# Bayesian approaches for Quantifying Clinicians' Variability in Medical Image Quantification

**Jaeik Jeon** [*1] **Yeonggul Jang** [*1 2] **Youngtaek Hong** [1] **Hackjoon Shim** [1] **Sekeun Kim** [3]

## Abstract

Medical imaging, including MRI, CT, and Ultrasound, plays a vital role in clinical decisions. Accurate segmentation is essential to measure the structure of interest from the image. However, manual segmentation is highly operator-dependent, which leads to high inter and intra-variability of quantitative measurements. In this paper, we explore the feasibility that Bayesian predictive distribution parameterized by deep neural networks can capture the clinicians' inter-intra variability. By exploring and analyzing recently emerged approximate inference schemes, we evaluate whether approximate Bayesian deep learning with the posterior over segmentations can learn inter-intra rater variability both in segmentation and clinical measurements. The experiments are performed with two different imaging modalities: MRI and ultrasound. We empirically demonstrated that Bayesian predictive distribution parameterized by deep neural networks could approximate the clinicians' inter-intra variability. We show a new perspective in analyzing medical images quantitatively by providing clinical measurement uncertainty.

## 1. Introduction

In quantitative medical image analysis, there is large inter-intra observer variability. Although this uncertainty arises from a variety of external sources, such as lack of precision in imaging devices or systematic measurement errors, manual segmentation is also dependent on the experience and perception of radiologists. The ambiguities of quantitative measurement has the potential to mislead poor clinical decision-making. Due to this uncertainty, there has been a clinician consensus that a deep learning model should provide distributions over possible outcomes rather than point estimates.

Most previous works on modeling uncertainty from inter/intra-rater variability focus on quantifying uncertainty at the per-pixel level. That is, many literatures aim to quantify clinicians' inter-intra variability in segmentation. (Baumgartner et al., 2019) proposed a probabilistic model that can generate segmentation samples closely resembling several annotators' ground-truth distribution. Kendall & Gal (2017) proposed a method to quantify epistemic and aleatoric uncertainty in semantic segmentation tasks. In Hu et al. (2019), they have defined aleatoric uncertainty to the per-pixel variance among the multiple segmentation by clinicians, built upon Kohl et al. (2018); Gal & Ghahramani (2016). Kohl et al. (2018) proposed a probabilistic U-Net, a combination of a conditional variational auto-encoder and U-Net, for the segmentation of ambiguous images. It provides multiple plausible semantic segmentation hypotheses.

These prior works focus on returning calibrated uncertainty estimates to inform clinicians about the model confidence of its prediction. However, in quantitative medical image analysis, segmentation itself lacks clinical significance, and the uncertainty of segmentation is also lacking accordingly. Its endpoint is often to derive clinical measurements for disease diagnosis and decision-making, which is very important in the medical field. Therefore, in contrast to the previous approaches, we aim to model uncertainty on clinical indices, not per-pixel level uncertainty. Since the clinical indices such as volume and diameter are used in clinical decision in practice, the uncertainty of clinical indices has a greater clinical significance than pixel-wise uncertainty modeling.

In this paper, we study the feasibility of a Bayesian predictive distribution parameterized by deep neural networks to model expert variability. By exploring and analyzing neural linear, MC dropout and deep ensembles, we evaluate whether approximate Bayesian deep learning with the posterior over segmentations can learn inter-intra rater variability both in segmentation and clinical measurements. The

[*]Equal contribution [1]CONNECT-AI Research Center, Yonsei University College of Medicine, Seoul, South Korea [2]Graduate School of Medical Science, Brain Korea 21 Project, Yonsei University College of Medicine, Seoul, 03722, South Korea [3]Gordon Center for Medical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA. Correspondence to: Sekeun Kim <skim207@mgh.harvard.edu>.

*Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML)*. 

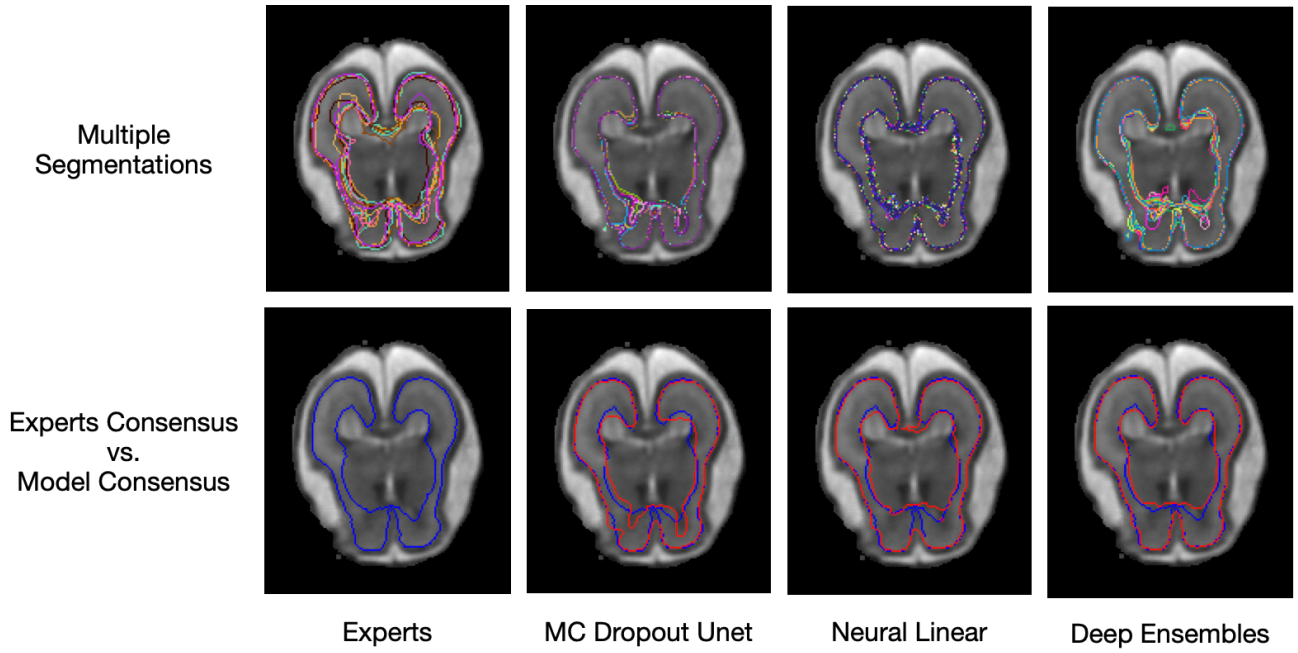



Figure 1: Visualization of the learned variability in segmentation over the 2D MRI image. First rows presents multiple segmentations generated from experts and various Bayesian segmentation models. We see that all presented Bayesian networks successfully approximate the inter-variability of experts' annotations. Different annotations due to inter-rater variability are indicated by randomly selected colors.

methods are evaluated on two datasets: a) the QUBIQ2021 dataset, which is used for evaluating the segmentation performance, and b) the IVUS dataset, which is used for evaluating a clinical endpoint. We do not propose new methods, but rather suggest that Bayesian approaches can approximate expert variability.

Our main contributions of this paper are:

1. We explore and analyze three recently emerged Bayesian uncertainty estimation techniques that are scalable to the medical imaging segmentation task.
2. We show that the explored Bayesian methods are effective to quantify clinician's inter-intra variability in segmentation, but also in clinical measurements over the various medical image quantification tasks.
3. We provide inclusion of clinical endpoints in the evaluation with different medical imaging modalities including MRI and Ultrasound using publicly available datasets.

## 2. Background and Related Works

In this section, we describe a general Bayesian approach to model the variability of experts' annotation of medical images. We hypothesize that the predictive distribution learned from data matches the variability of the expert's annotations. Specifically, we exploit Bayesian neural networks that place a probability distribution over the weight parameters with the hope that the distribution of predicted segmentation is identical to the distribution of multiple segmentation generated by multiple clinicians.

Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ be the training inputs and $Y = \{y_i\}_{i=1}^N$ be the outputs where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. We express uncertainty on the model parameters $\boldsymbol{\omega} \in \Omega$ by defining a likelihood distribution $p(Y|\mathbf{X}, \boldsymbol{\omega})$ and placing a prior distribution $p(\boldsymbol{\omega})$. From Bayes' rule, we then achieve the posterior distribution over the parameter space $p(\boldsymbol{\omega}|\mathbf{X}, Y) = \frac{p(Y|\mathbf{X},\boldsymbol{\omega})p(\boldsymbol{\omega})}{\int_\Omega p(Y|\mathbf{X},\boldsymbol{\omega})p(\boldsymbol{\omega})d\boldsymbol{\omega}}$ and the predictive distribution for a new input $\mathbf{x}^*$ $p(y^*|\mathbf{x}^*, \mathbf{X}, Y) = \int_\Omega p(y^*|\mathbf{x}^*, \boldsymbol{\omega})p(\boldsymbol{\omega}|\mathbf{X}, Y)d\boldsymbol{\omega}$. The tricky part is from the integration in the normalizing factor for which the closed form is often intractable, so is the predictive distribution.

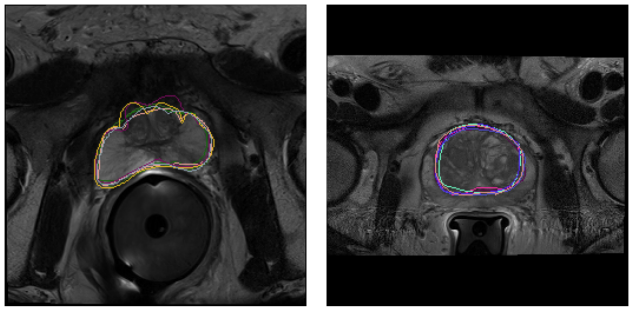
In this work, we focus on a variational inference which has been extensively used in Bayesian deep learning community (Blundell et al., 2015; Gal & Ghahramani, 2016; Kingma & Welling, 2013) which recasts marginalisation (integration) as an optimisation problem. The posterior distribution is approximated by a variational distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ parametrised by $\boldsymbol{\theta}$ such that the variational distribution is the closest distirbution to the posterior. We

consider Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) $\mathrm{KL}[q_{\boldsymbol{\theta}}(\boldsymbol{\omega})||p(\boldsymbol{\omega}|\mathbf{X}, Y)]$ between $q_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ and $p(\boldsymbol{\omega}|\mathbf{X}, Y)$ as a measure of closeness. The optimal variational distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ is achieved by minimizing KL divergence w.r.t the variational parameters $\boldsymbol{\theta}$, which is equivalent to maximizing evidence lower bound (ELBO) $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{\omega})}[\log p(Y|\mathbf{X}, \boldsymbol{\omega})] - \mathrm{KL}[q_{\boldsymbol{\theta}}(\boldsymbol{\omega})||p(\boldsymbol{\omega})]$. In order to achieve practical gradient estimator of the ELBO, we reparametrize the random parameter $\boldsymbol{\omega} \sim q_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ as $\boldsymbol{\omega} = g(\boldsymbol{\theta}, \boldsymbol{\epsilon})$, for a differentiable function $g(\cdot)$ and $\boldsymbol{\epsilon}$ being an auxiliary random variable drawn from $p(\boldsymbol{\epsilon})$, introduced in (Kingma & Welling, 2013). By applying the reparametrization trick, we are able to get a differentiable MC estimator of the ELBO w.r.t. $\boldsymbol{\theta}$ that is compatible with backpropagation:

$$\hat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \log p(Y|\mathbf{X}, \boldsymbol{\omega}_t = g(\boldsymbol{\theta}, \boldsymbol{\epsilon}_t)) - \mathrm{KL}[q_{\boldsymbol{\theta}}(\boldsymbol{\omega})||p(\boldsymbol{\omega})] \tag{1}$$

## 3. Methods

In this section, we will describe how we construct deep segmentation networks in Bayesian ways. We applied four scalable Bayesian methods created for different purposes in previous papers to Unet (Ronneberger et al., 2015).



Prostate Task01 Prostate Task02

Figure 2: Visualization of the inter-variability over the prostate images in QUBIQ 2021 challenge.

### 3.1. Variational Inference

#### 3.1.1. Estimate Uncertainty in the last layer

Riquelme et al. (2018) performs Bayesian linear regression on features extracted from the last layer of the neural network to achieve accurate uncertainty estimates for Thompson sampling, which is named neural linear. Inspired by the neural linear (Riquelme et al., 2018), we employ Unet (Ronneberger et al., 2015) as a deterministic feature extractor and obtain uncertainty estimates in the last layer. The last layer in Unet is a $1 \times 1$ convolution layer which projects the information over each channel to a class score, followed by the softmax function. We can think of this as a fully connected layer replicated for each pixel, which takes the corresponding channel vector as an input. By placing a probability distribution over the weights in the filter of the last layer, we capture the distribution in the aggregation process of information containing over channels in every pixel of the image.

We approximate the true posterior of weights in each filter with a fully factorised Gaussian invoking a mean field variational inference as in Hinton & Van Camp (1993):

$$q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) = \prod_{i=1}^{I}\prod_{j=1}^{O} q_{\mu_{ij}\sigma_{ij}}(\omega_{ij}) = \prod_{ij} N(\omega_{ij}; \mu_{ij}, \sigma_{ij}^2), \tag{2}$$

where $I$ is the number of input channel and $O$ number of output channel (classes). Following (Blundell et al., 2015), we reparameterize $\boldsymbol{\omega}$ by $\omega_{ij} = \mu_{ij} + \sigma_{ij}\epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, 1)$. The standard deviation $\boldsymbol{\sigma}$ is further reparameterized by $\rho$ through a softplus function $\sigma_{ij} = \log(1 + \exp(\rho_{ij}))$ to ensure $\sigma$ is always non-negative.

The local reparameterization trick (Kingma et al., 2015) is utilized for statistically efficient gradient estimation. Instead of sampling $\boldsymbol{\omega}$, we directly sample the random layer activation. Then the training procedure is described as follows:

1. Sample $\epsilon_{ij} \sim N(0, 1)$.
2. Calculate $y_j = \sum_{i=1}^{I} z_i^{(lk)}\mu_{ij} + \epsilon_{ij}\sqrt{{z_i^{(lk)}}^2\sigma_{ij}}$, where $z_i^{(lk)}$ is the output from the unet feature extractor corresponding to $(lk)^{th}$ pixel.
3. Calculate MC estimates of the ELBO according to (1).
4. The deterministic neural network parameters and the variational parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ are updated through gradient ascent.

When the training is done, the predictive distribution is achieved by MC approximation as follows:

$$\hat{p}(y^*|\mathbf{x}^*, \mathbf{X}, Y) \approx \frac{1}{T}\sum_{t=1}^{T} p(y^*|\mathbf{x}^*, \boldsymbol{\omega}_t), \;\; \boldsymbol{\omega}_t \sim q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) \tag{3}$$

#### 3.1.2. MC dropout

Gal & Ghahramani (2016) interprets the dropout training (Hinton et al., 2012; Srivastava et al., 2014) as a variational inference which approximates the posterior distribution by placing Bernoulli variational distribution in the Bayesian neural networks. Dropout training is a commonly used regularization technique that prevents neural networks from overfitting and co-adaption of features in practice. It is done by randomly removing units within the neural networks during training. Gal & Ghahramani (2016) relates dropout with

variational inference by defining the variational distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ as a Bernoulli distribution for each layer in the network, and shows that the obtained model is the approximation of Gaussian processes (Gal & Ghahramani, 2015). The KL divergence between $q_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ and $p(\boldsymbol{\omega})$ $\mathrm{KL}[q_{\boldsymbol{\theta}}(\boldsymbol{\omega})||p(\boldsymbol{\omega})]$ is minimized using stochastic gradient descent when train the network with cross entropy loss (Gal & Ghahramani, 2015). At test time, model parameters are sampled from the trained posterior distribution using dropout and the predictive distribution is achieved by the MC approximation as in equation 3.

### 3.2. Deep Ensembles (Lakshminarayanan et al., 2016)

In theory, Bayesian neural networks can capture uncertainty by learning a posterior distribution over the parameters of the network and exploring the space of solutions. However, it has been reported that they often fail to explore the entire weight space and capture the network uncertainty within a single-mode (Fort et al., 2019). Especially, Fort et al. (2019) empirically shows that subspace sampling methods along with a single training trajectory exhibit high functional similarity, and the disagreement of predictions from the sampled functions is low.

Instead, deep neural networks initialized at a different random point tend to end up at different modes in function space (Fort et al., 2019; Lakshminarayanan et al., 2016). Fort et al. (2019) shows that an ensemble of them that have different local minimums provides relatively larger benefits than subspace sampling methods such as MC dropout in terms of accuracy uncertainty in their experiment setting. We, therefore, construct an ensemble network that compensates for the insufficiency of a diverse set of predictions of the scalable variational inference methods by training the deterministic segmentation network multiple times with random initialization. The predictive distribution of an ensemble network then becomes

$$\hat{p}(y^*|\mathbf{x}^*, \mathbf{X}, Y) \approx \frac{1}{J}\sum_{j=1}^{J} p(y^*|\mathbf{x}^*, \hat{\boldsymbol{\omega}_j}) \tag{4}$$

where $\hat{\boldsymbol{\omega}_j}$, $j = 1, \ldots, J$ are $J$ independent model parameters trained from random initialization with different random seeds.

## 4. Experiments and Results

### 4.1. Dataset

We use three datasets where images have multiple annotations from multiple clinicians. From QUBIQ 2021 challenge, the Prostate and Brain-Growth datasets were used, which have 7 and 6 inter-observer annotations respectively. The prostate dataset has two tasks that segment different regions of interest. We use single 2D MRI images size of $256 \times 256$ pixels. All pixels are normalized to [0, 1]. The annotated mask has binary labels: region of interest is 1, and background is 0. Sample brain-growth images with annotated masks are presented in figure 1. Inter-observer annotations are exists both in the train and test sets. The detailed training strategy is discussed in section 4.3. Brain-growth dataset consists of 34 training set and 5 validation set, and prostate datset consists of 48 training set and 7 validation set. We also use intravascular ultrasound (IVUS) datasets to demonstrate the applicability of predicting clinical indicators with ranges. The dataset consists of two sub-datasets, datasets A and B, which are publicly available (Balocco et al., 2014). Dataset A consists of 77 images with 19 for training and 58 for testing with a 40 MHz catheter. Dataset B consists of 425 images, with 109 for training and 326 for testing with a 20 MHz catheter. Unlike QUBIQ 2021 challenge dataset, the IVUS datset consists of 3 inter-intra observer annotations only in the test set.

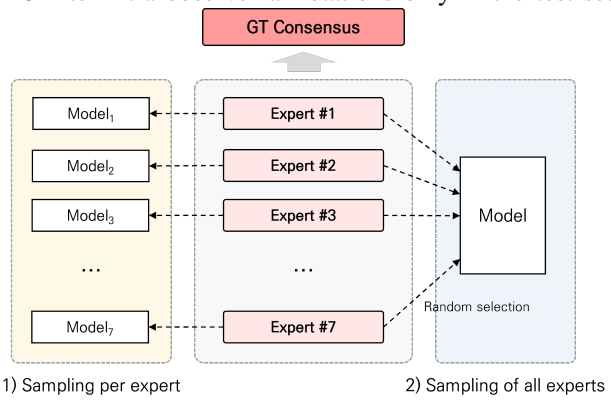



Figure 4: A schematic diagram of model training procedures using multiple annotated datasets from multiple experts.

### 4.2. Implementation Details

As described in 3.1.1, motivated by Riquelme et al. (2018), we construct a neural linear segmentation network, performing Bayesian inference only in a small part of the network. The neural linear Unet comprises two main components: feature extractor and Bayesian convolution layer. From the Unet feature extractor, 64 feature maps are produced from each pixel of an image. The produced feature maps are fed into the stochastic $1 \times 1$ convolution layer. The Bayesian layer will capture the uncertainty by aggregating the information over 64 feature maps. The mean-field approximation is used for the variational posterior $q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) = N(\omega_{ij}; \mu_{ij}, \sigma_{ij} = \log(1 + \exp(\rho_{ij})))$, where the variational parameters are initialized by $\mu_{ij} \sim N(0, 0.05)$, $\rho_{ij} \sim N(-4, 0.05)$. This initialization is from (Pinsler et al., 2019). We construct MC dropout Unet by allocating the dropout layer in the same Unet layer location as the Kendall et al. (2015). The only difference is that Kendall et al. (2015) doesn't have a skipconnection between encoder

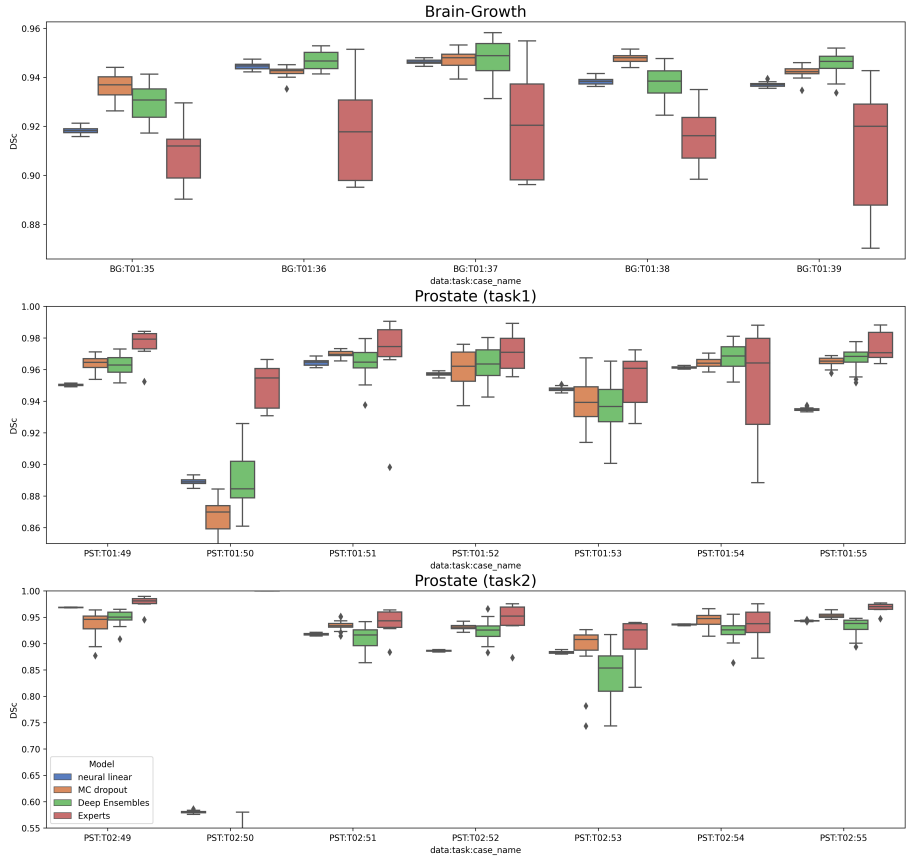



Figure 3: Corresponding dice coefficient score between clinicians' consensus and sampled segmentation, but also the consensus and individual clinicians' annotation.

Table 1: Generalized energy distance metric $D^2$ evaluations on MRI dataset.

| Data sampling Strategy | Methods | Prostate (task1) | Prostate (task2) | Brain-Growth (task1) | average |
|---|---|---|---|---|---|
| All experts | MC dropout (Kendall et al., 2015) | 0.117 ±0.07 | 0.305± 0.44 | 0.145 ±0.012 | 0.194 ± 0.287 |
| | Neural Linear (Riquelme et al., 2018) | 0.134 ± 0.05 | 0.313 ± 0.33 | 0.114 ± 0.02 | 0.194 ± 0.23 |
| | Deep Ensembles (Fort et al., 2019) | **0.079 ±0.05** | **0.245 ± 0.36** | **0.104±0.01** | **0.147 ± 0.24** |
| per expert model | MC dropout (Kendall et al., 2015) | 0.122 ± 0.07 | 0.34 ±0.42 | 0.203 ± 0.04 | 0.223 ± 0.27 |
| | Neural Linear (Riquelme et al., 2018) | 0.187 ± 0.09 | 0.398 ± 0.323 | 0.196 ± 0.05 | 0.264 ± 0.224 |
| | Deep Ensembles (Fort et al., 2019) | **0.1 ± 0.05** | **0.271 ± 0.32** | **0.129 ± 0.02** | **0.169 ± 0.2** |

and decoder. We use 0.5 dropout rate both in training and test phase. For training, we use Adam optimizer (Kingma & Ba, 2014) with a learning rate 0.001 and cosine annealing is applied. The network is trained for 3000 epochs with mini-batch size 8 and early stopped at the best validation dice similarity coefficient (DSC). No data augmentation is used. All architectures share consistent choices of other hyper-parameters that we explored. 20 MC samples and ensemble members are used throughout the experiments.

### 4.3. Data Sampling Strategy

In this paper, we adopted two strategies for training models using datasets independently annotated by multiple experts: 1) Sampling per expert; 2) Sampling of all experts. First, we build the same number of models as experts, and each model learns only the annotations of one of the multiple experts. After all, this is the same as building each expert-specific model. Second, only one model is built, and the model learns one randomly selected from multiple annotations on the same data during training. Additionally, we combine multiple annotations to make an annotation consensus. Figure 4.1 shows a visual summary of these processes.

### 4.4. Results Analysis

Our analysis focuses on two aspects. First, we validate the methods for approximating clinicians' variability by

evaluating the similarity of clinicians' inter annotations and sampled masks from the predictive distribution using the QUBIQ2021 challenege dataset. In specific, we assess how similar the variances of the calculated dice coefficients are from the learned predictive distributions to the clinicians' inter variability.

Second, we evaluate a clinical endpoint of the learned posterior (uncertainty) over the segmentation using the IVUS dataset (Balocco et al., 2014). We assess whether Bayesian approaches can approximate expert variability in the clinical measurement space. For a clinical endpoint, Lumen, EEM and plaque burden are calculated from the IVUS segmentation. We analyze whether the variance (clinical uncertainty) of the clinical values obtained through the Bayesian approaches and experts is consistent.

#### 4.4.1. Uncertainty Evaluation Metrics

**Generalized Energy Distance Metric** We aim to analyze if the approximate Bayesian deep learning with variational inference and deep ensembles can learn inter-rater variability in segmentation. To quantitatively assess the reproducibility of the rater variability of the method both in accuracy and diversity, we exploit the generalized energy distance metric (Kohl et al., 2018; Hu et al., 2019) $D^2(p_{\text{gt}}, p_{\text{out}}) = 2\mathbb{E}[d(S,Y)] - \mathbb{E}[d(S,S')] - \mathbb{E}[d(Y,Y')]$ where $d$ is an intersection over union (IoU), $Y$ and $Y'$ are ground truth sampled from clinicians, and $S$ and $S'$ are samples from the predictive distribution $\hat{p}$. Note that the first term quantify accuracy and the rest of the term quantify diversity.

The average generalized energy distance $D^2$ of the presented methods (i.e., neural linear (Riquelme et al., 2018), MC dropout (Kendall et al., 2015) and Deep ensembles (Fort et al., 2019) Unet) are described in table 1. Noticeably, in all tasks, deep ensemble had an overwhelmingly lower GED score and better performance than all other methods such as MC dropout and neural linear. The table 1 represents that tendency of deep ensembles to explore diverse modes in function space leads to diversity also in the segmentation space. Neither the neural linear nor the GED scores of mc dropout were particularly high or low for all three tasks. This result agrees with the argument of Fort et al. (2019) in that the different modes of the function space provide a greater advantage over the subspace sampling method. Although this is consistent regardless of how the data is sampled from inter-rater annotations, it can be seen that the general performance of all experts model is higher than that of per expert model. This may be because the random selection of various inter-rater annotations for each image acts as a regularizer from overfitting.

#### 4.4.2. Does methods learn experts' variability?

The presented method is capable of providing multiple segmentation results. We compared the presented models' variation with the posterior over the segmentation on three different medical datasets. In figure 3 we plot the corresponding DSc score between clinicians' consensus and sampled segmentation, but also individual clinician's annotation and the consensus.

From figure 3 it can be seen that the standard deviation of DSc between inter-raters is quite large, ranging from 0.01 to 0.04. This can be viewed as a meaningful estimate of the aleatoric uncertainty, the irreducible noise found in the data (Hu et al., 2019). The behavior of learned posteriors over the segmentation agrees with the intuition from (Fort et al., 2019). Even though neural linear and MC dropout Unet obtain reasonable variability of Dice coefficient scores, they lack diversity. However, deep ensembles successfully approximate expert variability in all tasks. Not only is the variance of DSc between the clinician's agreement and the sample split close to the expert's variance, but the mean DSc score is also close enough to the expert's mean. This can also be confirmed in Figure 1. Figure 1 is the visualization of the learned variability over the brain-growth image. Although MC dropout and neural linear approximate expert variability well enough, they are not as accurate and diverse as deep ensemble. From this, we conclude that deep ensembles perform best with respect to learning inter-rater variability in segmentation.

#### 4.4.3. Uncertainty in the Clinical Measurement Space

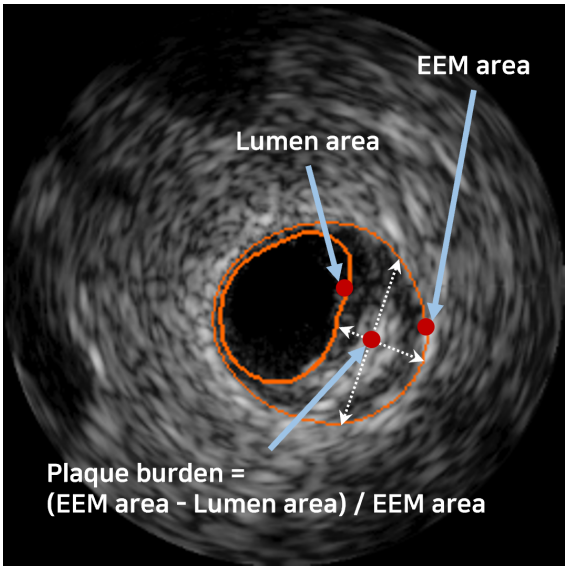



Figure 7: Visualization of measuring the cross-sectional plaque burden in IVUS imaging.

In medical images, segmentation work does not end simply

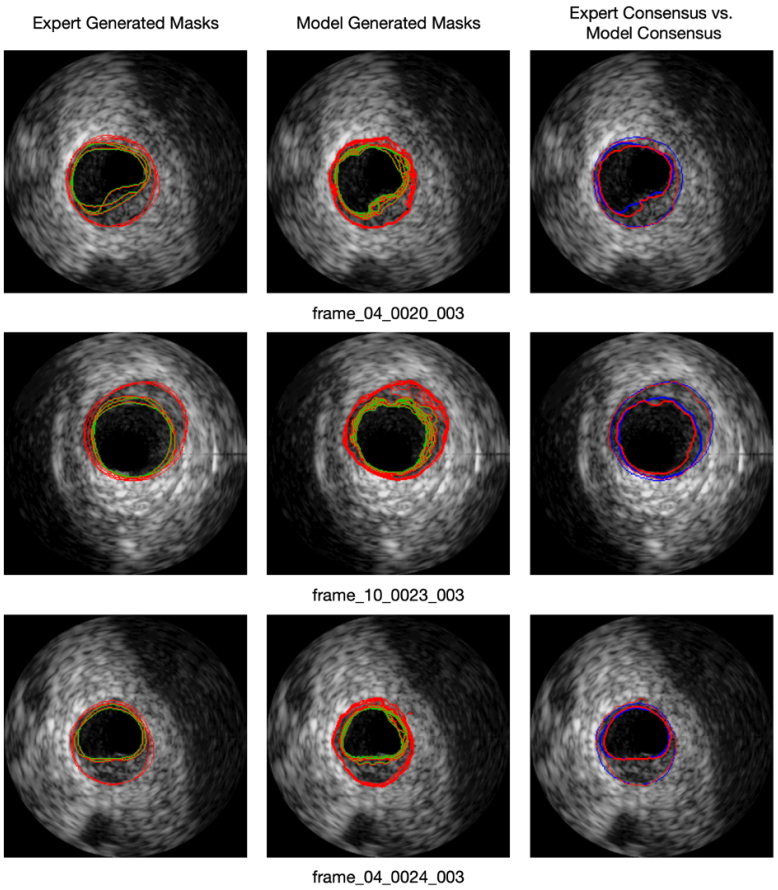



Figure 5: Visualization of the learned variability in segmentation over the IVUS dataset as well as inter-intra rater variability and their consensus.

Table 2: The uncertainty of (unscaled) clinical measurements for a single IVUS image in both experts and Bayesian models.

| | patient 04 (frame #20) | | | patient 10 (frame #23) | | | patient 04 (frame #24) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Lumen | EEM | Plaque burden | Lumen | EEM | Plaque burden | Lumen | EEM | Plaque burden |
| Experts | 4081 $\pm$ 510 | 7138 $\pm$224 | 0.43 $\pm$ 0.06 | 5669 $\pm$202 | 8675 $\pm$302 | 0.346 $\pm$ 0.01 | 3190 $\pm$205 | 5829 $\pm$153 | 0.45 $\pm$0.02 |
| Deep Ensembles | 4034 $\pm$232 | 6212 $\pm$230 | 0.39 $\pm$0.04 | 4772 $\pm$338 | 7671 $\pm$ 366 | 0.40 $\pm$0.04 | 2923 $\pm$166 | 5370 $\pm$301 | 0.47 $\pm$0.03 |

by itself. Its endpoint is to derive clinical measurements for disease diagnosis and decision-making, which is very important in the medical field. Therefore, to calculate the uncertainty of clinical measurements in the process, we propose to utilize multiple segmentations sampled from the posterior probability distribution, which is learned during training for the segmentation task.

In this section, we exploit the deep ensemble, best performed in previous experiments, to quantize uncertainty in clinical measurements and evaluate how effective it is. As the dataset for this experiment, we selected the IVUS dataset (Balocco et al., 2014), which is widely used as the gold standard for quantifying coronary artery stenosis. Coronary artery stenosis is quantified by measuring the cross-sectional plaque burden (McDaniel et al., 2011). The plaque burden is calculated by dividing the plaque area by the external elastic membrane (EEM) area as follows.

$$\text{Plaque burden} = \frac{(\text{EEM} - \text{Lumen area})}{\text{EEM}} \times 100 \qquad (5)$$

We first visualize the learned variability in segmentation over the IVUS dataset in figure 5. As with the previous MRI dataset, learned posterior from deep ensembles successfully approximates the inter-variability in segmentation. In particular, from the patient 04 (frame #20) in the first row, it can be seen that the variability of the depression of the lumen area is described as much as the inter-intra rater's variability. Furthermore, the variability of the boundary of the EEM area is also well approximated in all three cases. Figure 2

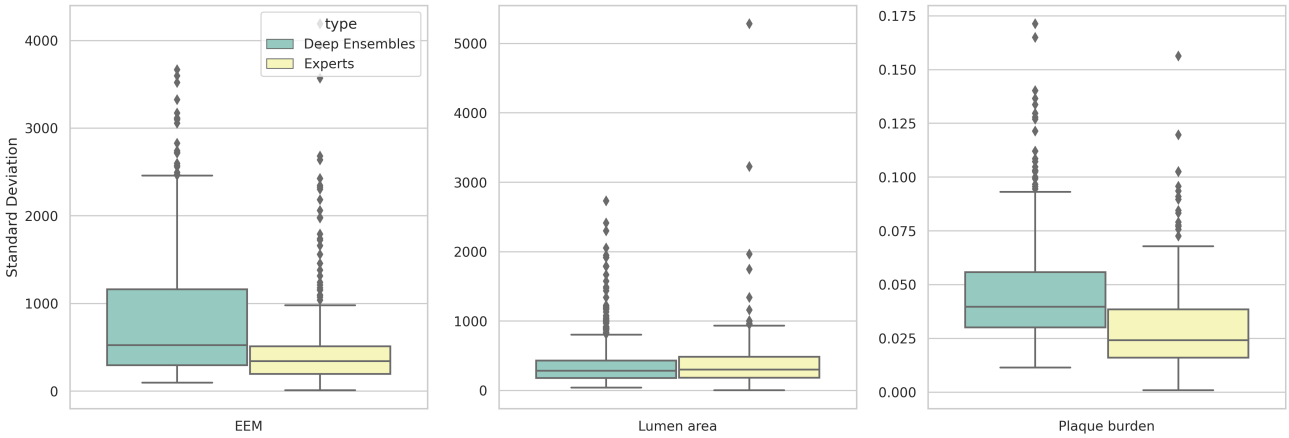



Figure 6: Distributions of the standard deviation of the calculated measurements (EEM, lumen area, plaque burden) in a single IVUS image for all test dataset.

represents the uncertainty of (unscaled) clinical measurements for a single IVUS image in both experts and Bayesian models. Clinical indices such as EEM, lumen and plaque burden are described with error range by standard deviation. Although the estimate of the mean of each measurement is slightly different in some cases, an estimate of uncertainty by standard deviation is reasonable for comparison with an expert's standard deviation. This information could be valuable tools in supporting clinicians' decision-making.

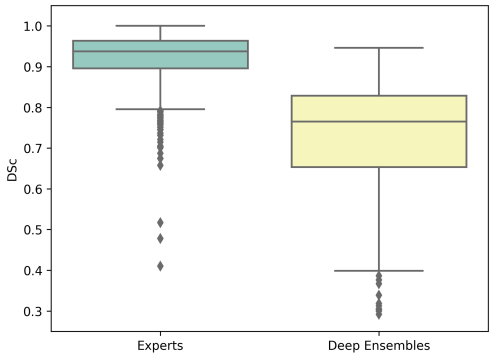



Figure 8: Distributions of dice coefficient between expert and expert consensus as well as trained model and expert consensus in IVUS test dataset.

We assess how well deep ensembles approximate the inter-intra variance (uncertainty) in the clinical measurements with figure 6. It presents the distribution of the standard deviation of the calculated measurements (EEM, lumen area, plaque burden) in a single IVUS image for all test datasets. The mean and variance of the standard deviation of the predicted lumen areas within an image agree well with the experts' calculated values. The statistic for the EEM and plaque burden agree with the experts to some extent, but overall, the standard deviation is over-estimated. A clue to the overestimation can be found in the figure 8, which shows distributions of dice coefficient between expert and expert consensus as well as the trained model and expert consensus. The DSc mean of experts is close to 0.93, whereas the DSc mean of the deep ensembles is close to 0.79. This indicates that the model may be less trained for several reasons, leaving room for further performance gains with additional data. Therefore, we can hypothesize that the obtained predictive distributions lead to larger variances than the clinician's inter-intra measurements because of the combination of epistemic and aleatoric uncertainty. We leave it for future research to test this hypothesis and to analyze the results and performance of separating aleatoric and epistemic uncertainty for modeling the rater variability over the clinical measurements.

## 5. Conclusion

We analyze if Bayesian predictive distribution learned from various approximate inference schemes can learn intra-rater variability both in segmentation and clinical measurements. We do not propose new methods, but rather to demonstrate that Bayesian neural networks are able to reproduce the rater variability with the posterior over the segmentations in four medical imaging tasks. Especially, the IVUS dataset is used for evaluating a clinical endpoint. Our result suggests that Bayesian approaches can approximate expert variability. For future research, we will analyze factors influencing uncertainty estimation performance in segmentation and clinical measurements. In specific, we will study how the epistemic and aleatoric uncertainties computed and separated by Bayesian neural networks in segmentation affect the approximation of uncertainty inherent in expert clinical measurements.

## References


Balocco, S., Gatta, C., Ciompi, F., Wahle, A., Radeva, P., Carlier, S., Unal, G., Sanidas, E., Mauri, J., Carillo, X., et al. Standardized evaluation methodology and reference database for evaluating ivus image segmentation. *Computerized medical imaging and graphics*, 38(2):70–90, 2014.

Baumgartner, C. F., Tezcan, K. C., Chaitanya, K., Hötker, A. M., Muehlematter, U. J., Schawkat, K., Becker, A. S., Donati, O., and Konukoglu, E. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 119–127. Springer, 2019.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

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

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Hinton, G. E. and Van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13, 1993.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., and Welling, M. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 137–145. Springer, 2019.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Kendall, A., Badrinarayanan, V., and Cipolla, R. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, pp. 2575–2583, 2015.

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., and Ronneberger, O. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

McDaniel, M. C., Eshtehardi, P., Sawaya, F. J., Douglas, J. S., and Samady, H. Contemporary clinical applications of coronary intravascular ultrasound. *JACC: Cardiovascular Interventions*, 4(11):1155–1167, 2011.

Pinsler, R., Gordon, J., Nalisnick, E., and Hernández-Lobato, J. M. Bayesian batch active learning as sparse subset approximation. In *Advances in Neural Information Processing Systems*, pp. 6359–6370, 2019.

Riquelme, C., Tucker, G., and Snoek, J. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.