. Guo *et al.*, "Curriculumnet: Weakly supervised learning from large-scale web images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 135–150.

[3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[4] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 590–597.

[5] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," 2020, *arXiv:2007.08199*.

[6] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101759.

[7] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2304–2313.

[8] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[9] J. Li, R. Socher, and S. C. H. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," 2020, *arXiv:2002.07394*.

[10] J. Shu *et al.*, "Meta-weight-net: Learning an explicit mapping for sample weighting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1919–1930.

[11] Y. Liu, "Understanding instance-level label noise: Disparate impacts and treatments," 2021, *arXiv:2102.05336*.

[12] G. Dawson and R. Polikar, "Rethinking noisy label models: Labeler-dependent noise with adversarial awareness," 2021, *arXiv:2105.14083*.

[13] Y. Zhang and M. Sugiyama, "Approximating instance-dependent noise via instance-confidence embedding," 2021, *arXiv:2103.13569*.

[14] Y. Zhang, S. Zheng, P. Wu, M. Goswami, and C. Chen, "Learning with feature-dependent label noise: A progressive approach," 2021, *arXiv:2103.07756*.

[15] H. Cheng, Z. Zhu, X. Li, Y. Gong, X. Sun, and Y. Liu, "Learning with instance-dependent label noise: A sample sieve approach," 2020, *arXiv:2010.02347*.

[16] C. Xue, Q. Dou, X. Shi, H. Chen, and P.-A. Heng, "Robust learning at noisy labeled medical images: Applied to skin lesion classification," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1280–1283.

[17] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[18] J. Huang, L. Qu, R. Jia, and B. Zhao, "O2U-Net: A simple noisy label detection approach for deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3326–3334.

[19] B. Han *et al.*, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8527–8537.

[20] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," 2018, *arXiv:1803.09050*.

[21] I. Challenge. (2049). *Isic 2019*. [Online]. Available: https://challenge2019.isic-archive.com/

[22] G. Challenge. (2019). *Gleason*. [Online]. Available: https://gleason2019.grand-challenge.org/

[23] California Healthcare Foundation and EyePACS. (2015). *Diabetic Retinopathy Detection*. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection/

[24] X. Yu, T. Liu, M. Gong, K. Zhang, K. Batmanghelich, and D. Tao, "Transfer learning with label noise," 2017, *arXiv:1707.09724*.

[25] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1765–1773.

[26] J. Speth and E. M. Hand, "Automated label noise identification for facial attribute recognition," in *Proc. CVPR Workshops*, Jan. 2019, pp. 25–28.

[27] F. Yang *et al.*, "Asymmetric co-teaching for unsupervised cross-domain person re-identification," in *Proc. AAAI*, 2020, pp. 12597–12604.

[28] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5049–5059.

[29] Y. Yao *et al.*, "Dual T: Reducing estimation error for transition matrix in label-noise learning," 2020, *arXiv:2006.07805*.

[30] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," 2017, *arXiv:1712.09482*.

[31] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8778–8788.

[32] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 322–330.

[33] Y. Lyu and I. W. Tsang, "Curriculum loss: Robust learning and generalization against label corruption," 2019, *arXiv:1905.10045*.

[34] X. Wang, Y. Hua, E. Kodirov, and N. M. Robertson, "IMAE for noise-robust learning: Mean absolute error does not treat examples equally and gradient Magnitude's variance matters," 2019, *arXiv:1903.12141*.

[35] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," 2014, *arXiv:1406.2080*.

[36] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1237–1246.

[37] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[38] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," 2017, *arXiv:1701.06548*.

[39] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels," 2019, *arXiv:1911.06475*.

[40] F. C. Ghesu *et al.*, "Quantifying and leveraging classification uncertainty for chest radiograph assessment," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 676–684.

[41] Y. Dgani, H. Greenspan, and J. Goldberger, "Training a neural network based on unreliable human annotation of medical images," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 39–42.

[42] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[43] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[44] K. Krishna and M. N. Murty, "Genetic K-means algorithm," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 29, no. 3, pp. 433–439, Jun. 1999.

[45] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Loop: Local outlier probabilities," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 1649–1652.

[46] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon *et al.*, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 5574–5584.

[47] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6405–6416.

[48] J.-X. Zhong and H. Zhang, "Uncertainty-aware INVASE: Enhanced breast cancer diagnosis feature selection," 2021, *arXiv:2105.02693*.

[49] L. Wang *et al.*, "Medical matting: A new perspective on medical segmentation with uncertainty," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*. Cham, Switzerland: Springer, 2021, pp. 573–583.

[50] F. C. Ghesu *et al.*, "Quantifying and leveraging predictive uncertainty for medical image assessment," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101855.

[51] A. JSANG, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Cham, Switzerland: Springer, 2018.

[52] M. S. Ayhan, L. Kühlewein, G. Aliyeva, W. Inhoffen, F. Ziemssen, and P. Berens, "Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101724.

[53] M. Raghu *et al.*, "Direct uncertainty prediction for medical second opinions," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5281–5290.

[54] F. L. Wauthier and M. I. Jordan, "Bayesian bias mitigation for crowdsourcing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1800–1808.

[55] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J. Jpn. Soc. Artif. Intell.*, vol. 14, nos. 771–780, p. 1612, Sep. 1999.

[56] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

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

[58] C. Tan, J. Xia, L. Wu, and S. Z. Li, "Co-learning: Learning from noisy labels with self-supervision," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1405–1413.

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

[60] E. Malach and S. Shalev-Shwartz, "Decoupling 'when to update' from 'how to update,"' in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 961–971.

[61] E. Arvaniti *et al.*, "Automated Gleason grading of prostate cancer tissue microarrays via deep learning," *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, Dec. 2018.

[62] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, "Learning from noisy labels by regularized estimation of annotator confusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11244–11253.

[63] J. L. Fleiss and J. Cohen, "The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability," *Educ. Psychol. Meas.*, vol. 33, no. 3, pp. 613–619, 1973.

[64] J. L. Fleiss, J. Cohen, and B. S. Everitt, "Large sample standard errors of Kappa and weighted Kappa," *Psychol. Bull.*, vol. 72, no. 5, p. 323, 1969.

[65] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Represent. Learn., (ICML)*, vol. 3, no. 2, 2013, pp. 1–6.

[66] X. Wang, L. Ju, X. Zhao, and Z. Ge, "Retinal abnormalities recognition using regional multitask learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 30–38.

[67] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 702–703.