Probabilistic Modeling of Inter- and Intra-observer Variability in Medical Image Segmentation

Arne Schmidt, Pablo Morales-Álvarez, and Rafael Molina
Universidad de Granada, Granada, Spain
{arne, pablomorales, rms}@decsai.ugr.es

Abstract

Medical image segmentation is a challenging task, particularly due to inter- and intra-observer variability, even between medical experts. In this paper, we propose a novel model, called Probabilistic Inter-Observer and iNtra-Observer variation NetwOrk (Pionono). It captures the labeling behavior of each rater with a multidimensional probability distribution and integrates this information with the feature maps of the image to produce probabilistic segmentation predictions. The model is optimized by variational inference and can be trained end-to-end. It outperforms state-of-the-art models such as STAPLE, Probabilistic U-Net, and models based on confusion matrices. Additionally, Pionono predicts multiple coherent segmentation maps that mimic the rater's expert opinion, which provides additional valuable information for the diagnostic process. Experiments on real-world cancer segmentation datasets demonstrate the high accuracy and efficiency of Pionono, making it a powerful tool for medical image analysis.

1. Introduction

Artificial Intelligence (AI) algorithms have shown remarkable progress in image analysis, holding great promise for faster and more accurate diagnostic procedures [13, 27, 26, 3]. Nevertheless, in medical practice, there exists a high degree of variability among the opinions of different medical experts, even when the same expert assesses the same data at different times. This inter- and intra-observer variability has been reported across various tasks, including MRI-based segmentation of HCC lesions [7], lung cancer segmentation in CT scans [14], and multiple fields in pathology [21, 2, 5, 22]. It leads to uncertainties when applying AI models because in contrast to other classification tasks, there is not a single ground truth.

Especially in the medical domain, the careful modeling of uncertainties in its different forms has a high priority to minimize the risk of relying on incorrect predictions

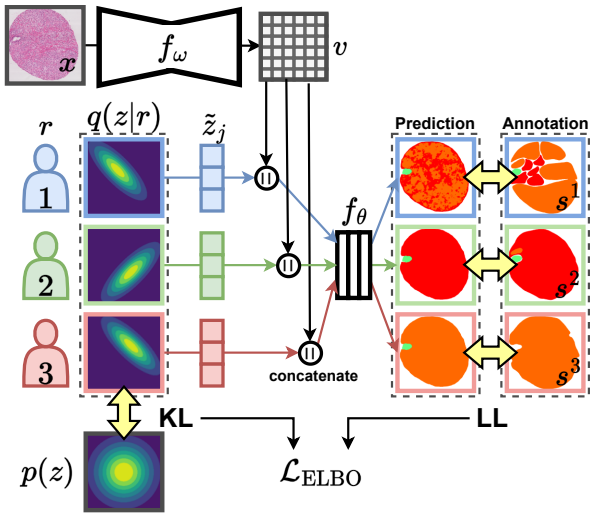


Figure 1. The proposed Pionono model. The labeling behavior of each rater r is represented by a multivariate Gaussian distribution $q(z|r)$. The drawn samples $\tilde{z}_j$ are concatenated with the extracted features v of f_ω and then fed into the segmentation head f_θ. The output simulates the inter- and intra-observer variability of annotations and is optimized using the real annotations s^r of each rater. The model is trained end-to-end with a combination of log-likelihood loss (LL) and Kulback Leibler (KL) divergence between posterior and prior, combined in the overall loss $\mathcal{L}_{\text{ELBO}}$.

[17, 26, 15, 13, 10]. In recent years, probabilistic methods, such as Bayesian Neural Networks [8] and sparse Gaussian processes [10, 33] have gained more and more attention, because they are able to account for uncertainties in a sound manner. They showed promising results when modeling uncertainty in the network weights [8], data ambiguities[15] or attention weights [26]. Although inter- and intra-observer variability is often mentioned as a key challenge when applying AI to medical data [21, 16, 28, 18], to the best of our knowledge there is no method that explicitly models these two types of uncertainty for medical image segmentation.

To address this gap, we propose a novel approach called the Probabilistic Inter-Observer and iNtra-Observer varia-

tion NetwOrk (Pionono), depicted in Figure 1. This model accurately accounts for inter- and intra-observer variability using probabilistic deep learning. Specifically, each rater's labeling behavior is represented as a probability distribution in latent space, and optimized using the log Evidence Lower BOund (ELBO) in an end-to-end training process. The variance of each rater's distribution models the intra-observer variability, while the differences between the distributions models the inter-observer variability. When two raters exhibit similar labeling behavior, their probability distributions overlap substantially, while different labeling behavior results in a small overlap of distributions.

The approach is validated in extensive experiments of prostate and breast cancer segmentation, using 'gold' labels. They reflect the expert agreement to show that our probabilistic modeling improves the predictive performance and estimates the predictive uncertainty. Furthermore, we also test its capability to model each rater's labeling behavior. As shown in the experiments, it can simulate expert opinions for a given test image in a consistent manner, providing a realistic estimation of "what expert X would say in this case". Our contributions can be summarized as follows:

- We propose Pionono, a probabilistic deep learning model that uses probability distributions in latent space to represent inter- and intra-observer variability. It can be trained with labels of multiple raters.
- The model is able to provide accurate segmentation predictions (compared to the expert agreement and different expert opinions), outperforming existing state-of-the-art algorithms such as STAPLE, Probabilistic U-Net and models based on global or local confusion matrices.
- Pionono provides uncertainty estimations that indicate areas where the predictions are not conclusive.
- The proposed model can provide several coherent segmentation hypotheses, simulating different medical experts.

2. Related Work

In this section, we review existing methods of probabilistic deep learning and crowdsourcing for medical images and highlight the differences to our model.

Probabilistic Deep Learning. As already indicated, probabilistic approaches such as Bayesian neural networks [8, 1, 15, 17] and sparse Gaussian processes [10, 33, 26] have shown promising results in a multitude of tasks in the medical image domain, modeling different sources of uncertainties. Often, a general predictive uncertainty is addressed using probabilistic weight parameters [1]. This uncertainty can be bisected into model and data uncertainty which originate from model parameters or data ambiguities, respectively [15]. Other approaches have modeled the uncertainty of missing instance labels in multiple instance learning [20, 26] or uncertainty of out of distribution samples [17]. The uncertainty in annotations has previously been addressed by the Probabilistic U-Net [13] (*Prob U-Net*), which encodes the labeling behavior in a latent random variable. The model is trained as a variational autoencoder with an encoder network predicting the latent distribution. This approach models a general variability in annotations but lacks the explicit modeling of inter- and intra-observer variability. Therefore, it is not able to incorporate the rater information during training and cannot simulate expert opinions.

Method	(i) Prob. Uncert.	(ii) Coh. Segm.	(iii) Exp. Opinion	(iv) Scale
STAPLE	✗	✗	✗	✓
Prob U-Net	✓	✓	✗	✓
CM global	✗	✗	✓	✓
CM pixel	✗	✗	✓	✗
Pionono	✓	✓	✓	✓

Table 1. For AI segmentation models to achieve the best possible diagnostic support, they should address four key issues: (i) provide a *probabilistic uncertainty* estimation, not only a single prediction for a test image; (ii) provide multiple *coherent segmentation* hypotheses; (iii) simulate different *expert opinions* for better explainability and decision support; (iv) *scale* to a higher amount of raters in the case that more data from different hospitals can be integrated.

Crowdsourcing. While existing crowdsourcing methods aim to capture inter-observer variability in the training labels, this variability is often not reflected in the test predictions by probabilistic outputs [13]. The intra-observer variability is often not modeled at all, although it is often mentioned as a challenge in literature [21, 16, 28, 2, 7].

One way to handle multiple annotations is label fusion. With this method, the annotations of different raters are merged to a single set of labels. The "Simultaneous Truth and Performance Level Estimation" (*STAPLE*) mechanism performs label fusion with a probabilistic estimate of the true labels by weighting each segmentation depending upon the estimated performance level of each rater [32]. A supervised network can then be trained on these fused labels. More dedicated approaches incorporate different rater labels using confusion matrices (CM), for example for classification of image patches with Gaussian processes [18] or image segmentation with global confusion matrices [31] (*CM global*). In this direction, also pixel-wise confusion matrices were explored for semantic segmentation that are estimated by a dedicated deep neural network [34] (*CM pixel*). These models have shown promising results, but come with a conceptual problem: They assume that the pixels are statistically independent of each other, although

neighboring pixels have a high correlation. Therefore, the output of different segmentation hypotheses is not coherent. Furthermore, the predictions of the mentioned approaches [31, 34] are not modeled by a predictive distribution, but by a deterministic point estimate. While the global confusion matrix approach [31] has a limited expressiveness, the pixel-wise calculation [34] is hard to scale for multiple raters, because for each rater, a complete deep neural network must be trained and stored.

Pionono unites the advantages of probabilistic and crowdsourcing methods. We summarize a comparison of different characteristics of Pionono and related methods for image segmentation in Table 1.

3. Methods

In this section, we outline the background of the proposed method. It is implemented in the Pytorch [23] framework and is publicly available at https://github.com/arneschmidt/pionono_segmentation.

3.1. Problem Definition

Let $X = \{x_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1,..,N}$ be a set of images, $S^r = \{s_i^r \in \mathbb{R}^{H\times W\times C}\}_{i=1,..,N}$ the corresponding segmentation maps with image dimensions $H \times W$ and C the number of classes. The segmentation maps are provided by different raters $r \in R = \{1, 2, .., M\}$. Some or all images can be segmented by multiple raters, such that some segmentation maps s_i^r can be empty. The proposed model does not require any overlap of the sets of annotated images.

If there are images available with segmentations assigned by expert agreements (so-called gold labels), the model should be able to predict a gold distribution over outputs $p(S^{\text{gold}})$ with the mean estimating the segmentation and the variance estimating the uncertainty. In any case, the model should model different segmentation hypotheses for the raters $\{p(S^r); r = 1, 2, .., M\}$ for diagnostic decision support.

3.2. Proposed Model

First, we introduce the common segmentation backbone f_ω with trainable weights ω. We use the well-known U-Net architecture [25] with a Resnet34 feature extractor [9]. This model takes an image x_i and extracts a feature map $v_i \in \mathbb{R}^{H\times W\times L}$ with $H \times W$ being the image resolution and L the dimensions of the feature vectors ($L = 16$ in the case of U-Net). We denote the feature extraction as

$$v_i = f_\omega(x_i). \tag{1}$$

Based on these feature vectors, we could perform segmentation with a segmentation head f_θ:

$$s_i = f_\theta(v_i). \tag{2}$$

Now we extend this model to incorporate the *inter- and intra-observer variation*. The segmentation maps are influenced by the rater's experience, assessment, and personal choices. To encode the labeling behavior, we use a random vector $z \in \mathbb{R}^D$. In practice, $D = 8$ are enough dimensions to reflect different labeling behaviors. We define a prior distribution $p(z) = \mathcal{N}(z|0, \sigma_{\text{prior}} * I)$ which encodes a generic labeling behavior without further information about the rater. It is possible to encode prior knowledge in this distribution, but we present a general model and leave this for future work. We set $\sigma^2_{\text{prior}} = 2.0$, because we observe a realistic variability in the output for this value. In section 4.5 we prove that the model is robust for different settings of hyperparameters D and σ^2_{prior}.

Now, the posterior distribution of $p(z|r)$ that depends on the rater r should be found. We approximate it with one multivariate Gaussian distribution for each rater:

$$q(z|r) = \mathcal{N}(z|\mu^r, \Sigma^r)\ \forall r = 1, .., M \tag{3}$$

where $\{\mu^r, \Sigma^r\}^{r=1,..,M}$ are trainable parameters. The variance of each distribution $q(z|r)$ models the $intra - observer$ variability. The differences between the distributions for different raters model the $inter - observer$ variability. To obtain the predictive gold distribution we add another 'rater' $r = M + 1$ represented by an additional gold distribution q which is trained with the available gold segmentations. During prediction, this distribution provides the estimated agreement between experts.

The segmentation head f_θ, parametrized by weights θ, must be adapted to take the random vector z into account. The approximated predictive distribution is then obtained by:

$$q(s_i|x_i, r, \omega, \theta) = \int f_\theta(v_i, z)q(z|r)dz. \tag{4}$$

The closed-form calculation is not feasible and therefore we approximate it by Monte Carlo (MC) sampling:

$$\tilde{s}_{i,j}|x_i, r = f_\theta(v_i, \tilde{z_j});\ \tilde{z_j} \sim q(z|r) \tag{5}$$

with $j = 1, ..., K$ indexing the MC samples. In practice, we concatenate the feature maps v_i and the latent vector $\tilde{z_j}$, which is broadcasted to the image size, leading to a feature map with dimensions $H \times W \times (L + D)$. The segmentation head consists of three layers with 1x1 convolutions and 16 filters in the first two layers and C filters in the last layer.

3.3. Training

First, all posterior distributions $q(z|r)$ are initialized randomly. Each initial value of the mean vectors μ^r is independently drawn from a distribution $\mathcal{N}(0, \sigma^2_{\text{post}})$. We set $\sigma^2_{\text{post}} = 8$, because this initializes the mean vectors sufficiently different for a good optimization. In section 4.5 we

show, that the model is robust to other settings of this value. The covariance matrices Σ^r are initialized with $\sigma_{\text{prior}} * I$.

To optimize the parameters $\{\mu^r, \Sigma^r\}^{r=1,..,M}$ of the probability distribution $q(z|r)$, we maximize the ELBO:

$$\mathcal{L}_{ELBO} = \mathbb{E}_q log\ p(S^r|X, r, \omega, \theta) - \lambda KL(q(Z|r)|p(Z)). \tag{6}$$

with distribution q as defined in eq. 4. The first term defines a log-likelihood (LL) loss, making the model fit to the annotations of each rater. The second term defines the KL-divergence between the posterior distribution $q(Z|r)$ and the prior $p(Z)$ and works as a regularization of the latent distributions. The factor λ weights the regularization term and is set to 0.0005 to balance the magnitudes of the log-likelihood and the KL (we will check the robustness of this hyperparameter in Section 4.5). While the KL term can be optimized analytically, the log likelihood term must be approximated. We use the reparametrization trick [12] to split each probabilistic sample $\tilde{z}^r$ into its probabilistic component and deterministic parameters μ^r and Σ^r. These parameters can be optimized by backpropagation of gradients, together with the CNN parameters ω and θ. For numerical stability, we train the covariance matrix parameters by using the lower triangular matrix L of the Cholesky decomposition $\Sigma^r = L^r L^{r\top}$. The log-likelihood can be optimized with standard methods like the categorical cross-entropy. We found that the general dice loss [30] leads to better results, so all final results are reported with this loss.

We use the Adam optimizer [11] for 100 epochs with a learning rate of 0.0001. The model parameters μ^r, Σ^r are optimized with a higher learning rate of $\nu = 0.02$, because else the gradient was not strong enough to properly learn the rater distributions. We tested $\nu = 0.01, 0.02, 0.04$ and include the results in section 4.5. Both learning rates are decreased after 40 epochs by dividing them by 1.1 in each epoch.

3.4. Predicting

For a test image x^*, the predictive gold distribution can again be obtained by drawing Monte-Carlo samples

$$\tilde{s}_j^*|x^*, r = f_\theta(v^*, \tilde{z}_j); \tilde{z}_j \sim q(z|r = M + 1) \tag{7}$$

with $j = 1, .., K$ indexing the MC samples and $q(z|r = M + 1)$ representing the gold distribution as described in section 3.2. The **mean** of these samples provides the segmentation hypothesis that approximates the expert agreements. The **variance** of the samples indicates uncertainties in the prediction.

Furthermore, the model is able to simulate *intra-observer variations* of rater r' by drawing multiple samples of the distribution $\tilde{z}'_j \sim q(z|r = r')$ for the final prediction. The *inter-observer variations* between rater r' and r'' can be simulated by using samples $\tilde{z}'_j \sim q(z|r = r')$ and $\tilde{z}''_k \sim q(z|r = r'')$ and finally taking the mean of both output distributions.

The model can therefore simulate **expert opinions** for a given test image. Other AI methods typically aggregate the expertise provided by all annotators to make predictions (e.g., using STAPLE, Prob U-Net). However, in such approaches, the knowledge of highly specialized experts can be diluted or lost among the less experienced annotators' knowledge. In our framework, we provide consistent predictions for each individual expert, thereby preserving their unique expertise and contributions.

4. Experiments

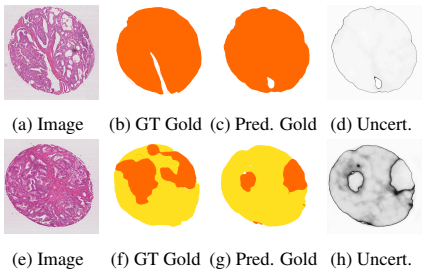
In several experiments we demonstrate that the uncertainty estimation of the model indicates areas of false predictions (4.2), the model is able to capture the inter and intra-observer variations (4.3) and outperforms other related methods (4.4). Additionally, we analyze the robustness to hyperparameters (4.5), required resources (4.6), and limitations (4.7).

4.1. Datasets

For empirical validation, three public histopathological datasets were used. The first dataset, "Gleason 2019" [22] was published as a MICCAI grand challenge for pathology and includes 333 Tissue Micro Arrays (TMA) of prostate cancer, labeled by 6 different pathologists. The TMAs were scanned with a magnification of 40x and have a size of approximately 4000×4000 pixels. Of the 333 images, 244 are publicly available with labels (the test annotations of the challenge are not available). Each pathologist annotated between 61 and 241 TMAs with segmentation masks and the gold labels were obtained using the STAPLE algorithm [32], following the original work of the dataset [22]. We resize all images to 1024×1024 pixels and create 4 cross-validation splits.

The second dataset, which we will refer to as "Arvaniti TMA" was published in 2018 [4] and includes a total of 886 TMAs of prostate cancer of which 245 images were annotated by two pathologists (while the other images only have annotations of one pathologist and are therefore discarded in our study). The TMAs were scanned with a magnification of 40x but the scanned area is smaller than for the Gleason19 dataset. The images have a resolution of 3100×3100 pixels and we resize them to 512×512 such that the magnification matches the resized images of the Gleason 2019 dataset. Again, we split the dataset into 4 cross-validation splits for the experimental setup.

For the classification of prostate cancer, the tissue is segmented in the Gleason Grading (GG) scheme. The classes are 'Non-cancerous' (NC), 'Gleason 3' (G3), 'Gleason 4' (G4), and 'Gleason 5' (G5) depending on the architectural growth patterns of the tumor [28, 29]. To visualize the



(a) Image (b) GT Gold (c) Pred. Gold (d) Uncert.

(e) Image (f) GT Gold (g) Pred. Gold (h) Uncert.

Figure 2. Gold prediction and uncertainty of the Pionono model. The first row shows a confident prediction as the uncertainty in 2d is low (white) for almost all the area. Indeed, the segmentation prediction 2c is very accurate, see the ground truth (GT) 2b. The second row shows an example of an uncertain prediction. Some parts of the area classified as G3 (yellow) in 2g are labeled as G4 in the ground truth 2f. These areas are estimated with a high uncertainty (black) in 2h, warning that these predictions are unreliable.

segmentations we use the colors: green for NC, yellow for G3, orange for G4, and red for G5. For the evaluation of algorithms for prostate cancer classification, previous works used the Cohen's kappa coefficient [22, 4, 28] which measures the agreement of two raters or a rater and an AI model. To compare to previously reported results for the two datasets, we use the unweighted Cohen's kappa κ for the Gleason 2019 dataset [22] and the quadratic weighted Cohen's kappa κ for the Arvaniti TMA dataset [4]. The main difference is that the quadratic kappa takes the class order into account and weighs the errors based on the quadratic distance of the predicted and the real class.

The third dataset contains 151 WSIs for breast cancer segmentation that were sliced into 11,836 patches of 512x512 pixels annotated by 25 raters [3, 19]. We will refer to this dataset as "bc segmentation". The tissue was segmented into "tumor", "inflammation", "necrosis", "stroma", and "other". Here, the gold labels were obtained by an actual discussion of experts. We use the predefined train/validation/test splits [19].

For all datasets we use image augmentation with the albumentations library [6] by applying random flip, rotation, shear, zoom, blur, and shifts in brightness, contrast, hue, and saturation. This leads to a broad range of realistic transformations of the image to avoid overfitting.

4.2. Uncertainty estimation

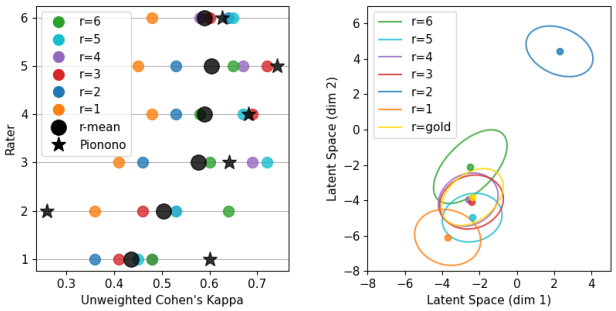
The proposed model provides probabilistic predictions that allow an accurate assessment of the predictive uncertainty. Fig. 2 shows the model predictions and uncertainties obtained with the gold distribution as described in section 3.2. For the first image (2a), the prediction of the model (2c) is accurate and matches the real gold annotation (2b) very well. The uncertainty (2d) for this predictions is low (white), which means that there is a low risk of a wrong prediction. Therefore, the model correctly indicates that this prediction is reliable. For the second image 2e, some areas that are predicted as G3 (yellow) in (2g) are actually G4 in the ground truth gold prediction (2f). The model's uncertainty estimation indicates that this prediction is not reliable: the misclassified areas are marked with a high uncertainty (dark) in the image (2h). Therefore, the probabilistic output adds valuable information to the diagnostic process. It estimates if a prediction is reliable - or unreliable and should be double-checked.

4.3. Inter- and Intra-observer Variation

The Pionono model is able to capture the inter- and intra-observer variability. This accurate probabilistic modeling of the annotations does not only improve the predictive results (see section 4.4), but also allows to simulate specific experts at test time. In this section, we empirically show that the model learns the different label behaviors of the raters and is able to reproduce them.

In Fig. 3a we plot the *inter-observer variations* between the raters. The figure shows that there is indeed a high variability among the raters, with a Cohen's kappa ranging from 0.36 to 0.72. The simulated test predictions by Pionono show a higher agreement with each rater than the average agreement of the other raters, except for rater 2. For two raters (1 and 5), the simulated predictions of Pionono are even more than 15 percentage points higher than the average rater agreement. We also measured the IoU metric, which was $0.574, 0.540, 0.619, 0.649, 0.692, 0.507$, for the 6 raters respectively, compared to a mean inter-pathologist IoU of 0.361. The results confirm that most raters are modeled with high accuracy.

Fig. 3b shows the posterior distributions $q(z|r)$ of the proposed model, encoding the labeling behavior of each rater. The following observations confirm, that these learned distributions approximate well the real-world labeling behavior of the raters: (i) The four raters $3, 4, 5$ and 6 show a high overlap of the distributions and corresponding to a high labeling agreement shown in Fig. 3a. (ii) The gold distribution (simulating raters agreement) overlaps significantly with the distribution of these four raters. (iii) The distribution of rater 2 is far away from all other distributions. This rater shows a different labeling behavior due to frequent under-segmentation of images, assigning the 'background' class to areas that contain tissue. (iv) Raters 1 and 6 often deviate from the other raters, especially for the differentiation of classes G3 and G4. Their distribution accordingly has a smaller overlap with the gold distribution and the other raters. Fig. 4 shows some visual image examples of Pionono test predictions, simulating each rater r by drawing samples from the corresponding distribution $q(z|r)$ and then taking the mean of the output samples. The examples confirm that the rater differences are modeled well.



(a) Rater simulation statistic (b) Latent distributions

Figure 3. Analyzing the labeling behavior. In Fig. (a) the agreement of each rater with all other raters is depicted, measured by the unweighted Cohen's Kappa of the true labels [22]. The mean agreement of each rater with all other raters is represented by a black dot and the agreement with Pionono test predictions, simulating the corresponding rater, by a star. This confirms that the model accurately models each rater, reaching even a higher agreement than the other rater's average, except for rater 2. Fig. (b) shows the first two dimensions of the posterior distributions $q(z|r)$ with mean and covariance after training. The distributions of raters $3, 4, 5$ and 6 overlap significantly with the gold distribution and with each other, indicating a similar labeling behavior. Indeed, these raters show the highest labeling agreement in Fig. (a).

Next, we analyze the *intra-observer variations*. As the dataset does not contain multiple annotations of the same rater for the same image, the assessment of this quality is more difficult. Still, certain intra-observer variability can be assessed by observing the general labeling behavior of one annotator. For example, rater 6 tends to over-assign class G5 (red), and rater 2 tends to not segment all image parts that contain tissue. Interestingly, these intra-observer variations are present in the model predictions when multiple samples are drawn from their corresponding distribution. Fig. 5 shows visual examples of the simulated variations of raters 2 and 6.

4.4. Model Comparison

The proposed model is compared to previously reported results and several state-of-the-art approaches (see section 2) for medical image segmentation with labels from multiple raters. For fair comparison, we use the same backbone architecture[1], epochs, learning rate, and optimizer for all experiments. We have tuned the model-specific hyperparameters to obtain the best possible results for each method.

First, we perform experiments with the Gleason 2019 dataset with a 4-fold crossvalidation. For comparison with previous works, we report the unweighted Cohen's kappa metric comparing gold predictions with gold ground truth. Additionally, we report the accuracy. As the results in Table 2 show, the proposed Pionono model outperforms the previously reported results [22, 24], including the winner of the Gleason 2019 challenge [24], by a large margin of over 20 percentage points. This accounts for the exact modeling of the raters by Pionono, but also for the different choices of backbone architecture and other training details. Compared to other state-of-the-art methods with the same architecture and training details, Pionono still shows a considerably better performance.

Next, we compare the generalization capabilities of the models by using the Arvaniti TMA dataset as an external test set, as reported in Table 3. This means, that the models are trained with all images from Gleason 2019 and tested with all images from Arvaniti TMA. As the Arvaniti TMA dataset does not contain gold labels, the model's gold predictions are compared to both raters independently, as previously done by Arvaniti *et al.* [4]. We observe that the model

[1]Only the model *CM pixel* uses ResNet18 to fit on the GPU.

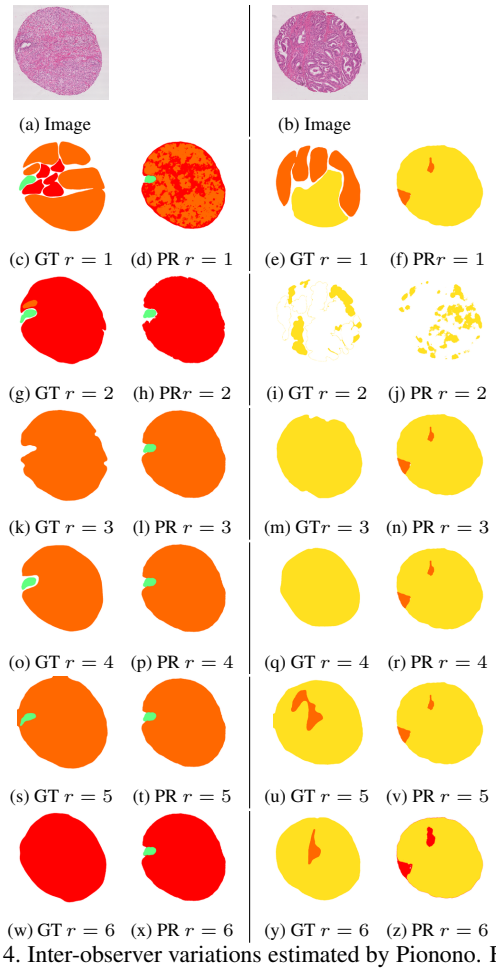


Figure 4. Inter-observer variations estimated by Pionono. For two test images we depict the ground truth (GT) segmentations of all raters and the predicted segmentations (PR), simulating each rater. The proposed model is able to simulate certain labeling behavior like the tendency of assigning class G5 (red) for raters 2 and 6 (see g and w) where other raters assigned G4 (orange). Furthermore, the model captures the under-segmentation by rater 2 (see i).

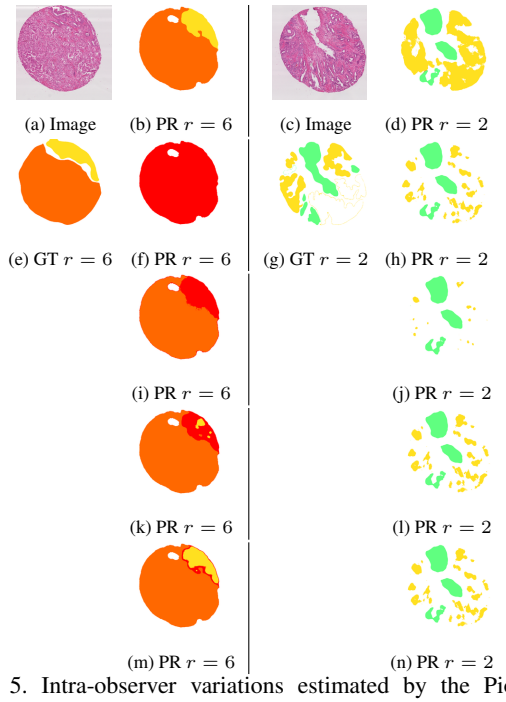


Figure 5. Intra-observer variations estimated by the Pionono model. For two example images we depict the true annotation (GT) of two different raters ($r = 2$ and $r = 6$). On the right sides we show different coherent segmentation predictions (PR) for each rater estimated by our model. The differences in each column reflect possible intra-observer variations. The first example (a) shows that rater 6 might show some variations in the assigned classes. While the segmentation prediction (b) is composed by classes G3 (yellow) and G4 (orange), the segmentation sample (f) only consists of G5 (red). Indeed, this rater often assigns class G5 (red) in areas where other raters assign G4 (see Fig. 4) such that this is a plausible hypothesis. In the second example (c) we see variations due to the under-segmentation of rater 2 in some images. Our model captures this behavior and provides hypothesis of more (d) or less (j) segmentation of class G3 (yellow).

generalizes better than all other methods, achieving a higher agreement in terms of Cohen's quadratic kappa with both raters. In terms of accuracy, only 'CM global' outperforms Pionono by a small margin.

In the third experiment, we use the Arvaniti TMA dataset for training and testing with the two rater annotations. Again, the proposed model is able to outperform previously reported results as well as other state-of-the-art methods in terms of quadratic Cohen's kappa. In terms of accuracy, only the "Prob U-Net" model obtains a better result for rater 1, while Pionono reaches the best accuracy for rater 2.

To validate the model on a different kind of data, we performed the fourth experiment on the "bc segmentation" dataset [3]. The results are reported in Table 5 and confirm the strong performance of the proposed Pionono model.

The results support our hypothesis that explicitly modeling the inter- and intra-observer variations improves the model's performance. Pionono takes the different labeling behavior into account during training which leads to accu-

Method	Unweighted κ	Accuracy
Nir *et al.* [22]	0.51	N.A.
Qiu *et al.* [24]	0.524	N.A.
STAPLE	0.75 ± 0.006	0.834 ± 0.005
Prob U-Net	0.741 ± 0.002	0.83 ± 0.001
CM global	0.721 ± 0.018	0.814 ± 0.012
CM pixel	0.692 ± 0.019	0.791 ± 0.012
Pionono	$\mathbf{0.758} \pm 0.011$	$\mathbf{0.84} \pm 0.007$

Table 2. Cohens Kappa Comparison for the 4-fold crossvalidation experiment of the Gleason 2019 dataset, reported by mean and standard error.

Rater 1 Method	Quadratic κ	Accuracy
STAPLE	0.629 ± 0.002	0.718 ± 0.001
Prob U-Net	0.629 ± 0.005	0.73 ± 0.003
CM global	0.624 ± 0.003	0.728 ± 0.002
CM pixel	0.618 ± 0.007	0.72 ± 0.004
Pionono	$\mathbf{0.641} \pm 0.006$	$\mathbf{0.736} \pm 0.004$
Rater 2 Method	Quadratic κ	Accuracy
STAPLE	0.563 ± 0.002	0.621 ± 0.002
Prob U-Net	0.56 ± 0.003	0.626 ± 0.006
CM global	0.557 ± 0.003	$\mathbf{0.638} \pm 0.006$
CM pixel	0.551 ± 0.006	0.626 ± 0.008
Pionono	$\mathbf{0.569} \pm 0.005$	0.633 ± 0.005

Table 3. Cohens Kappa Comparison with the two raters of the Arvaniti TMA dataset trained on the Gleason 2019 dataset, reported by mean and standard error.

Rater 1 Method	Quadratic κ	Accuracy
Arvaniti *et al.* [4]	0.55	N.A.
Silva-R. *et al.* [28]	0.536	N.A.
supervised	0.658 ± 0.025	0.734 ± 0.008
Prob U-Net	0.697 ± 0.008	$\mathbf{0.762} \pm 0.004$
CM global	0.677 ± 0.028	0.745 ± 0.011
CM pixel	0.647 ± 0.016	0.731 ± 0.012
Pionono	$\mathbf{0.716} \pm 0.011$	0.751 ± 0.02
Rater 2 Method	Quadratic κ	Accuracy
Arvaniti	0.49	N.A.
supervised	0.521 ± 0.014	0.678 ± 0.014
Prob U-Net	0.534 ± 0.002	0.68 ± 0.005
CM global	0.533 ± 0.022	0.676 ± 0.011
CM pixel	0.508 ± 0.013	0.663 ± 0.008
Pionono	$\mathbf{0.548} \pm 0.008$	$\mathbf{0.697} \pm 0.012$

Table 4. Cohens Kappa Comparison with the two raters of the Arvaniti TMA dataset trained and validated by 4-fold crossvalidation of the Arvaniti data, reported by mean and standard error.

Method	Unweighted κ	Accuracy
STAPLE	0.647 ± 0.003	0.755 ± 0.002
Prob U-Net	0.685 ± 0.023	0.734 ± 0.004
CM global	0.654 ± 0.005	0.761 ± 0.004
CM pixel	0.689 ± 0.010	0.784 ± 0.007
Pionono	$\mathbf{0.711} \pm 0.002$	$\mathbf{0.799} \pm 0.001$

Table 5. Results for the breast cancer segmentation of WSIs, reported as mean and standard error of 4 runs.

rate predictions.

4.5. Robustness to Hyperparameter Settings

To measure the sensitivity of the model regarding different hyperparameters, we performed studies on the 4-fold cross-validation experiment of the Gleason 19 dataset. Table 6 shows that the model is robust to variations of all analyzed hyperparameters. We observe minor performance drops for different values of the regularization factor λ and the initialization variance σ^2_{post}. In both cases, wrong choices of the hyperparameters can hinder the correct optimization of the latent distributions. Furthermore, we tested different backbone architectures, indicating a limited performance with a VGG16 backbone. Overall the performance drops are minor and for all other settings, the model shows highly accurate results of $\kappa > 0.75$.

4.6. Required Resources

For the Gleason 2019 dataset with images of 1024×1024, the model can be trained with a batch size of 3 on a single NVIDIA GeForce RTX 3090 with 24Gb memory. The training takes less than 1.5h in total and test predictions less than 0.2s per image. The trained model occupies less than 350Mb when saved to the disk. As each additional annotator adds only one additional vector $\mu^r \in \mathbb{R}^8$ and one covariance matrix $\Sigma^r \in \mathbb{R}^{8\times8}$, it is scalable to a large number of annotators. The model's quick runtime and excellent scalability make it easily applicable in clinical practice.

4.7. Limitations

As semantic segmentation itself is a challenging task, some details of the annotator segmentations are not captured well by the model, such as the variations of class NC (green) in the GT of Fig. 4d - 4x or class G4 (orange) in the GT of Fig. 4f - 4z. Here, the model tends to predict similar shapes for the raters. A possible solution is to use more layers in the segmentation head f_θ with a wider kernel (e.g. 5×5 convolutions). This would increase the complexity of the model and might enable it to capture the different labeling behavior in even more detail.

Hyperp.	Value	Unweighted κ	Accuracy
D	4	0.752 ± 0.005	0.836 ± 0.003
	8	0.758 ± 0.011	0.84 ± 0.007
	16	0.752 ± 0.006	0.836 ± 0.004
σ^2_{prior}	1	0.758 ± 0.007	0.839 ± 0.004
	2	0.758 ± 0.011	0.84 ± 0.007
	4	0.757 ± 0.007	0.839 ± 0.004
σ^2_{post}	4	0.757 ± 0.007	0.839 ± 0.004
	8	0.758 ± 0.011	0.84 ± 0.007
	16	0.745 ± 0.009	0.83 ± 0.005
λ	0.0001	0.744 ± 0.003	0.829 ± 0.003
	0.0005	0.758 ± 0.011	0.84 ± 0.007
	0.001	0.745 ± 0.008	0.837 ± 0.005
ν	0.01	0.757 ± 0.004	0.839 ± 0.002
	0.02	0.758 ± 0.011	0.84 ± 0.007
	0.04	0.753 ± 0.01	0.836 ± 0.005
Backbone	VGG16	0.734 ± 0.01	0.823 ± 0.005
	Resnet34	0.758 ± 0.011	0.84 ± 0.007
	Eff.netB2	0.754 ± 0.01	0.836 ± 0.004

Table 6. Study of hyperparameter robustness using the Gleason 2019 dataset. The default hyperparameter value is marked with bold letters. While varying one hyperparameter, all other values are set to the default value. We observe consistent and robust performance across all settings of tested hyperparameters.

5. Conclusions

In this work we present "Pionono", a method for medical image segmentation that models the inter- and intra-observer variability explicitly with a probabilistic approximation. This is especially relevant for tasks where the labeling behavior of medical experts is known to vary widely, such as in the case of prostate cancer segmentation. Our experiments on real-world cancer segmentation data demonstrate that Pionono outperforms state-of-the-art models such as STAPLE, Probabilistic U-Net, and models based on confusion matrices. Apart from the improved predictive performance, it provides a probabilistic uncertainty estimation and the simulation of expert opinions for a given test image. This makes it a powerful tool for medical image analysis and has the potential to improve the diagnostic process considerably.

6. Acknowledgements

This work was funded by the European Union's H2020 research and innovation programme (Marie Skłodowska Curie grant agreement No 860627, CLARIFY Project), the Spanish Ministry of Science and Innovation (project PID2019-105142RB-C22), and FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades (project P20_00286).

References

[1] Abdullah A Abdullah, Masoud M Hassan, and Yaseen T Mustafa. A review on bayesian deep learning in healthcare: Applications and challenges. 2022. 2

[2] Felicia D. Allard, Jeffrey D. Goldsmith, Gamze Ayata, Tracy L. Challies, Robert M. Najarian, Imad A. Nasser, Helen Wang, and Eric U. Yee. Intraobserver and interobserver variability in the assessment of dysplasia in ampullary mucosal biopsies. 42(8):1095–1100, 2018. 1, 2

[3] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai A T Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem S E Salem, Ahmed F Ismail, Anas M Saad, Joumana Ahmed, Maha A T Elsebaie, Mustafijur Rahman, Inas A Ruhban, Nada M Elgazar, Yahya Alagha, Mohamed H Osman, Ahmed M Alhusseiny, Mariam M Khalaf, Abo-Alela F Younes, Ali Abdulkarim, Duaa M Younes, Ahmed M Gadallah, Ahmad M Elkashash, Salma Y Fala, Basma M Zaki, Jonathan Beezley, Deepak R Chittajallu, David Manthey, David A Gutman, and Lee A D Cooper. Structured crowdsourcing enables convolutional segmentation of histology images. 35(18):3461–3467, 2019. 1, 5, 7

[4] Eirini Arvaniti, Kim S. Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter J. Wild, Jan H. Rüschoff, and Manfred Claassen. Automated gleason grading of prostate cancer tissue microarrays via deep learning. 8(1):12054, 2018. Dataset link: https://doi.org/10.7910/DVN/OCYCMP. 4, 5, 6, 7

[5] Lieve Brochez, Evelien Verhaeghe, Edouard Grosshans, Eckhart Haneke, Gérald Piérard, Dirk Ruiter, and Jean-Marie Naeyaert. Inter-observer variation in the histopathological diagnosis of clinically suspicious pigmented skin lesions: Observer variation in pigmented lesion diagnosis. 196(4):459–466, 2002. 1

[6] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. 11(2), 2020. 5

[7] Elise C. Covert, Kellen Fitzpatrick, Justin Mikell, Ravi K. Kaza, John D. Millet, Daniel Barkmeier, Joseph Gemmete, Jared Christensen, Matthew J. Schipper, and Yuni K. Dewaraja. Intra- and inter-operator variability in MRI-based manual segmentation of HCC lesions and its impact on dosimetry. 9(1):90, 2022. 1, 2

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, 2016. 1, 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[10] Melih Kandemir, Manuel Haussmann, Ferran Diego, Kumar Rajamani, JeroenVanDer Laak, and Fred Hamprecht. Variational weakly supervised gaussian processes. In *British Machine Vision Conference (BMVC)*, pages 71.1–71.12, 2016. 1, 2

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 4

[12] Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *International Conference on Neural Information Processing Systems - NIPS*, pages 2575–2583, 2015. 4

[13] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. 31, 2018. 1, 2

[14] Nikolas S. Kulberg, Roman V. Reshetnikov, Vladimir P. Novik, Alexey B. Elizarov, Maxim A. Gusev, Victor A. Gombolevskiy, Anton V. Vladzymyrskyy, and Sergey P. Morozov. Inter-observer variability between readers of CT images: all for one and one for all. 2(2):105–118, 2021. 1

[15] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. 142:106816, 2020. 1, 2

[16] Jiayun Li, William Speier, King Chung Ho, Karthik V. Sarma, Arkadiusz Gertych, Beatrice S. Knudsen, and Corey W. Arnold. An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. 69:125–133, 2018. 1, 2

[17] Jasper Linmans, Stefan Elfwing, Jeroen van der Laak, and Geert Litjens. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. 83:102655, 2023. 1, 2

[18] Miguel López-Pérez, Mohamed Amgad, Pablo Morales-Álvarez, Pablo Ruiz, Lee A. D. Cooper, Rafael Molina, and Aggelos K. Katsaggelos. Learning from crowds in digital pathology using scalable variational gaussian processes. 11(1):11612, 2021. 1, 2

[19] Miguel López-Pérez, Pablo Morales-Álvarez, Lee A. D. Cooper, Rafael Molina, and Aggelos K. Katsaggelos. Crowdsourcing segmentation of histopathological images using annotations provided by medical students. 13897:245–249. 5

[20] Miguel López-Pérez, Arne Schmidt, Yunan Wu, Rafael Molina, and Aggelos K. Katsaggelos. Deep gaussian processes for multiple instance learning: Application to CT intracranial hemorrhage detection. 219:106783. 2

[21] Amirreza Mahbod, Gerald Schaefer, Benjamin Bancher, Christine Löw, Georg Dorffner, Rupert Ecker, and Isabella Ellinger. CryoNuSeg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images. 132:104349, 2021. 1, 2

[22] Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F. Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F. Villamil, Gang Wang, R. Storey Wilson, Kenneth A. Iczkowski, M. Scott Lucia, Peter C. Black, Purang Abolmaesumi, S. Larry Goldenberg, and Septimiu E. Salcudean. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. 50:167–180, 2018.

Dataset link: https://gleason2019.grand-challenge.org/. 1, 4, 5, 6, 7

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems (NeurIPS)*, 32, pages 8024–8035. 2019. 3

[24] 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. 12:772403, 2022-04-08. 6, 7

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351, pages 234–241, 2015. 3

[26] Arne Schmidt, Pablo Morales-Álvarez, and Rafael Molina. Probabilistic attention based on gaussian processes for deep multiple instance learning. pages 1–14, 2023. 1, 2

[27] Arne Schmidt, Julio Silva-Rodriguez, Rafael Molina, and Valery Naranjo. Efficient cancer classification by coupling semi supervised and multiple instance learning. 10:9763–9773. 1

[28] Julio Silva-Rodríguez, Adrián Colomer, María A. Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. 195:105637, 2020. 1, 2, 4, 5, 7

[29] Julio Silva-Rodríguez, Arne Schmidt, María A. Sales, Rafael Molina, and Valery Naranjo. Proportion constrained weakly supervised histopathology image classification. 147:105714. 4

[30] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248, 2017. 4

[31] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11244–11253, 2019. 2, 3

[32] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. 23(7):903–921, 2004. 2, 4

[33] Yunan Wu, Arne Schmidt, Enrique Hernández-Sánchez, Rafael Molina, and Aggelos K. Katsaggelos. Combining attention-based multiple instance learning and gaussian processes for CT hemorrhage detection. 12902:582–591. 1, 2

[34] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof, and Daniel C. Alexander. Disentangling human error from the ground truth in segmentation of medical images. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3