Unsupervised Confidence Approximation: Trustworthy Learning from Noisy Labelled Data

Navid Rabbani, Adrien Bartoli
EnCoV, Institut Pascal, Université Clermont Auvergne, CNRS, France
DRCI, CHU Clermont-Ferrand, France
navid_rabbani@yahoo.com, adrien.bartoli@gmail.com

Abstract

Training neural networks with noisy labels presents a challenge due to inherent errors in label annotations. Concurrently, selectively predicting outputs from neural networks involves identifying confidently predicted results. These challenges are particularly important in the medical domain, as they often occur jointly. Existing techniques address either the training of models with noisy labels or the task of selective prediction in isolation, often neglecting their intrinsic interdependence. We establish a relationship between these challenges and propose a novel framework called Unsupervised Confidence Approximation (UCA) to address them simultaneously. UCA facilitates the concurrent training of neural networks for a main task such as image segmentation and classification while also predicting confidence levels. This is all done while accommodating datasets containing noisy labels. Remarkably, UCA operates autonomously, eliminating the need for labelled confidence information and qualifying as an unsupervised solution. Furthermore, UCA is versatile, integrating with diverse network architectures. Our evaluation of UCA's efficacy covers the general CIFAR-10N dataset as well as the medical image datasets CheXpert and Gleason-2019. In our experiments, incorporating UCA into existing networks enhances performance in both aspects of noisy label training and selective prediction. Moreover, networks equipped with UCA demonstrate comparable performance to state-of-the-art methods for noisy label training when operating in the conventional full coverage mode. By design, these UCA-equipped networks incorporate a risk-management mechanism, as evidenced by flawless risk-coverage curves. Additionally, UCA-equipped networks outperform existing selective prediction techniques, leading to substantial performance improvements and reinforcing its utility and impact within the context of trustworthy medical deep learning.

1. Introduction

Deep learning has been very successful in many domains. Effectively training a deep neural network (DNN) generally requires a large amount of carefully labelled data. Medical image datasets, like any real-world dataset, may include noise in the labels. Noisy labels arise when the annotators give a wrong label to the image, either as a random mistake or owing to the ambiguity of the image, leading to inconclusiveness of the annotation task. The rate of label noise can be substantial when the annotators are non-expert humans, automated systems or when the diagnostic uncertainty is intrinsically high, see figure 1. While recklessly training a DNN with noisy labels severely degrades performance, specific robust training methods exist [4, 14, 25]. Aside, potential errors are inherent and inevitable in the outputs of any given DNN. To manage the risk caused by these errors, a selective predictor abstains from making predictions when it detects high uncertainty in the DNN predictions. A reliable uncertainty or confidence measure is at the core of selective prediction methods [6]. We claim that the engineering of clinical and healthcare systems would strongly benefit the concurrent features of 1) training from noisy labels and 2) making selective prediction. Both features are well-known but have not been realised concurrently. We show that they are interdependent and solvable in an integrated framework.

We propose Unsupervised Confidence Approximation (UCA), a method to train a DNN for its main task and for confidence prediction, from noisy datasets without confidence labels. UCA gives, for the first time, concurrent solutions for the two mentioned features. It is a major contribution as existing methods solve one of these two problems but fail when they are concurrent. UCA adds a confidence prediction head to the main DNN, whose role is to approximate the confidence for the main task. It is generic, in the sense that it can be used with any neural architecture. The proposed UCA loss makes it possible to train the main network and the UCA head concurrently. It does not require

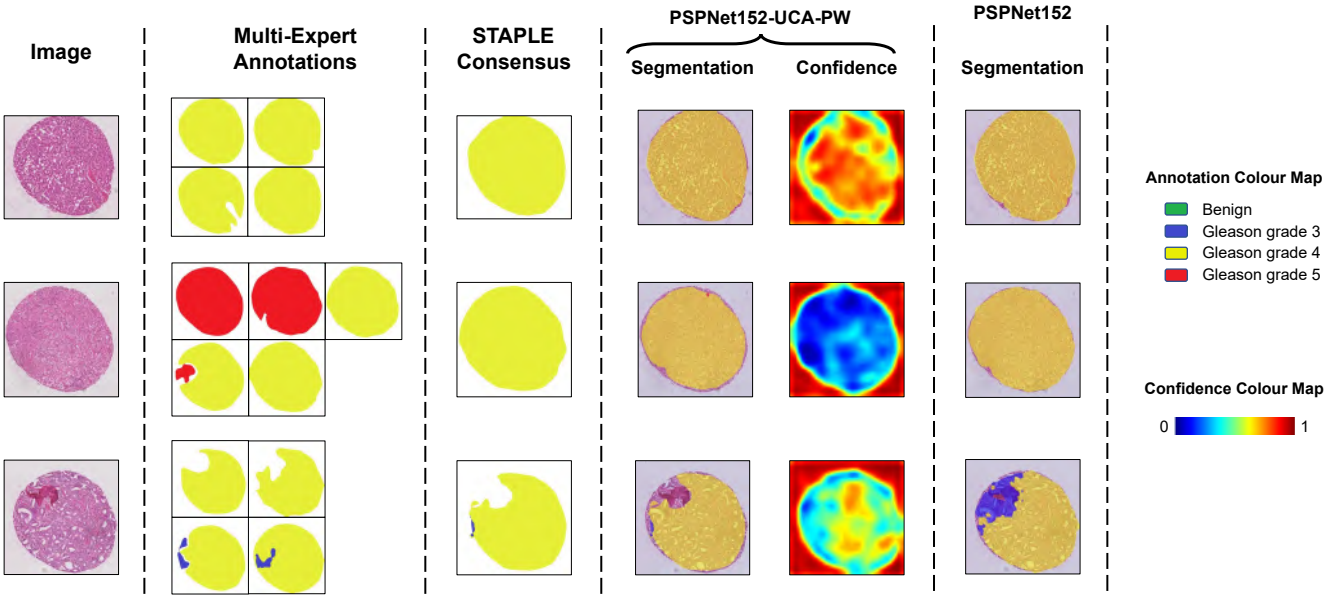

Figure 1. Test samples from the Gleason-2019 dataset [20] for cancer grading. PSPNet152 [22] trained from the STAPLE consensus is to date the best performing method (Gleason-2019 challenge). PSPNet152-UCA-PW is PSPNet152 with the proposed UCA trained from STAPLE (very similar results are obtained when trained from the multi-expert annotations). First row: test case with multi-expert agreement, PSPNet152 and PSPNet152-UCA-PW give similar results, PSPNet152-UCA-PW has high confidence. Second row: test case with strong multi-expert disagreement, PSPNet152 and PSPNet152-UCA-PW give similar results, however PSPNet152-UCA-PW indicates low confidence. Third row: test case with mild multi-expert disagreement, PSPNet152 fails to predict the STAPLE consensus, while PSPNet152-UCA-PW does succeed and also indicates low confidence in the disagreement area.

the confidence labels and is thus unsupervised in this respect. We show experimental results on the CIFAR-10N, CheXpert and Gleason-2019 datasets, where UCA shows a strong performance gain over existing selective prediction methods and is on par with the state-of-the-art in noisy label training when used in full coverage mode.

2. Related Work

We review related work in training with noisy labels and predictive uncertainty estimation. *There is no concurrent solution to these two problems.*

Training with noisy labels has been a research focus in machine learning for a decade, see the surveys [4, 14, 25]. The first approach weights the contribution of samples to the loss. A straightforward method is the confidence-scored instance-dependent noise (CSIDN) weight, which however requires the confidence labels [2]. The weights can also be found during training by constrained optimisation [15]. The second approach iteratively selects samples that are likely to be noise-free [11, 33, 17, 30, 3]. These methods use two networks selecting the clean data samples for each other to mitigate the confirmation bias [26]. The third approach uses a noise-resistant loss. The mean absolute error (MAE) was shown to be more robust to noise than cross-entropy (CE) [8]. A generalised cross-entropy (GCE) loss was proposed that combines the advantages of MAE and CE [34]. A loss exploiting class switching probabilities was used [12, 21, 9]. However, the probabilities are assumed class-dependent and feature-independent, which is not realistic in many cases. The fourth approach uses early training stopping, assuming that the clean data have more impact in the early training steps whilst the noisy samples start corrupting in the later training steps [1].

Predictive uncertainty estimation has recently gained an increased interest, see the survey [6]. The first approach uses the ultimate softmax value of a DNN to predict confidence. A DNN is deemed calibrated when this prediction is valid. A straightforward method is to directly train a calibrated DNN, which however requires the confidence labels [23]. Calibration can also be done by post-processing from a clean validation dataset [10]. The mixup method regularises the DNN to favour a simple linear behaviour across the training examples, resulting in an improved calibration [27]. The second approach uses a stochastic model. The parameters of Bayesian DNNs are explicitly modelled as random variables, leading to stochastic predictions, from which the confidence can be estimated. Bayesian inference in DNNs is however intractable. This was addressed by Deep Ensembles [16, 18, 24] and Monte Carlo Dropout (MC-Dropout) [5]. Both techniques are highly resource-

intensive and require several forward passes.

3. Method

We predict confidence as a measure of prediction uncertainty [19]. We first describe the 'global UCA', which implements a per sample confidence.

3.1. Noisy Labels and Confidence Score Approximation

We formulate the problem of learning with noisy labels following [32]. Let D be the distribution of the noise-free samples, modelled as a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the input space and $\mathcal{Y} = \{1, 2, \ldots, C\}$ is the target set. In contrast, the samples of a noisy dataset $(X, \bar{Y}) \in \mathcal{X} \times \mathcal{Y}$ are drawn from the noisy distribution $\bar{D}$. A relationship between the two distributions is given by the clean probability of the sample $(x, \bar{y})$:

$$r(x, \bar{y}) = P(Y = \bar{y} \,|\, \bar{Y} = \bar{y}, X = x). \tag{1}$$

We assume the label noise is structered, image-dependent and label-independent [28, 35], which holds very well for human annotations [31]. It means the label noise statistics are highly correlated to the visual features, hence images with similar features have similar noise statistics. Concretely, the human-annotated label noise is due in large part to the image being ambiguous, low quality, inconclusive or confusing, and in small part to random mistakes. The clean probability (1) thus becomes independent of $\bar{y}$; we propose to model it by a DNN $\bar{r}(x; \phi) \approx P(Y = \bar{Y} \,|\, X = x_i)$ with parameters ϕ. Assuming an effective training of $\bar{r}(x; \phi)$, it provides the average clean probability distribution. As the reliability of the DNN's output for the main task is compromised in regions where training samples have a low clean probability, we can consider $\bar{r}(x; \phi)$ as an approximation for the confidence score.

3.2. Unsupervised Confidence Approximation Loss

We model the DNN for the main task as $y = f(x; \theta)$ with parameters θ. We denote the loss for the main task and the i-th training sample as $L(x_i, \bar{y}_i; \theta) \geq 0$, for $i = 1, \ldots, N$. For per-sample weights $\{w_i\}$, the DNN parameters θ^* are classically found by solving:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} w_i \, L(x_i, \bar{y}_i; \theta). \tag{2}$$

We propose to use $w_i = \alpha \, \bar{r}(x_i; \phi)$ as sample weights so as to downweight the samples prone to noise. Considering that:

$$\begin{aligned} \sum_{i=1}^{N} \bar{r}(x_i; \phi) &\approx N \mathbb{E}_X \bar{r}(x_i; \phi) \\ &= N \sum_i P(Y = \bar{Y} | X = x_i) P(X = x_i) \\ &= N P(Y = \bar{Y}), \end{aligned} \tag{3}$$

and normalising the weights to $\sum_{i=1}^{N} w_i = 1$, we have:

$$\alpha = \frac{1}{\sum_{i=1}^{N} \bar{r}(x_i; \phi)} \approx \frac{1}{N P(Y = \bar{Y})} = \frac{1}{NA}, \tag{4}$$

where A is the total labelling accuracy of the training data, considered as a hyperparameter if not known a priori. A naive approach is then to train θ, ϕ by solving:

$$\theta^*, \phi^* = \underset{\theta, \phi}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{\bar{r}(x_i; \phi)}{NA} L(x_i, \bar{y}_i; \theta). \tag{5}$$

This has trivial spurious solutions, such as weighting all samples with zero except one. We thus add a regularisation term $D(w, u)$ penalising divergence of the discrete weight distribution w, with $w_i = \frac{1}{NA} \bar{r}(x_i; \phi)$, to a prior weight distribution u. We use the non-informative uniform distribution $u_i = \frac{1}{N}$ by default; any other distribution constructed for instance from inter-expert variability may be used instead. We arrive at the proposed UCA loss for training in the presence of noisy data with hyperparameter $\beta > 0$ as:

$$\begin{aligned} \theta^*, \phi^* = \quad & \underset{\theta, \phi}{\operatorname{argmin}} \sum_{i=1}^{N} \bar{r}(x_i; \phi) L(x_i, \bar{y}_i; \theta) + \beta D(w, u) \\ \text{subject to:} \quad & \bar{r}(x; \phi) > 0, \\ & \sum_{i=1}^{N} \bar{r}(x_i; \theta) = NA \end{aligned} \tag{6}$$

The UCA loss is the core of our approach: it allows one to train $f(x; \theta)$ and $\bar{r}(x; \phi)$ end-to-end without needing confidence labels while handling noisy data.

3.3. Unsupervised Confidence Approximation Architecture

We name the DNN $\bar{r}(x; \phi)$ as UCA head, as it learns the instance-based confidence without requiring its label. The UCA head is connected to the features of the main network $f(x; \theta)$, as shown in figure 2. We present two versions of the UCA head. The global UCA head implements the method as described thus far, with a per-sample weight $\bar{r}(x; \phi)$. It has a global averaging layer and K fully connected hidden layers with ReLU activation. We use a sigmoid as last activation, enforcing $\bar{r}(x; \phi) > 0$. We use a special batch normalisation layer in the output, enforcing $\sum_i \bar{r}(x_i; \theta) = NA$ in each training batch. The pixelwise UCA head is described in section 3.5.

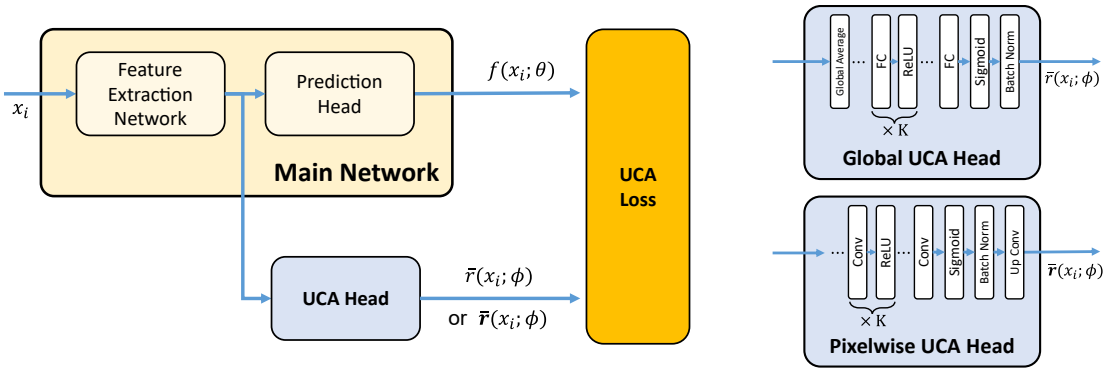

Figure 2. Unsupervised Confidence Approximation (UCA) architecture.

3.4. Confidence-selective Prediction

Following the concept of selective classifiers [7], we define the confidence-selective predictor $\tilde{f}$ as a pair of functions (f, r) where $f : \mathcal{X} \to \mathcal{Y}$ is the prediction function and $r : \mathcal{X} \to [0, 1]$ is the confidence function. Defining $t \in [0, 1]$ as the minimum acceptable confidence, the confidence-selective predictor is:

$$\tilde{f}(x) = \begin{cases} f(x), & r(x) \geq t \\ \text{reject} & \text{otherwise.} \end{cases} \tag{7}$$

Concretely, $r(x)$ is obtained by UCA, softmax confidence or any other confidence measure. By varying t, one controls the coverage and consequently the risk. Coverage is the probability mass of the non-rejected region in $\mathcal{X}$ and risk is the expected value of $l(f(x), y)$ on the same region, where $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a given evaluation loss function. For classification, we use the classification error and for segmentation, we use Jaccard dissimilarity. A risk-coverage (RC) curve is a plot of prediction risk and coverage for a varying t. The RC curve can be used to choose a balancing point with acceptable trade-off between risk and coverage. We use area under RC curve (AURC) as a performance metric of selective predictors.

3.5. Pixelwise UCA

The above described UCA, which we name global UCA, estimates a single confidence per sample. This is very restricted for complex images and pixelwise tasks such as segmentation, for which one may be interested in accessing the local confidence of the DNN prediction, as shown in figure 1. We propose an extension named pixelwise UCA, which predicts a per-pixel confidence map $\bar{\mathbf{r}}_q(x_i;\phi)$ per sample, where $q \in \mathcal{I}$ is the pixel coordinates within the set of image pixel coordinates $\mathcal{I}$. We write the training loss as $\mathbf{L}_q(x_i, \bar{y}_i; \theta)$ for training sample x_i at pixel q. Defining the number of pixels as $M = \text{card}(\mathcal{I})$, we set the weights as $\mathbf{w}_{i,q} = \frac{1}{MNA}\bar{\mathbf{r}}_q(x_i;\phi)$ and the uniform prior distribution as $\mathbf{u}_{i,q} = \frac{1}{MN}$. We arrive at the proposed pixelwise UCA loss as:

$$\begin{aligned} \theta^*, \phi^* = \quad & \underset{\theta,\phi}{\text{argmin}} \sum_{i=1}^{N} \sum_{q \in \mathcal{I}} \bar{\mathbf{r}}_q(x_i;\phi)\, \mathbf{L}_q(x_i, \bar{y}_i; \theta) + \beta\, D(\mathbf{w}, \mathbf{u}) \\ \text{subject to:} \quad & \bar{\mathbf{r}}_q(x;\phi) > 0, \\ & \sum_{i=1}^{N} \sum_{q \in \mathcal{I}} \bar{\mathbf{r}}_q(x_i;\theta) \end{aligned} \tag{8}$$

The confidence is modelled by the pixelwise UCA head shown in figure 2, which is similar to the global UCA head without the global averaging layer and with convolutional hidden layers instead of fully connected ones. Pixelwise UCA allows pixelwise selective prediction. Concretely, the selective predictor can reject the predicted class for low confidence pixels. The Jaccard index is then computed on the selected pixel set $\mathcal{I}_c = \{q \in \mathcal{I} \,|\, \bar{\mathbf{r}}_q(x;\phi) \geq t\}$.

4. Experimental Results

4.1. Evaluation metrics

We use standard metrics. We evaluate the ability to cope with noisy labels using the Full Coverage Accuracy (FC-Acc), Full Coverage Area Under Curve (FC-AUC) and Full Coverage Jaccard index (FC-Jac), for classification and segmentation respectively. Full Coverage metrics are computed averaging over the complete test dataset. We evaluate selective prediction using the RC curve and AURC. An effective method must both cope with noisy labels and perform well in selective prediction.

4.2. Image Classification on CIFAR-10N

We use CIFAR-10N [31]. This dataset uses the same images as CIFAR-10 but the training dataset labels are substituted by human-annotated noisy labels. The test dataset labels are kept unchanged. We use ResNet34 trained with CE as baseline, named ResNet34-CE. We connect the global UCA head with $K = 1$ and 128 neurons to the output of

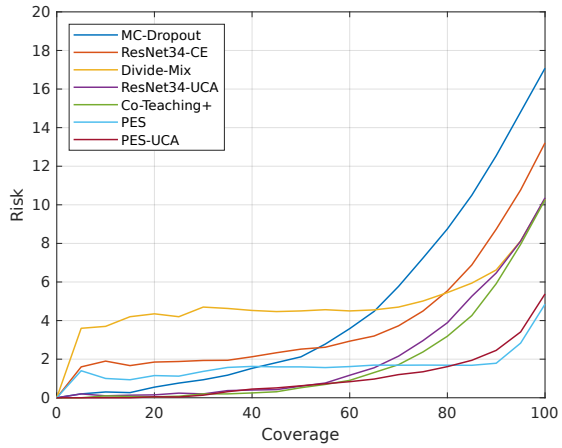

Figure 3. RC curves on CIFAR-10N.

Table 1. FC-Acc and AURC for image classification on CIFAR-10N.

Method	**FC-Acc↑**	**AURC↓**
MC-Dropout [5]	82.92%	4.43%
Divide-Mix [17]	89.64%	4.87%
Co-Teaching+ [33]	89.83%	1.75%
ResNet34-CE (Baseline)	86.79%	3.76%
ResNet34-UCA (ours)	89.64%	2.01%
PES (semi) [1]	95.12%	1.60%
PES-UCA (ours)	94.62%	0.96%

layer 4. We train using equation (6) and CE as main task loss, with fixed hyperparameters $A = 0.5$ and $\beta = 5$ forming method ResNet34-UCA. We also combined PES [1] and UCA, forming method PES-UCA. We trained in three steps: the main network using PES, then the UCA head and finally the complete DNN, both using equation (6). We compare UCA-equipped DNNs with existing methods, all trained on the Random1 subset of CIFAR-10N, whose noise rate is 17.23%. The results are in table 1 and figure 3. Comparing FC-Acc values between ResNet34-CE and ResNet34-UCA shows that UCA successfully downweights the impact of noisy samples. The performance of ResNet34-UCA is substantially better than ResNet34-CE and on par with PES [1], Co-Teaching+ [33] and Divide-Mix [17], which are solid methods in noisy label training. We also observe that MC-Dropout [5], representing uncertainty quantification methods, does not cope with noisy labels. The RC curves and AURCs show that ResNet34-CE performs poorly, but that ResNet34-UCA brings a significant boost. While PES has a satisfactory AURC, its RC curve is mostly flat, making it nearly impossible to trade off coverage for gaining accuracy. In contrast, PES-UCA shows the best RC performance. The AURC is considerably decreased compared to PES and the RC curve gives better control on the risk-coverage trade-off.

In figure 4, we present a visual analysis of the classification outcomes achieved by the ResNet34-UCA model on the CIFAR-10N dataset. The visualised samples are selected from the lowest 10% and highest 10% of estimated confidences. We observe that UCA assigns higher confidences to the unambiguous and clear samples and they are mostly classified correctly; in contrast, the ambiguous and unclear samples are classified with high error but they are assigned with lower confidences. Thus, a selective classifier can sensibly reject the uncertain cases where the input data is not sufficient to make a decision.

In addition to the aforementioned investigation, we conducted a series of additional experiments on the CIFAR-10N dataset. These experiments were geared towards understanding the impact of various factors, including alterations in hyperparameters, changes in layer architecture, modifications in training loss functions, and the use of inter-expert variability as a prior distribution of sample weights. The influence of these factors on the performance of UCA-equipped networks were examined.

In table 2, we evaluated the AURC metric for a ResNet34-UCA network trained on CIFAR-10N with different values of the hyperparameters A and β. The sensitivity of the UCA model to these hyperparameters is notably low. Consequently, there is not a strong necessity for careful hyperparameter tuning.

Table 2. Comparison of AURC for different hyperparameters A and β for training ResNet34-UCA on the Random1 subset of CIFAR-10N.

A \ β	0.5	1	2.5	5	10
0.5	3.59%	2.97%	2.25%	1.98%	2.14%
0.65	4.17%	2.51%	2.15%	2.35%	2.12%
0.8	3.53%	2.65%	2.27%	2.42%	2.22%

The architecture of the UCA head offers flexibility in terms of the number of hidden layers and the number of neurons in each layer. In table 3, we present an assessment of various UCA head architectures during the training of ResNet34-UCA on the CIFAR-10N dataset. Our evaluation highlights a notable trend favouring simpler network structures characterised by fewer learnable parameters.

Table 4 illustrates the comparative analysis of the ResNet34-UCA model's performance on the Random1 subset of CIFAR-10N dataset using two distinct loss functions: negative log likelihood (NLL) and cross entropy (CE). The performance of the model showcases minimal sensitivity to

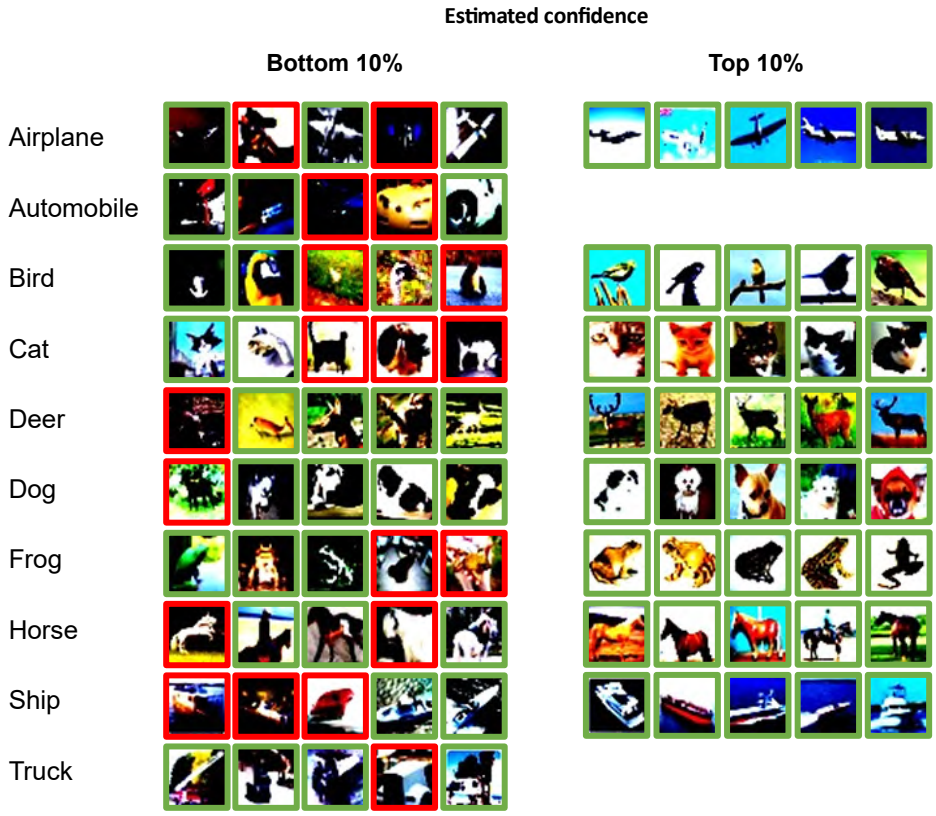

Figure 4. Test samples from the CIFAR-10N dataset. The correct classification and misclassification are shown with green and red borders respectively for ResNet34-UCA. The samples in the left column are selected from the lowest 10% of estimated confidences and the ones on the right are selected from the top 10%.

the choice of loss function.

In order to investigate the impact of selecting different sample uncertainty priors on the training of the ResNet34-UCA and PES-UCA networks, we conducted an investigation as illustrated in table 5. We introduced a novel sample uncertainty prior termed as the 'inter-expert prior', which is based on the inter-expert variability observed in the CIFAR-10N dataset. This inter-expert prior was used to assign uncertainty priors, p_1, p_2, and p_3, to training samples based on the presence of zero, one, or two affirmative labels from other annotation sets, respectively. The values of these uncertainty priors were determined empirically as 0.35, 0.5, and 0.55. Notably, the integration of the inter-expert prior led to enhancements in network performance. However, it is worth noting that the observed performance improvement was not significant enough to systematically justify multiple expert annotations.

Table 3. Comparison of AURC against UCA architecture, specifically, the number of hidden layers and the number of neurons in each layer.

Neurons in hidden layers	AURC
[64]	2.27%
[128]	1.98%
[512]	2.13%
[512, 128]	2.21%
[128, 64]	3.97%
[512, 128, 64]	2.16%

Table 4. Performance comparison of ResNet34-UCA trained with different loss functions.

	FC-Acc	AURC
ResNet34-UCA (CE loss)	90.49%	1.98%
ResNet34-UCA (NLL loss)	89.93%	2.25%

4.3. Image Classification on CheXpert

In order to further expand our assessments of UCA equipped networks, a series of experiments were conducted using the CheXpert dataset [13]. CheXpert is a huge dataset of chest X-ray images along with associated textual reports generated by radiologists. CheXpert is notable for its large

Table 5. Influence of sample uncertainty prior on ResNet34-UCA and PES-UCA training.

	FC-Acc	AURC
ResNet34-UCA (uniform prior)	90.49%	1.98%
ResNet34-UCA (inter-expert prior)	90.35%	1.67%
PES-UCA (uniform prior)	94.73%	0.90%
PES-UCA (inter-expert prior)	94.93%	0.81%

Table 6. FC-mAcc, FC-mAUC and AURC for image classification on CheXpert.

Method	β	**FC-mAcc**↑	**FC-mAUC**↑	**AURC**↓
SB [13]	-	87.00%	90.67%	6.27%
SB+UCA	0.1	87.60%	90.77%	5.58%
SB+UCA	0.5	87.92%	91.05%	5.41%
SB+UCA	1	86.83%	90.93%	5.79%
SB+UCA	5	86.52%	90.54%	5.81%

scale, over 200,000 chest X-rays from more than 65,000 patients, and its unique feature of handling uncertainty in radiology reports which makes it an ideal candidate for evaluating the efficacy of our proposed methodology. In establishing a foundation for our experiments, we adopt the baseline network proposed in [13] named Stanford baseline (SB). We extend SB by adding the proposed UCA extension which we call SB+UCA. We compare the performance of the baseline network with and without UCA extension in table 5 and figure 6.

As the CheXpert dataset is labelled for the presence of 14 common chest radiographic observations which might be concurrent, SB is configured as multiple binary classifiers, each one for detecting one of the diagnoses. Thus we evaluate the performance of the network by averaging the accuracy and AUC metrics over all the binary classifiers. In full coverage mode, we call them FC-mACC and FC-mAUC respectively. Both metrics are improved when the UCA extension is added to the baseline network, which shows the efficacy of the UCA method in dealing with noisy labels. The Area Under Risk-Coverage Curve (AURC) was also evaluated, which shows a significant improvement when UCA is added to SB, which confirms the capability of the proposed method to control the risk-coverage trade-off. The same trend is also seen in the RC curves shown in figure 5.

4.4. Image Segmentation on Gleason-2019

We used Gleason-2019 [20], figure 1. The dataset is tissue micro-array (TMA) images with multiple segmentation masks by up to six expert pathologists. Because of the large degree of heterogeneity in the cellular and glandular patterns associated with each Gleason grade, there is a significant inter-expert variability. We use PSPNet152 and UNet trained with CE as baselines, named PSPNet152-CE and UNet-CE. We connect the global UCA head with $K = 2$ hidden layers with 512 and 128 neurons to the last layer of PSPNet152 and to the last layer of the contracting path of UNet. We trained with STAPLE consensus [29] using equation (6) and CE as main task loss forming methods PSPNet152-UCA and UNet-UCA. We also connect the pixelwise UCA head to PSPNet152 with $K = 2$ convolutional layers with 512 and 128 filters and trained using equation (8) with STAPLE consensus and with the multi-expert annotations, forming methods PSPNet152-UCA-PW and PSPNet152-UCA-PW* respectively. We use the same hyperparameters $A = 0.75$ and $\beta = 12$ in all cases. The noisy label training methods evaluated above [17, 33, 1] are not applicable to segmentation. The results are in table 7 and figure 6. UNet-UCA has a similar FC-Jac as the original UNet but decreases AURC by more than 3pp. PSPNet152, as winner of the Gleason-2019 challenge [22], represents the state of the art for this dataset. PSPNet152-UCA boosts the FC-Jac and AURC by more than 5pp and 6pp respectively. UCA thus brings a significant boost to both baselines. PSPNet152-UCA-PW and PSPNet152-UCA-PW* have remarkably better AURCs with an FC-Jac on par with the global one. PSPNet152-UCA-PW* has the benefit of being self-sufficient and to not dependent on STAPLE.

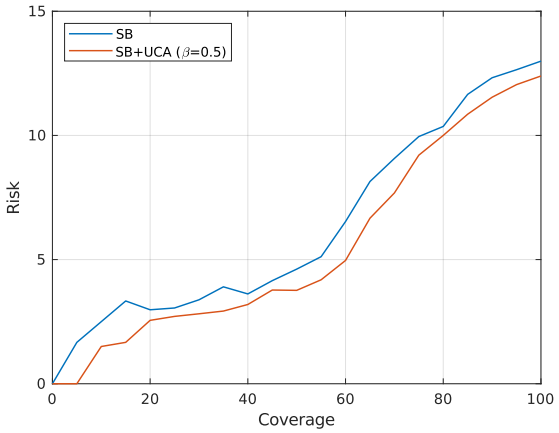

Figure 5. RC curves on CheXpert.

In the pursuit of determining the influence of hyperparameters on the training process of PSPNet152-UCA, we conducted an experiment to analyse the performance with respect to different hyperparameters, denoted as A and β, as shown in table 8. This evaluation was run in parallel with the experiment carried out on CIFAR10-N and reveals that the model's sensitivity to hyperparameter variations remains consistently low across different tasks and datasets.

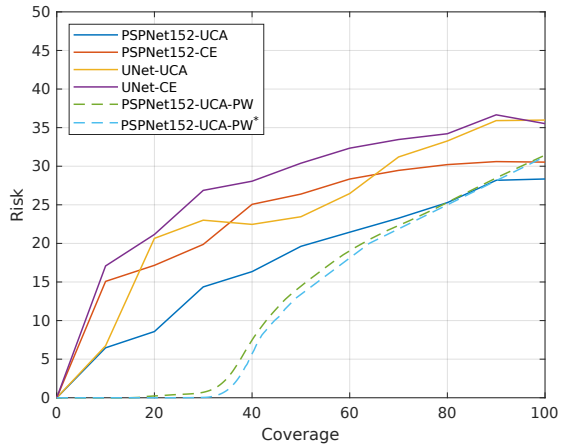

Figure 6. RC curves on Gleason-2019.

Table 7. FC-Jac and AURC for image segmentation on Gleason-2019.

Method	**FC-Acc↑**	**AURC↓**
UNet-CE	64.48%	27.79%
UNet-UCA (ours)	64.02%	24.11%
PSPNet152-CE	69.47%	23.74%
PSPNet152-UCA (ours)	71.65%	17.77%
PSPNet152-UCA-PW (ours)	68.56%	13.32%
PSPNet152-UCA-PW* (ours)	68.74%	12.74%

Table 8. AURC comparison for various hyperparameters A and β in PSPNet152-UCA training on Gleason-2019.

A \ β	8	12	16
0.5	17.88%	18.41%	21.20%
0.75	22.35%	17.77%	20.78%
1.0	25.43%	21.12%	19.14%

5. Conclusion

We have proposed UCA, the first method to handle training from noisy labels and confidence selective prediction simultaneously. UCA is generic: it does not require additional labels (specifically, confidence labels) and adapts to any existing neural architecture for various tasks, making it an adapted solution in the medical context. It shows a strong performance gain over existing selective prediction methods and is on par with the state-of-the-art in noisy label training when used in full coverage mode. Future work will test UCA in highly subjective medical image computing problems.

References

[1] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. In *NeurIPS*, 2021.

[2] Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *ICML*, 2021.

[3] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*, 2020.

[4] Filipe R Cordeiro and Gustavo Carneiro. A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations? In *SIBGRAPI*, 2020.

[5] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

[6] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

[7] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NeurIPS*, 2017.

[8] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017.

[9] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[11] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

[12] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, 2018.

[13] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI conference on artificial intelligence*, 2019.

[14] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

[15] Abhishek Kumar and Ehsan Amid. Constrained instance and class reweighting for robust learning under label noise. *arXiv preprint arXiv:2111.05428*, 2021.

[16] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

[17] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

[18] Jingya Liu, Bin Lou, Mamadou Diallo, Tongbai Meng, Heinrich von Busch, Robert Grimm, Yingli Tian, Dorin Comaniciu, Ali Kamen, David Winkel, et al. Detecting out-of-distribution via an unsupervised uncertainty estimation for prostate cancer diagnosis. In *MIDL*, 2021.

[19] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *NeurIPS*, 2018.

[20] Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, R Storey Wilson, et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical image analysis*, 50:167–180, 2018.

[21] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017.

[22] Yali Qiu, Yujin Hu, Peiyao Kong, Hai Xie, Xiaoliu Zhang, Jiuwen Cao, Tianfu Wang, and Baiying Lei. Automatic prostate gleason grading using pyramid semantic parsing network in digital histopathology. *Frontiers in Oncology*, 12:1–13, 2022.

[23] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. In *ICML*, 2019.

[24] Javier Rodriguez-Puigvert, David Recasens, Javier Civera, and Ruben Martinez-Cantin. On the uncertain single-view depths in colonoscopies. In *MICCAI*, 2022.

[25] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[26] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

[27] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *NeurIPS*, 2019.

[28] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *ACM FAccT*, 2021.

[29] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

[30] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, 2020.

[31] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

[32] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, 2019.

[33] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, 2019.

[34] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018.

[35] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *ICML*, 2021.