Direct Uncertainty Prediction for Medical Second Opinions

Maithra Raghu [*1,2] **Katy Blumer** [*2] **Rory Sayres** [2] **Ziad Obermeyer** [3] **Robert Kleinberg** [1]
Sendhil Mullainathan [4] **Jon Kleinberg** [1]

Abstract

The issue of disagreements amongst human experts is a ubiquitous one in both machine learning and medicine. In medicine, this often corresponds to doctor disagreements on a patient diagnosis. In this work, we show that machine learning models can be trained to give *uncertainty scores* to data instances that might result in high expert disagreements. In particular, they can identify patient cases that would benefit most from a *medical second opinion*. Our central methodological finding is that *Direct Uncertainty Prediction* (DUP), training a model to predict an uncertainty score directly from the raw patient features, works better than *Uncertainty Via Classification*, the two-step process of training a classifier and postprocessing the output distribution to give an uncertainty score. We show this both with a theoretical result, and on extensive evaluations on a large scale medical imaging application.

1. Introduction

In both the practice of machine learning and the practice of medicine, a serious challenge is presented by disagreements amongst human labels. Machine learning classification models are typically developed on large datasets consisting of (x_i, y_i) (data instance, label) pairs. These are collected (Russakovsky & Fei-Fei, 2010; Welinder & Perona, 2010) by assigning each raw instance x_i to multiple human evaluators, yielding several labels $y_i^{(1)}, y_i^{(2)}, ...y_i^{(n_i)}$. Unsurprisingly, these labels often have disagreements amongst them and must be carefully aggregated to give a single target value.

This label disagreement issue becomes a full-fledged clinical problem in the healthcare domain. Despite the human labellers now being highly trained medical experts (doctors), disagreements (on the diagnosis) persist (Van Such et al., 2017; Abrams et al., 1994; AAO, 2002; Gulshan et al., 2016; Rajpurkar et al., 2017). One example is (Van Such et al., 2017), where agreement between referral and final diagnoses in a cohort of two hundred and eighty patients is studied. *Exact* agreement is only found in 12% of cases, but more concerningly, 21% of cases have significant disagreements. This latter group also turns out to be the most costly to treat. Other examples are given by (Daniel, 2004), a study of tuberculosis diagnosis, showing that radiologists disagree with colleagues 25% of the time, and with *themselves* 20% of the time, and (Elmore et al., 2015), studying disagreement on cancer diagnosis from breast biopsies.

These disagreements arise not solely from random noise (Rolnick et al., 2017), but from expert judgment and bias. In particular, some patient cases x_i intrinsically contain features that result in greater expert uncertainty (e.g. Figure 2.) This motivates applying machine learning to *predict* which patients are likely to give rise to the most doctor disagreement. We call this the *medical second opinion* problem. Such a model could be deployed to automatically identify patients that might need a second doctor's opinion.

Mathematically, given a patient instance x_i, we are interested in assigning a scalar uncertainty score to x_i, $h(x_i)$ that reflects the amount of expert disagreement on x_i. For each x_i, we have multiple labels $y_i^{(1)}, y_i^{(2)}, ...y_i^{(n_i)}$, each corresponding to a different individual doctor's grade.

One natural approach is to first train a classifier mapping x_i to the $y_i^{(j)}$, e.g. via the empirical distribution of labels $\hat{\mathbf{p}}_i$. For ungraded examples, a measure of spread of the output distribution of the classifier (e.g. variance) could be used to give a score. We call this *Uncertainty via Classification (UVC)*.

An alternate approach, *Direct Uncertainty Prediction (DUP)*, is to learn a function h_{dup} directly mapping x_i to a scalar uncertainty score. The basic contrast with Uncertainty via Classification is illustrated in Figure 1. Our central methodological finding is that Direct Uncertainty Prediction (provably) works better than the two step process of Uncertainty via Classification.

[*]Equal contribution [1]Department of Computer Science, Cornell University [2]Google Brain [3]UC Berkeley School of Public Health [4]Chicago Booth School of Business. Correspondence to: Maithra Raghu <maithrar@gmail.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

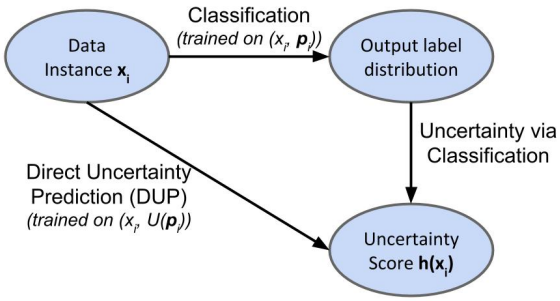

Figure 1. **Different ways of computing an uncertainty scores.** An uncertainty score $h(x_i)$ for x_i can be computed by the two step process of Uncertainty via Classification: training a classifier on pairs (data instance, empirical grade distribution from $y_i^{(j)}$) $(x_i, \mathbf{p}_i)$, and then post processing the classifier output distribution to get an uncertainty score. $h(x_i)$ can also be learned directly on x_i, i.e. Direct Uncertainty Prediction. DUP models are trained on pairs (data instance, target uncertainty function on empirical grade distribution), $(x_i, U(\mathbf{p}_i))$. Theoretical and empirical results support the greater effectiveness of Direct Uncertainty Prediction.

In particular, our three main contributions are the following:

1. We define simple methods of performing Direct Uncertainty Prediction on data instances x_i with multiple noisy labels. We prove that under a natural model for the data, DUP gives an *unbiased* estimate of the true uncertainty scores $U(x_i)$, while Uncertainty via Classification has a bias term. We then demonstrate this in a synthetic setting of mixtures of Gaussians, and on an image blurring detection task on the standard SVHN and CIFAR-10 benchmarks.

2. We train UVC and DUP models on a large-scale medical imaging task. As predicted by the theory, we find that DUP models perform better at identifying patient cases that will result in large disagreements amongst doctors.

3. On a small gold standard *adjudicated* test set, we study how well our existing DUP and UVC models can identify patient cases where the individual doctor grade disagrees with a consensus *adjudicated* diagnosis. This adjudicated grade is a proxy for the best possible doctor diagnosis. All DUP models perform better than all UVC models on all evaluations on this task, in both an uncertainty score setting and a ranking application.

2. Direct Uncertainty Prediction

Our core prediction problem, motivated by identifying patients who need a *medical second opinion*, centers around

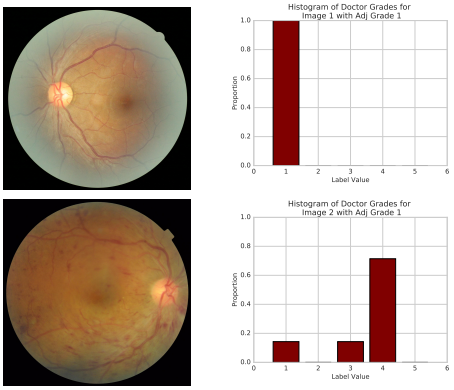

Figure 2. **Patient cases have features resulting in higher doctor disagreement.** The two rows give example datapoints in our dataset. The patient images x_i, x_j are in the left column, and on the right we have the empirical probability distribution (histogram) of the multiple individual doctor DR grades. For the top image, all doctors agreed that the grade should be 1, while there was a significant spread for the bottom image. When later performing an *adjudication* process (Section 5), where doctors discuss their initial diagnoses with each other and come to a consensus, both patient cases were given an *adjudicated* DR grade of 1.

learning a scalar *uncertainty scoring function* $h(x)$ on patient instances x, which signifies the amount of expert disagreement arising from x.

To do so, we must first define a *target* uncertainty scoring function $U(\cdot)$. Our data consists of pairs of the form (patient features, multiple individual doctor labels), $(x_i; y_i^{(1)}, y_i^{(2)}, ...y_i^{(n_i)})$ (Figure 2). Letting $c_1, ..., c_k$ be the different possible doctor grades, we can define the empirical grade distribution – the empirical *histogram*: $\hat{\mathbf{p}}_\mathbf{i} = [\hat{p}_i^{(1)}, ..., \hat{p}_i^{(k)}]$, with

$$\hat{p}_i^{(l)} = \frac{\sum_j \mathbb{1}_{y_i^{(j)}=c_l}}{n_i}$$

Our target uncertainty scoring function $U(\cdot)$ then computes an uncertainty score for x_i using the empirical histogram $\hat{\mathbf{p}}_\mathbf{i}$. One such function, which computes the probability that two draws from the empirical histogram will disagree is

$$U_{disagree}(x_i) = U_{disagree}(\hat{\mathbf{p}}_\mathbf{i}) = 1 - \sum_{l=1}^{k} (\hat{p}_i^{(l)})^2 \quad (1)$$

Another uncertainty score, which penalizes larger disagreements more, is the variance:

$$U_{var}(x_i) = U_{var}(\hat{\mathbf{p}}_\mathbf{i}) = \sum_{l=1}^{k} c_l \cdot (\hat{p}_i^{(l)})^2 - \left(\sum c_l \cdot \hat{p}_i^{(l)}\right)^2 \quad (2)$$

For a large family of these uncertainty scoring functions (including entropy, variance, etc) we can show that Direct

Uncertainty Prediction gives an unbiased estimate of $U(\hat{\mathbf{p}}_i)$, whereas uncertainty via classification has a bias term.

The key observation is that while we want our model to predict doctor disagreement, it does not see all the patient information the doctors do. In particular, the model must predict doctor disagreement based off of only x_i (in our setting, images). In contrast, human doctors see not only x_i but other extensive information, such as patient medical history, family medical history, patient characteristics (age, symptom descriptions, etc) (De Fauw et al., 2018).

Letting o denote all patient features seen by the doctors, we can think of x_i as being the image of o under a (many to one) mapping g, which hides the additional patient information, i.e. $x_i = g(o)$. Suppose there are k possible doctor grades, $c_1, .., c_k$. Let f denote the joint distribution over patient features and doctor grades. In particular, let O be a random variable representing patient features, and Y the doctor label for O. Then our density function assigns a probability to (patient features, doctor grade) pairs (o, y).

This can also be defined with a vectorized version of the grades: let $Y_l = \mathbf{1}_{Y=c_l}$, the event that O is diagnosed as c_l. Then we define the vector $\mathbf{Y} = [Y_1, ..., Y_k]$. f is therefore also a density over the points $f(O = o, \mathbf{Y} = \mathbf{y})$. Let the marginal probability of the patient features be f_O, with $f_O = \int_{\mathbf{y}} f(O, \mathbf{y})$.

Given an uncertainty scoring function $U(\cdot)$, we would like to predict the disagreement in labels amongst doctors who have seen the patient features O. But as the patient features O and doctor grades $\mathbf{Y}$ are jointly distributed according to f, this is just the uncertainty of the expected value of $\mathbf{Y}$ under the posterior of $\mathbf{Y}$ given o. In particular, we are interested in predicting:

$$U\left(\int_{\mathbf{y}} \mathbf{y} \cdot f(\mathbf{Y} = \mathbf{y}|O)\right) = U(\mathbf{E}[\mathbf{Y}|O])$$

This is a function taking as input a patient's features. For a particular patient's features o, we get a scalar uncertainty score given by

$$U(\mathbf{E}[\mathbf{Y}|O = o])$$

However, our model doesn't see o, but only $x = g(O)$. We make the mild assumptions that Y is conditionally independent of $g(O)$ given O, and that $g(\cdot)$ truly hides information, loosely that $O|g(O) = x$ is not a point mass (see Appendix for details.) In this setting, direct uncertainty prediction, h_{dup} computes the expectation of the uncertainty scores of all the possible posteriors, i.e.

$$\begin{aligned} h_{dup}(x) &= \mathbb{E}\left[U(\mathbb{E}[\mathbf{Y}|O])|g(O) = x\right] \\ &= \int_o U(\mathbb{E}[\mathbf{Y}|O = o]) f_O(o|g(O) = x) \end{aligned}$$

Uncertainty via classification h_{uvc} does this in reverse order, first computing the expected posterior, and assigning an uncertainty score to that:

$$\begin{aligned} h_{uvc}(x) &= U(\mathbb{E}[\mathbf{Y}|g(O) = x]) \\ &= U\left(\int_o \mathbb{E}[\mathbf{Y}|O = o] f_O(o|g(O) = x)\right) \end{aligned}$$

In this setting we can show

Theorem 1. *Using the above notation*

(i) h_{dup} *is an unbiased estimate of the true uncertainty*

(ii) For any concave uncertainty scoring function $U(\cdot)$ *(which includes* $U_{disagree}, U_{var}$*), uncertainty via classification,* h_{uvc} *has a bias term.*

The full proof is the Appendix. A sketch is as follows: the unbiased result arises from the tower rule (law of total expectations). The bias of h_{uvc} follows by the concavity of $U()$, Jensen's inequality, and the fact that $g(\cdot)$ truly hides some patient features. For $U_{disagree}$ and U_{var}, we can compute this bias term exactly (full computation in Appendix):

Corollary 1. *For* $U_{disagree}, U_{var}$ *the bias term is:*

(i) Bias of h_{uvc} *with* $U_{disagree}$*:*

$$\mathbb{E}_{g(O)}\left[\sum_l Var_{O|g(O)}\left(\mathbb{E}[Y_l|O]\middle|g(O)\right)\right]$$

(ii) Bias of h_{uvc} *with* U_{var}*:*

$$\mathbb{E}_{g(O)}\left[Var_{O|g(O)}\left(\sum_l l \cdot \mathbb{E}[Y_l|O]\middle|g(O)\right)\right]$$

In Sections 4, 5 we train both Direct Uncertainty Prediction (DUP) models and Uncertainty Via Classification (UVC) models on a large scale medical imaging task. However, to gain intuition for the theoretical results, we first study a toy case on a mixture of Gaussians.

2.1. Toy Example on Mixture of Gaussians

To illustrate the formalism in a simplified setting, we consider the following pedagogical toy example. Suppose our data is generated by a mixture of k Gaussians. Let $f_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, and q_i be mixture probabilities. Then $f(o, y = i) = q_i f_i(o)$ and the marginal $f_O(o) = \sum_{i=1}^k q_i f_i(o)$. Additionally, the probability of a particular class l given o, $f(y = l|o)$ is simply $\frac{q_l f_l(o)}{\sum_{i=1}^k q_i f_i(o)}$.

Two 1-D Gaussians: As the first, most simple case, suppose we have two one dimensional Gaussians, the first, $f_1 =$

Model Type	$(3d, 5G)$	$(5d, 4G)$	$(10d, 4G)$
UVC	69.1%	62.0%	56.0%
DUP	**74.6**%	**71.2**%	**63.4**%

Table 1. **DUP and UVC trained to predict disagreement on mixtures of Gaussians.** We train DUP and UVC models on different mixtures of Gaussians, with(nd, mG) denoting a mixture of m Gaussians in m dimensions. Results are in percentage AUC over 3 repeats. The means of the Gaussians are drawn iid from a multivariate normal distribution (full setup in Appendix.) We see that the DUP model performs much better than the UVC model at identifying datapoints with high disagreement in the labels.

$\mathcal{N}(-1, 1)$ and the second, $f_2 = \mathcal{N}(1, 1)$. Assume that the mixture probabilities q_1, q_2 are equal to 0.5. Given o drawn from this mixture $q_1 f_1 + q_2 f_2$, we'd like to estimate $U(f(y|o))$. Suppose the model sees $x = g(o) = |o|$, the absolute value of o. Then, DUP can estimate the uncertainty exactly:

$$\begin{aligned}\mathbb{E}\left[U(\mathbb{E}[\mathbf{Y}|O])|x = |o|\right] =& 0.5 \cdot U(\mathbb{E}[\mathbf{Y}|O = o]) \\ &+ 0.5 \cdot U(\mathbb{E}[\mathbf{Y}|O = -o]) \\ &= U(\mathbb{E}[\mathbf{Y}|O = o]) \\ &= U\Big([f(1|o), 1 - f(1|o)]\Big)\end{aligned}$$

where the third line follows by the symmetry of the two distributions, with

$$f(1|o) = \frac{0.5 f_1(o)}{0.5 f_1(o) + 0.5 f_2(o)}$$

On the other hand, the expected posterior over labels in UVC, $\mathbb{E}[\mathbf{Y}|x = |o|]$, is just $[0.5, 0.5]$, as by symmetry, given $x = g(o) = |o|$, o is equally likely to come from f_1 or f_2. So UVC outputs a constant uncertainty score $U([0.5, 0.5])$ for all $x = |o|$, despite the true varying uncertainty scores.

Training DUPs and UVCs on Mixture of Gaussians: In Table 1 we train DUPs and UVCs on a few different mixture of Gaussian settings. We generate data o from a Gaussian mixture with iid centers, and labels for the data using the posterior over the different centers given o. We use these labels to score o on its uncertainty (using $U_{disagree}$). We then train a model on $x = g(o) = |o|$ to predict whether x is low or high uncertainty. (Full details in Appendix.) We see that DUP performs consistently better than UVC.

2.2. Example on SVHN and CIFAR-10

Another empirical demonstration is given by training DUP and UVC to predict label agreement in an image blurring setting. For a source image in SVHN or CIFAR-10, we first apply a Gaussian blur, with a variance chosen for that source image. Then, we draw three noisy labels for the source image, where the noise distribution over labels corresponds to the severity of the image blur. For example, for an image that has a Gaussian blur of variance 0 (i.e. no blurring), the distribution over labels is a point mass on the true label. For an image that has been blurred severely, there is significant mass on incorrect labels. (Exact distributional details are given in the Appendix.) We train DUP and UVC models on this dataset and evaluate their ability to predict label disagreement. We again find that DUP models outperfom UVC models. This is despite the setting not directly mapping onto the statement of Theorem 1 – there is no obscuring function g. This suggests the benefits of DUP are more general than the precise theoretical setting. We also observe that the DUP and UVC models learn different features (see Appendix.)

Model	**SVHN** (disagree)	**CIFAR-10** (disagree)
UVC	75.8%	79.1%
DUP	**88.0**%	**85.3**%

Table 2. **DUP and UVC trained to predict label disagreement corresponding to image blurring on SVHN and CIFAR-10.** DUP outperforms UVC on predicting label disagreement on SVHN and CIFAR-10, where the labels are drawn from a noisy distribution that varies depending on how much blurring the source image has been subjected to. Full details in Appendix.

3. Related Work

The challenges posed by expert disagreement is an important one, and prior work has put forward several approaches to address some of these issues. Under the assumption that the noise distribution is conditionally independent of the data instance given the true label, (Natarajan et al., 2013; Sukhbaatar et al., 2014; Reed et al., 2014; Sheng et al., 2008) provide theoretical analysis along with algorithms to denoise the labels as training progresses, or efficiently collect new labels. However, the conditional independence assumption does not hold in our setting (Figure 2.) Other work relaxes this assumption by defining a domain specific generative model for how noise arises (Mnih & Hinton, 2012; Xiao et al., 2015; Veit et al., 2017) with some methods using additional clean data to pretrain models to form a good prior for learning. Related techniques have also been explored in semantic segmentation (Gurari et al., 2018; Kohl et al., 2018). Modeling uncertainty in the context of noisy data has also been looked at through Bayesian techniques (Kendall & Gal, 2017; Tanno et al., 2017), and (for different models) in the context of crowdsourcing by (Werling et al., 2015; Wauthier & Jordan, 2011). A related line of work (Dawid et al., 1979; Welinder & Perona, 2010) has looked at studying the per labeler error rates, which also requires the additional information of labeler ids, an assumption we relax. Most related is (Guan et al., 2018), where a multiheaded neural network is used to model different labelers. Surprisingly however, the best model is independent of image features, which is the source of signal in our experiments.

Task		Model Type	Performance (AUC)
Variance Prediction	UVC	Histogram-E2E	70.6%
Variance Prediction	UVC	Histogram-PC	70.6%
Variance Prediction	DUP	Variance-E2E	72.9%
Variance Prediction	DUP	Variance-P	74.4%
Variance Prediction	DUP	Variance-PR	74.6%
Variance Prediction	DUP	Variance-PRC	**74.8%**
Disagreement Prediction	UVC	Histogram-E2E	73.4%
Disagreement Prediction	UVC	Histogram-PC	76.6%
Disagreement Prediction	DUP	Disagree-P	**78.1%**
Disagreement Prediction	DUP	Disagree-PC	**78.1%**
Variance Prediction	DUP	Disagree-PC	73.3%
Disagreement Prediction	DUP	Variance-PRC	77.3%

Table 3. **Performance (percentage AUC) averaged over three runs for UVC and DUPs on Variance Prediction and Disagreement Prediction tasks**. The UVC baselines, which first train a classifier on the empirical grade histogram, are denoted Histogram-. DUPs are trained on either $T_{train}^{(disagree)}$ or $T_{train}^{(var)}$, and denoted Disagree-, Variance- respectively. The top two sets of rows shows the performance of the baseline (and a strengthened baseline Histogram-PC using Prelogit embeddings and Calibration) compared to Variance and Disagree DUPs on the (1) Variance Prediction task (evaluation on $T_{test}^{(var)}$) and (2) Disagreement Prediction task (evaluation on $T_{test}^{(disagree)}$). We see that in both of these settings, the DUPs perform better than the baselines. Additionally, the third set of rows shows tests a Variance DUP on the disagreement task, and vice versa for the Disagreement DUP. We see that both of these also perform better than the baselines.

4. Doctor Disagreements in DR

Our main application studies the effectiveness of Direct Uncertainty Predictors (DUPs) and Uncertainty via Classification (UVC) in identifying patient cases with high disagreements amongst doctors in a large-scale medical imaging setting. These patients stand to benefit most from a medical second opinion.

The task contains patient data in the form of retinal fundus photographs (Gulshan et al., 2016), large (587 x 587) images of the back of the eye. These photographs can be used to diagnose the patient with different kinds of eye diseases. One such eye disease is Diabetic Retinopathy (DR), which, despite being treatable if caught early enough, remains a leading cause of blindness (Ahsan, 2015).

DR is graded on a 5-class scale: a grade of 1 corresponds to *no* DR, 2 to *mild* DR, 3 to *moderate* DR, 4 to *severe* and 5 to *proliferative* DR (AAO, 2002). There is an important clinical threshold at grade 3, with grades 3 and above corresponding to *referable* DR (needing immediate specialist attention), and $1, 2$ being non-referable. Clinically, the most costly mistake is not referring referable patients, which poses a high risk of blindness.

Our main dataset T has many features typical of medical imaging datasets. T has larger but much fewer images than in natural image datasets such as ImageNet. Each image x_i has a few (typically one to three) individual doctor grades $y_i^{(1)}, ..., y_i^{(n_i)}$. These grades are also very noisy, with more than 20% of the images having large (referable/non-referable) disagreement amongst the grades.

4.1. Task Setup

In this section we describe the setup for training variants of DUPs and UVCs using a train test split on T. We outline the resulting model performances in Table 3, which measure how successful the models are in identifying cases where doctors most disagree with each other and consequently where a medical second opinion might be most useful. In Section 5, we perform a different evaluation (disagreement with consensus) of the best performing DUPs and UVCs on a special, gold standard *adjudicated* test set. In both evaluation settings, we find that DUPs noticeably outperform UVCs.

The DUP and UVC models are trained and evaluated using a train/test split on T, T_{train}, T_{test}. This split is constructed using the patient ids of the $x_i \in T$, with 20% of patient ids being set aside to form T_{test} and 80% to form T_{train} (of which 10% is sometimes used as a validation set.) Splitting by patient ids is important to ensure that multiple images $x_i, x_j \in T$ corresponding to a single patient are correctly split (Gulshan et al., 2016).

We apply $U_{disagree}(\cdot)$ to the x_i in T_{train}, T_{test} with more than one doctor label to form a new train/test split $T_{train}^{(disagree)}, T_{test}^{(disagree)}$. We repeat this with $U_{var}(\cdot)$ to also form a train/test split $T_{train}^{(var)}, T_{test}^{(var)}$. These two datasets capture the two different medical interpretations of DR grades:

Categorical Grade Interpretation: The DR grades can be interpreted as categorical classes, as each grade has specific features associated with it. A grade of 2 always means microaneurysms, while a grade of 5 can refer to lesions or laser scars (AAO, 2002). The $T_{train}^{(disagree)}, T_{test}^{(disagree)}$ data measures disagreement in this categorical setting.

Continuous Grade Interpretation: While there are specific features associated with each DR grade, patient conditions tend to progress sequentially through the different DR grades. The $T_{train}^{(var)}, T_{test}^{(var)}$ data thus accounts for the magnitude of differences in doctor grades.

Having formed $T_{train}^{(disagree)}, T_{test}^{(disagree)}$ and $T_{train}^{(var)}, T_{test}^{(var)}$, which consist of pairs $(x_i, U_{disagree}(\hat{\mathbf{p}}_\mathbf{i}))$ and $(x_i, U_{var}(\hat{\mathbf{p}}_\mathbf{i}))$ respectively, we binarize the uncertainty scores $U_{disagree}(\hat{\mathbf{p}}_\mathbf{i}), U_{var}(\hat{\mathbf{p}}_\mathbf{i})$ into 0 (low uncertainty) or 1 (high uncertainty) to form our final prediction targets. We denote these $U_{disagree}^B(\hat{\mathbf{p}}_\mathbf{i}), U_{var}^B(\hat{\mathbf{p}}_\mathbf{i})$. More details on this can be found in Appendix Section D.

4.2. Models and First Experimental Results

We train both UVCs and DUPs on this data. All models rely on an Inception-v3 base that, following prior work (Gulshan et al., 2016), is initialized with pretrained weights on ImageNet. The UVC is performed by first training a classifier h_c on $(x_i, \hat{\mathbf{p}}_\mathbf{i})$ pairs in T_{train}. The output probability distribution of the classifier, $\tilde{\mathbf{p}}_i = h_c(x_i)$ is then used as input to the uncertainty scoring function $U(\cdot)$, i.e. $h_{uvc}(x_i) = U \circ h_c(x_i)$ In contrast, the DUPs are trained directly on the pairs $(x_i, U_{disagree}^B(\hat{\mathbf{p}}_\mathbf{i}))$, $(x_i, U_{var}^B(\hat{\mathbf{p}}_\mathbf{i}))$, i.e. $h_{dup}(x_i)$ directly tries to learn the value of $U^B(\hat{\mathbf{p}}_\mathbf{i})$

The results of evaluating these models (on $T_{test}^{(disagree)}$ and $T_{var}^{(disagree)}$) are given in Table 3. The *Variance Prediction* task corresponds to evaluation on $T_{var}^{(disagree)}$, and the *Disagreement Prediction* task to evaluation on $T_{test}^{(disagree)}$. Both tasks correspond to identifying patients where there is high disagreement amongst doctors. As is typical in medical applications due to class imbalances, performance is given via area under the ROC curve (AUC) (Gulshan et al., 2016; Rajpurkar et al., 2017).

From the first two sets of rows, we see that DUP models perform better than their UVC counterparts on both tasks. The third set of rows shows the effect of using a variance DUP (Variance-PRC) on the disagreement task and a disagree DUP (Disagree-PC) on the variance task. While these don't perform as well as the best DUP models on their respective tasks, they still beat both the baseline and the strengthened baseline. Below we describe some of the different UVC and DUP variants, with more details in Appendix Section D.

UVC Models The UVC models are trained on (image, empirical grade histogram) $(x_i, \hat{\mathbf{p}}_i)$ pairs, and denoted *Histogram-* in Table 3. The simplest UVC is *Histogram-E2E*, the same model used in (Gulshan et al., 2016). We improved this baseline by instead taking the prelogit embeddings of Histogram-E2E, and training a small neural network (fully connected, two hidden layers width 300) with temperature scaling (as in (Guo et al., 2017)) only on x_i with multiple labels. This gave the strengthened baseline *Histogram-PC*.

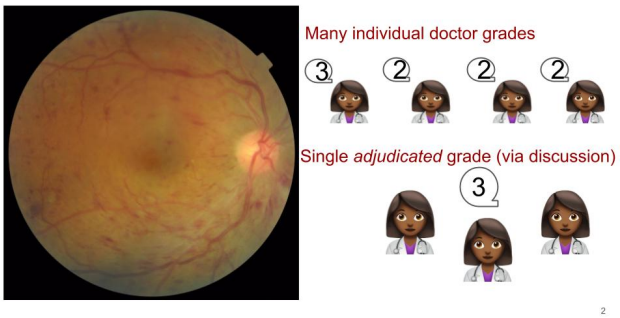

Figure 3. **Labels for the adjudicated dataset A.** The small, gold standard adjudicated dataset A has a very different label structure to the main dataset T. Each image has many individual doctor grades (typically more than 10 grades). These doctors also tend to be specialists, with higher rates of agreement. Additionally, each image has a single *adjudicated* grade, where three doctors first grade the image individually, and then come together to discuss the diagnosis and finally give a single, *consensus diagnosis*.

DUP Variance Models The simplest Variance DUP is *Variance-E2E*, which is analogous to Histogram-E2E, except trained on $T_{train}^{(var)}$. This performed better than Histogram-E2E, but as $T_{train}^{(var)}$ is small for an Inception-v3, we trained a small neural network (fully connected, two hidden layers width 300) on the prelogit embeddings, called *Variance-P*. Small variants of *Variance-P* (details in Appendix Section D) give *Variance-PR*, and *Variance-PRC*.

DUP Disagreement Models Informed by the variance models, the *Disagree-P* model was designed exactly like the *Variance-P* model (a small fully connected network on prelogit embeddings), but trained on $T_{train}^{(disagree)}$. A small variant of this with calibration gave *Disagree-PC*.

In the Appendix, we demonstrate similar results using entropy as the uncertainty function, as well as experiments studying convergence speed and finite sample behaviour of DUP and UVC. We find that the performance gap between DUP and UVC is robust to train data size, and manifests early in training.

	Model Type	Majority	Median	Majority $= 1$	Median $= 1$	Referable
UVC	Histogram-E2E-Var	78.1%	78.2%	81.3%	78.1%	85.5%
UVC	Histogram-E2E-Disagree	78.5%	78.5%	80.5%	77.0%	84.2%
UVC	Histogram-PC-Var	77.9%	78.0%	80.2%	77.7%	85.0%
UVC	Histogram-PC-Disagree	79.0%	78.9%	80.8%	79.2%	84.8%
DUP	Variance-PR	80.0%	79.9%	83.1%	80.5%	85.9%
DUP	Variance-PRC	79.8%	79.7%	82.7%	80.2%	85.9%
DUP	Disagree-P	**81.0%**	80.8%	**84.6%**	**81.9%**	**86.2%**
DUP	Disagree-PC	80.9%	**80.9%**	84.5%	81.8%	**86.2%**

Table 4. **Evaluating models (percentage AUC) on predicting disagreement between an average individual doctor grade and the adjudicated grade.** We evaluate our models's performance using multiple different aggregation metrics (majority, median, binarized non-referable/referable median) as well as special cases of interest (no DR according to majority, no DR according to median). We observe that **all** direct uncertainty models (Variance-, Disagree-) outperform *all* classifier-based models (Histogram-) on *all* tasks.

5. Predicting Disagreement with Consensus: Adjudicated Evaluation

Section 4 trained DUPs and UVCs on T_{train}, and evaluated them on their ability to identify patient cases where individual doctors were most likely to disagree with each other. Here, we take these trained DUPs/UVCs, and perform an *adjudicated* evaluation, to satisfy two additional goals.

Firstly, and most importantly, the clinical question of interest is not only in identifying patients where individual doctors disagree with each other, but cases where a more thorough diagnosis – the *best possible* doctor grade – would disagree significantly with the *individual* doctor grade. Evaluation on a gold-standard *adjudicated* dataset A enables us to test for this: each image $x_i \in A$ not only has many individual doctor grades (by specialists in the disease) but also a single *adjudicated* grade. This grade is determined by a group of doctors seeking to reach a *consensus* on the diagnosis through discussion (Krause et al., 2018). Figure 3 illustrates the setup.

We can thus evaluate on this question by seeing if high model uncertainty scores correspond to disagreements between the (average) individual doctor grade and the adjudicated grade. More specifically, we compute different aggregations of the individual doctor grades for $x_i \in A$, and give a binary label for whether this aggregation agrees with the adjudicated grade (0 for agreement, 1 for disagreement). We then see if our model uncertainty scores is predictive of the binary label.

Secondly, our evaluation on A also provides a more accurate reflection of our models's performance, with less confounding noise. The labels in A (both individual and adjudicated) are much cleaner, with greater consistency amongst doctors. As A is used *solely* for evaluation (all evaluated models are trained on T_{train}, Section 4), this introduces a distribution shift, but the predicted uncertainty scores transfer well. The results are shown in Table 4. We evaluate on several different aggregations of individual doctor grades. Like (Gulshan et al., 2016), we compare agreement between the majority vote of the individual grades and the adjudicated grades. To compensate for a bias of individual doctors giving lower DR grades (Krause et al., 2018), we also look at agreement between the median individual grade and adjudicated grade. Additionally, we look at referable/non-referable DR grade agreement. We binarize both the individual doctor grades and the adjudicated grade into $0/1$ non-referable/referable, and check agreement between the median binarized grades and adjudicated grade. Finally, we also look at the special case where the average doctor grade is 1 (no DR), and compare agreement with the adjudicated grade.

We evaluate both baseline models (Histogram-E2E, Histogram-PC) as well as the best performing DUPs, (Variance-PR, Variance-PRC, Disagree-P, Disagree-PC.) The additional -Var, -Disagree suffixes on the baseline models indicate which uncertainty function (U_{var} or $U_{disagree}$) was used to postprocess the classifier output distribution $\tilde{\mathbb{p}}$ to get an uncertainty score. We find that *all* DUPs outperform *all* the baselines on *all* evaluations.

5.1. Ranking Evaluation

A frequent practical challenge in healthcare is to *rank* cases in order of hardest (needing most attention) to easiest (needing least attention), (Harrell Jr et al., 1984). Therefore, we evaluate how well our models can rank cases from greatest disagreement between the adjudicated and individual grades to least disagreement between the adjudicated and individual grades. To do this however, we need a *continuous* ground truth value reflecting this disagreement, instead of the binary $0/1$ agree/disagree used above. One natural way to do this is to compute the *Wasserstein* distance between the empirical histogram (individual grade distribution) and the adjudicated grade.

At a high level, the Wasserstein distance computes the minimal cost required to move a probability distribution $\mathbf{p}^{(1)}$ to

	Prediction Type	Absolute Val	2-Wasserstein	Binary Disagree
UVC	Histogram-E2E-Var	0.650	0.644	0.643
UVC	Histogram-E2E-Disagree	0.645	0.633	0.643
UVC	Histogram-PC-Var	0.638	0.639	0.619
UVC	Histogram-PC-Disagree	0.660	0.655	0.649
DUP	Variance-PR	0.671	0.664	0.660
DUP	Variance-PRC	0.665	0.658	0.656
DUP	Disagree-P	**0.682**	**0.670**	**0.676**
DUP	Disagree-PC	0.680	0.669	0.675
	2 Doctors	0.460	0.448	0.455
	3 Doctors	0.585	0.576	0.580
	4 Doctors	0.641	0.634	0.644
	5 Doctors	0.676	0.670	0.675
	6 Doctors	0.728	0.712	0.718

Table 5. **Ranking evaluation of models uncertainty scores using Spearman's rank correlation.** In the top set of rows, we compare the ranking induced by the model uncertainty scores to the (ground truth) ranking induced by the Wasserstein distance between the empirical grade histogram and the adjudicated grade. We use three different metrics for evaluating Wasserstein distance: absolute value distance, 2-Wasserstein and Binary agree/disagree (more details in Appendix F.) Again, we see that *all* DUPs outperform *all* baselines on *all* metrics. The second set of rows provides another way to interpret these results. We subsample n doctors to create a new subsampled empirical grade histogram, and compare the ranking induced by the Wasserstein distance between this and the adjudicated grade to the ground truth ranking. We can thus say that the average DUP ranking corresponds to having 5 doctor grades, and the average UVC ranking corresponds to 4 doctor grades.

a probability distribution $\mathbf{p}^{(2)}$ with respect to a given metric $d(\cdot)$. In our setting, $\mathbf{p}^{(1)}$ is the empirical histogram $\hat{\mathbf{p}}_i$ of x_i, and $\mathbf{p}^{(2)}$ is the point mass at the adjudicated grade a_i. When one distribution is a point mass, the Wasserstein distance has a simple interpretation:

Theorem 2. *Let $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ be two probability distributions, with $\mathbf{p}^{(2)}$ a point mass with non-zero value a. Let $d(\cdot)$ be a given metric. The Wasserstein distance between $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$, $||\mathbf{p}^{(1)} - \mathbf{p}^{(2)}||_w$ with respect to $d(\cdot)$ can be written as*

$$||\mathbf{p}^{(1)} - \mathbf{p}^{(2)}||_w = \mathbb{E}_{C\sim\mathbf{p}^{(1)}}[d(C, a)]$$

The proof is in Appendix F. In our setting, the theorem says that the (continuous) disagreement score for $x_i \in A$ is just the expected distance between a grade drawn from the empirical histogram and the adjudicated grade. We consider three different distance functions $d(\cdot)$: (a) the absolute value of the grade difference, (b) the $2 - Wasserstein$ distance, a metrization of the squared distance penalizing large grade differences more (details in Appendix F) and (c) a $0/1$ binary agree/disagree metric, in line with the categorical and continous interpretations of DR grades, Section 4.

We compare the ranking induced by this continuous disagreement score on A with the ranking induced by the model's predicted uncertainty scores. To evaluate the similarity of these rankings, we use Spearman's rank correlation (Spearman, 1904), which takes a value between $[-1, 1]$. A -1 indicates perfect negative rank correlation, 1 a perfect positive rank correlation and 0 no correlation. The results are shown in Table 5. Similar to Table 4, we observe strong performance with DUPs: all DUPs beat all the baselines on all the different distances.

This task also enables a natural comparison between the models and doctors. In particular, we can compute a third ranking over A, by sampling n individual doctor grades, and computing the Wasserstein distance between this subsampled empirical histogram and the adjudicated grade. This experiment tells us how many doctor grades are needed to give a ranking as accurate as the models. For DUPs, we need on average 5 doctors, while for the UVC baseline, we need on average 4 doctors.

6. Discussion

In this paper, we show that machine learning models can successfully be used to predict data instances that give rise to high expert disagreement. The main motivation for this prediction problem is the medical domain, where some patient cases result in significant differences in doctor diagnoses, and may benefit greatly from a medical second opinion. We show, both with a formal result and through extensive experiments, that Direct Uncertainty Prediction, which learns an uncertainty score directly from the raw patient features, performs significantly better than Uncertainty via Classification. Future work might look at transferring these techniques to different data modalities, and extending the applications to machine learning data denoising.

ACKNOWLEDGMENTS

We thank Varun Gulshan and Arunachalam Narayanaswamy for detailed advice and discussion on the model, and David Sontag and Olga Russakovsky for specific suggestions. We also thank Quoc Le, Martin Wattenberg, Jonathan Krause, Lily Peng and Dale Webster for general feedback. We thank Naama Hammel and Zahra Rastegar for helpful medical insights. Robert Kleinberg's work was partially supported by NSF grant CCF-1512964.

References

AAO. *International Clinical Diabetic Retinopathy Disease Severity Scale Detailed Table*. American Academy of Ophthalmology, 2002.

Abrams, L. S., Scott, I. U., Spaeth, G. L., Quigley, H. A., and Varma, R. Agreement among optometrists, ophthalmologists, and residents in evaluating the optic disc for glaucoma. *Ophthalmology*, 101(10):1662–1667, 1994.

Ahsan, H. Diabetic retinopathy – biomolecules and multiple pathophysiology. *Diabetes and Metabolic Syndrome: Clincal Research and Review*, pp. 51–54, 2015.

Chen, I., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 3543–3554. Curran Associates, Inc., 2018.

Daniel, T. M. *Toman's tuberculosis. Case detection, treatment and monitoring: questions and answers*. ASTMH, 2004.

Dawid, P., Skene, A. M., Dawidt, A. P., and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, pp. 20–28, 1979.

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., ODonoghue, B., Visentin, D., et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.

Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt, S. J., et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313(11):1122–1132, 2015.

Guan, M. Y., Gulshan, V., Dai, A. M., and Hinton, G. E. Who said what: Modeling individual labelers improves classification, 2018. URL https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16970.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. Q., Mega, J., and Webster, D. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316 (22):2402–2410, 2016.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. abs/1706.04599, 2017. URL http://arxiv.org/abs/1706.04599.

Gurari, D., He, K., Xiong, B., Zhang, J., Sameki, M., Jain, S. D., Sclaroff, S., Betke, M., and Grauman, K. Predicting foreground object ambiguity and efficiently crowdsourcing the segmentation (s). *International Journal of Computer Vision*, 126(7):714–730, 2018.

Harrell Jr, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., and Rosati, R. A. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3 (2):143–152, 1984.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, volume 30, pp. 5580–5590, 2017.

Kohl, S. A., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K. H., Eslami, S., Rezende, D. J., and Ronneberger, O. A probabilistic u-net for segmentation of ambiguous images. *arXiv preprint arXiv:1806.05034*, 2018.

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., Peng, L., and Webster, D. R. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125 8:1264–1272, 2018.

Mnih, V. and Hinton, G. Learning to label aerial images from noisy data. *International Conference on Machine Learning*, 2012.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26*, pp. 1196–1204. 2013.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., and Ng, A. Y. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. abs/1711.05225, 2017. URL http://arxiv.org/abs/1711.05225.

Reed, S. E., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. abs/1412.6596, 2014. URL http://arxiv.org/abs/1412.6596.

Rolnick, D., Veit, A., Belongie, S. J., and Shavit, N. Deep learning is robust to massive label noise. *CoRR*, abs/1705.10694, 2017. URL http://arxiv.org/abs/1705.10694.

Russakovsky, O. and Fei-Fei, L. Attribute learning in large-scale datasets. In *European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes*, 2010.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 614–622. ACM, 2008. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401965.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology*, pp. 72–101, 1904.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. *CoRR*, abs/1406.2080, 2014. URL http://arxiv.org/abs/1406.2080.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328. JMLR. org, 2017.

Tanno, R., Worrall, D., Ghosh, A., Kaden, E., N. Sotiropoulos, S., Criminisi, A., and C. Alexander, D. Bayesian image quality transfer with cnns: Exploring uncertainty in dmri super-resolution. *Medical Image Computing and Computer Assisted Intervention*, pp. 611–619, 2017.

Van Such, M., Lohr, R., Beckman, T., and Naessens, J. M. Extent of diagnostic agreement among medical referrals. *Journal of evaluation in clinical practice*, 23(4):870–874, 2017.

Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., and Belongie, S. J. Learning from noisy large-scale datasets with minimal supervision. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6575–6583, 2017.

Wauthier, F. L. and Jordan, M. I. Bayesian bias mitigation for crowdsourcing. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, pp. 1800–1808. 2011.

Welinder, P. and Perona, P. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2010, San Francisco, CA, USA, 13-18 June, 2010*, pp. 25–32, 2010. doi: 10.1109/CVPRW.2010.5543189. URL https://doi.org/10.1109/CVPRW.2010.5543189.

Werling, K., Chaganty, A. T., Liang, P. S., and Manning, C. D. On-the-job learning with bayesian decision theory. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 3465–3473. 2015.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2691–2699, 2015.