Supplementary Material

A. Clustering Algorithm

The algorithms presented in Algorithm .1 and Algorithm .2 are utilized to cluster predicted instances obtained from multiple forward passes through the StarDist model.

```
Algorithm .1: Clustering with Pixel Approach
Data: Set of samples $\mathbb{S} = \{S_1, S_2, ..., S_F\}$
Result: Clusters $\mathbb{O} = \{O_1, O_2, ..., O_M\}$
$\mathbb{O} = \emptyset$
$\theta_{IoU} = 0.5$ (IoU threshold)
$z = 0$
for $S_f$ in $\mathbb{S}$ do
    for $P_{u_f}$ in $S_f$ do
        if $z > 0$ then
            for $O_m$ in $\mathbb{O}$ do
                for $P_{v_f}$ in $O_m$ do
                    if $IoU(P_{u_f}, P_{v_f}) \geq \theta_{IoU}$ then
                        Add $P_{u_f}$ in $O_m$
                    else
                        Add $P_{u_f}$ in $O_{m+1}$
        else
            Add $P_{u_f}$ in $O_m$
            $z+ = 1$
    Add $O_m$ in $\mathbb{O}$
```

```
Algorithm .2: Clustering with Radial Approach
Data: Set of samples
      $\mathbb{G} = \{\{D_1, R_1\}, \{D_2, R_2\}, ..., \{D_F, R_F\}\}$
Result: Clusters $\mathbb{O} = \{O_1, O_2, ..., O_M\}$
$\mathbb{O} = \emptyset$
$\theta_d = 0.5$ (Object probability threshold)
$\mu_G$ = ComputeMean($\mathbb{G}$)
$\mathbb{C}$ = NonMaxSuppression($\mu_G$)
for $m$ in $\{1, 2, ..., \mid \mathbb{C} \mid= M\}$ do
    Create a new cluster $O_m$
    for $f$ in $\{1, 2, ..., F\}$ do
        $P_{m_f} =$
          CreateInstance($(x_m, y_m), \{\{r^n_{x_m,y_m}\}^n_{i=1}\}_f$
        if $\{d_{x_m,y_m}\}_f \geq \theta_d$ then
            Add $P_{m_f}$ to $O_m$
    Add $O_m$ to $\mathbb{O}$
```

B. Quality of Certainty Score

The calibration diagrams in Figure 1 and Figure 2 with bin size $B = 10$, show the calibration quality of the different certainty scores using the Radial Approach with Monte-Carlo Dropout technique ($d_{rate} = 0.8$, $F = 20$) on the validation set for the DSB2018 and GlaS datasets. The diagrams compare hybrid certainty score (c_{hyb}), spatial certainty score (c_{spl}), and fractional certainty score (c_{frac}).

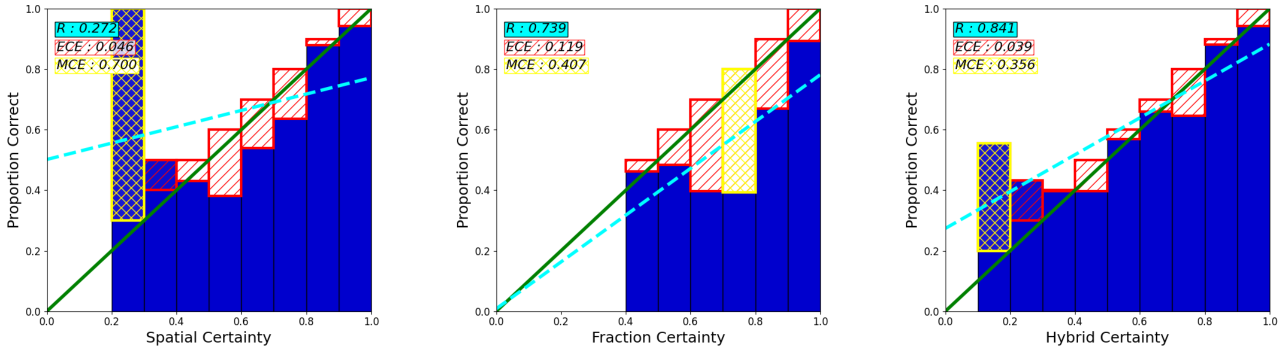
In Figure 1, we observe that the hybrid certainty score c_{hyb} for the DSB2018 dataset exhibits better calibration as the certainty score closely approximates the expected accuracy.

However, in Figure 2, we observe that the GlaS dataset exhibits unsatisfactory and notably elevated calibration errors. This phenomenon can be attributed to the dataset's incongruity with the StarDist model, resulting in higher certainty scores for incorrect predictions.

C. Effects of Forward Passes on Certainties Quality

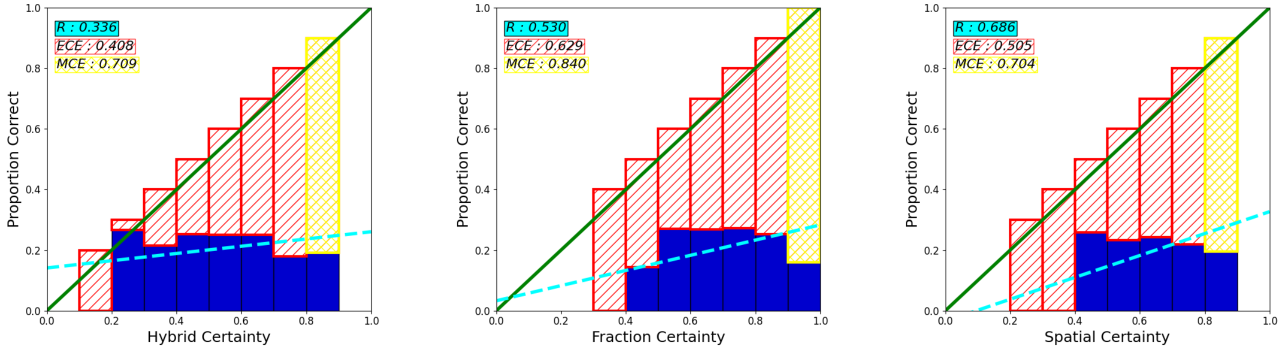
The influence of the number of forward passes F on the calibration of the hybrid certainty (c_{hyb}) for the DSB2018 and GlaS datasets was assessed and the results are visualized in Figure 3 and Figure 4 . The calibration errors, measured by Pearson's R, Expected Calibration Error (ECE), and Maximum Calibration Error (MCE), are plotted against F.

A consistent pattern emerges, resembling the observations made with the Bubble dataset. As the number of forward passes increases, the calibration errors tend to converge, aligning with the principles of the Central Limit Theorem. Furthermore, distinct convergence behaviors are observed for each dropout rate.



(a) (b) (c)

Figure 1. Calibration diagrams illustrating certainty score estimation for the DSB2018 dataset. Panels (a), (b), and (c) show spatial certainty (c_{spl}), fractional certainty (c_{frac}), and hybrid certainty (c_{hyb}) scores, respectively. These scores are calculated using the Pixel Approach and Monte-Carlo Dropout with a dropout rate of 0.8, and $F = 20$ forward passes. Notably, the hybrid certainty scores (c_{hyb}) demonstrate superior calibration compared to individual certainty scores across three calibration error metrics: Pearson Correlation Coefficient (R), Expected Calibration Error (ECE), and Maximum Calibration Error (MCE).



(a) (b) (c)

Figure 2. Calibration diagrams illustrating certainty score estimation for the GlaS dataset. Panels (a), (b), and (c) show spatial certainty (c_{spl}), fractional certainty (c_{frac}), and hybrid certainty (c_{hyb}) scores, respectively. These scores are calculated using the Pixel Approach and Monte-Carlo Dropout with a dropout rate of 0.8, and $F = 20$ forward passes. It is evident that the GlaS dataset demonstrates unsatisfactory calibration for all certainty scores across three calibration error metrics: Pearson Correlation Coefficient (R), Expected Calibration Error (ECE), and Maximum Calibration Error (MCE).

D. Effects of Dropout Location on Certainties Quality

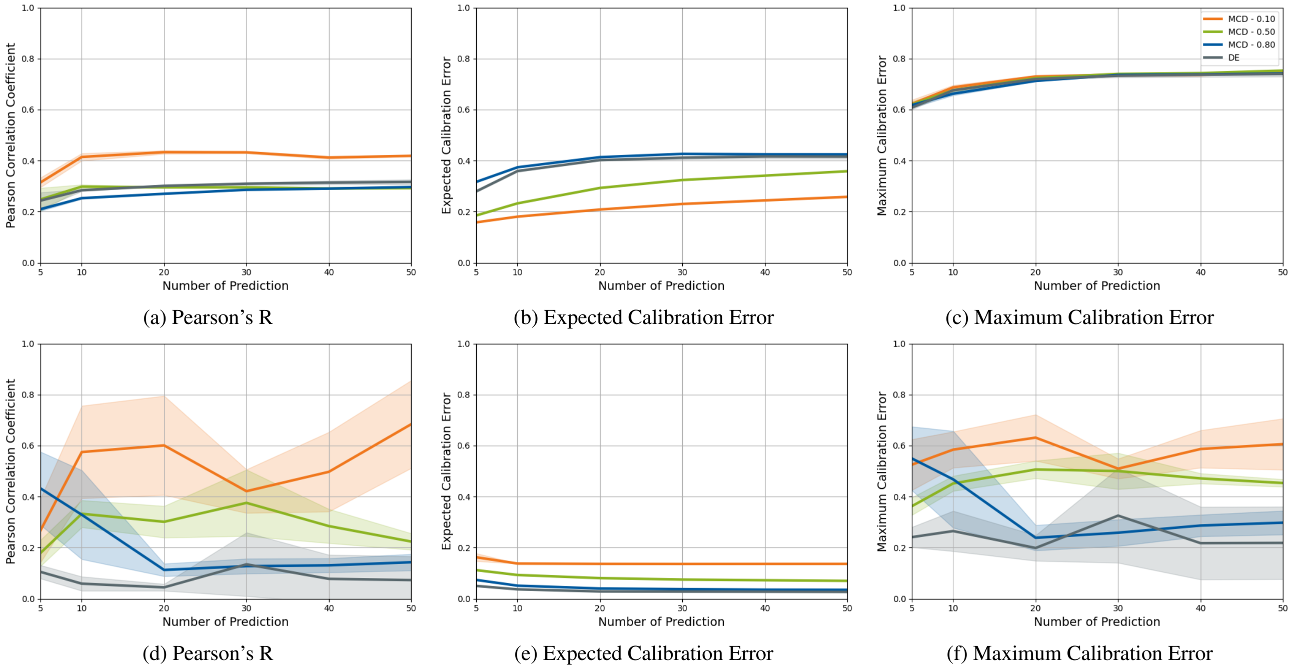
We conducted an investigation into the calibration of hybrid certainty (u_{hyb}) obtained from various dropout layer positions, as depicted in Figure 5. The number of forward passes was kept constant at $F = 20$.

The calibration errors of hybrid certainty (u_{hyb}) derived from different dropout layer positions is presented in Figure 6, Figur 7, and Figure 8 for the Bubble, DSB2018, and GlaS datasets. The influence of dropout layer locations on calibration is assessed through calibration metrics, including the Pearson Correlation Coefficient, Expected Calibration Error, and Maximum Calibration Error.

The relationship between the location of the dropout layer and calibration performance does not exhibit consistent patterns.

E. Intersection over Union

Intersection over Union (IoU) is a metric used to assess the accuracy of two masks, often in tasks like image segmentation. It measures the extent of overlap between the regions represented by the two masks. In simpler terms, IoU helps you understand how much the areas covered by two masks align with each other. It's a way to quantify the similarity between the shapes or regions defined by the masks. This metric is especially valuable when dealing with tasks where you want to compare how well two masks match each other. A higher IoU score indicates that the masks closely match, while a lower score suggests a greater discrepancy between the regions defined by the masks.



(a) Pearson's R

(b) Expected Calibration Error

(c) Maximum Calibration Error

(d) Pearson's R

(e) Expected Calibration Error

(f) Maximum Calibration Error

Figure 3. Plot showing calibration errors as a function of the number of forward passes for the DSB2018 dataset using the Monte-Carlo Dropout and Deep Ensemble techniques (calibration errors as a function of the number of models in the case of Deep Ensemble). Panels (a) to (c) depict certainty estimates using the Pixel Approach, while panels (d) to (f) represent certainty estimates using the Radial Approach. Notably, the Deep Ensemble technique exhibits faster convergence of calibration errors compared to the Monte-Carlo Dropout technique. Additionally, distinctive convergence patterns are observed for each dropout rate.

$$p^{x,y} = \begin{cases} 1 & \text{if } (x,y) \text{ belong to an instance} \\ 0 & \text{else} \end{cases} \qquad (1)$$

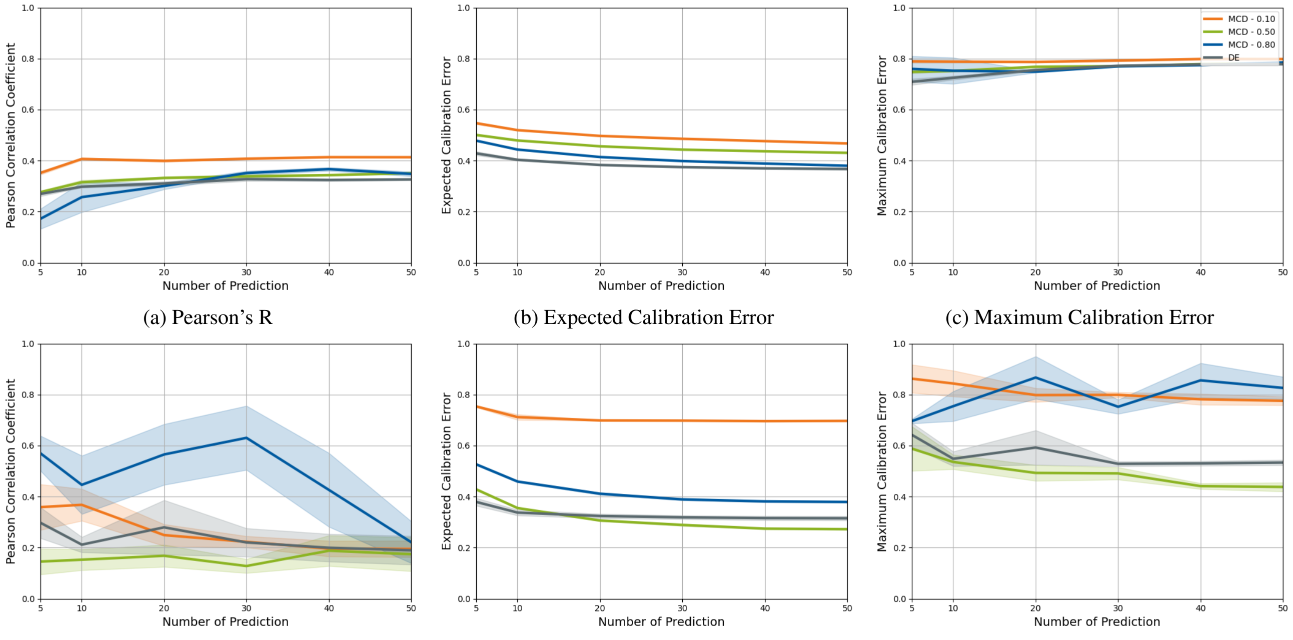
$$IoU(P_u, P_v) = \sum_{y=1}^{Y} \sum_{x=1}^{X} \frac{p_u^{x,y} \cdot p_v^{x,y}}{p_u^{x,y} + p_v^{x,y} - p_u^{x,y} \cdot p_v^{x,y}} \qquad (2)$$

Equation 1 defines the binary masks for the pixels. In this equation, $p^{x,y}$ represents the pixel value at coordinates (x, y). If the pixel belongs to an instance of interest, $p^{x,y}$ is assigned a value of 1. Otherwise, if the pixel does not belong to the instance, $p^{x,y}$ is assigned a value of 0. This binary representation helps distinguish between the pixels that are part of the instance and those that are not. Equation 2 calculates the IoU between two binary masks P_u and P_v. The goal of this equation is to quantify how well the mask P_u aligns with the mask P_v. The summation over x and y iterates through all pixels in the masks. The terms $p_u^{x,y}$ and $p_v^{x,y}$ represent the binary pixel values in masks P_u and P_v at the same coordinates (x, y), respectively. The IoU is then calculated by dividing the sum of the element-wise product $p_u^{x,y} \cdot p_v^{x,y}$ by the sum of the pixel values in P_u and P_v, minus the sum of the element-wise product. This calculation effectively captures the overlap between the two masks and provides a value that indicates the degree of similarity between the two masks.

F. Choice of Dropout Rates

To investigate the effect of dropout rates on uncertainty estimation in instance segmentation with the StarDist model, we carefully selected three distinct dropout probabilities, namely $0.1, 0.5$, and 0.8. The decision to use a limited number of dropout rates was primarily driven by computational complexity, as uncertainty estimation using the sampling techniques can be computationally demanding. By focusing on these representative dropout rates, we aimed to strike a balance between resource efficiency and comprehensive analysis.

Our choice of 0.1 and 0.8 as two extremes of the dropout rates, and a dropout rate of 0.5, situated between the two extremes, ensure that our investigation spans a relevant and informative range of dropout rates. This selection enables us to analyze uncertainty estimation with varying degrees of dropout rates while maintaining a manageable computational load. The insights gained from studying these specific dropout rates in combination with diverse layer locations within the U-Net architecture will provide valuable contributions to the understanding of uncertainty estimation for instance segmentation in the context of the StarDist model.



(a) Pearson's R

(b) Expected Calibration Error

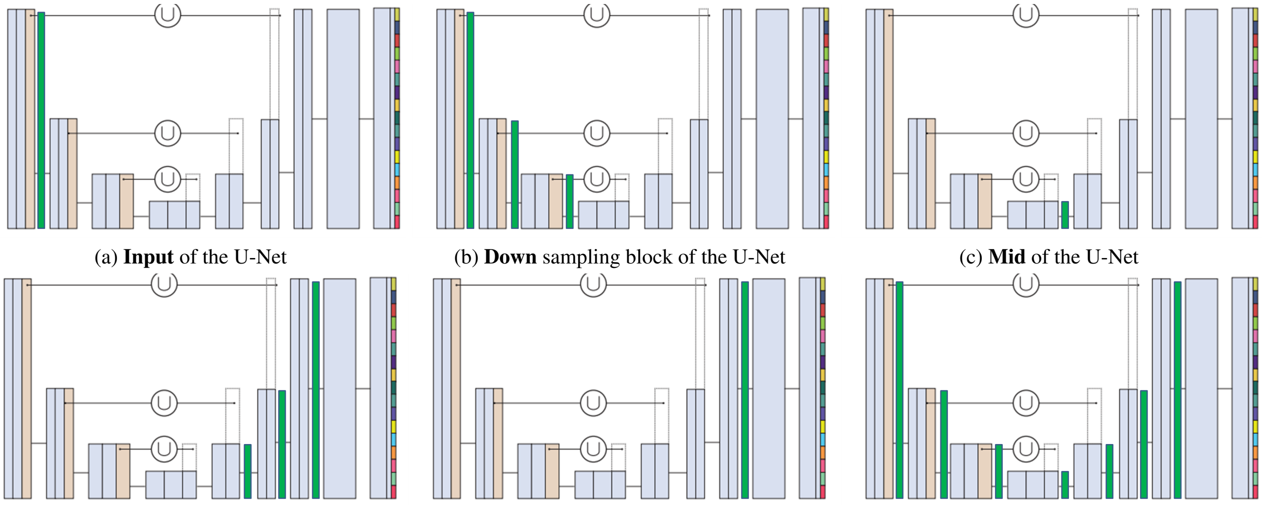
(c) Maximum Calibration Error

(d) Pearson's R

(e) Expected Calibration Error

(f) Maximum Calibration Error

Figure 4. Plot showing calibration errors as a function of the number of forward passes for the GlaS dataset using the Monte-Carlo Dropout and Deep Ensemble techniques (calibration errors as a function of the number of models in the case of Deep Ensemble). Panels (a) to (c) depict certainty estimates using the Pixel Approach, while panels (d) to (f) represent certainty estimates using the Radial Approach. Notably, the Deep Ensemble technique exhibits faster convergence of calibration errors compared to the Monte-Carlo Dropout technique. Additionally, distinctive convergence patterns are observed for each dropout rate.



(a) **Input** of the U-Net

(b) **Down** sampling block of the U-Net

(c) **Mid** of the U-Net

(d) **Up** sampling block of the U-Net

(e) **Output** of the U-Net

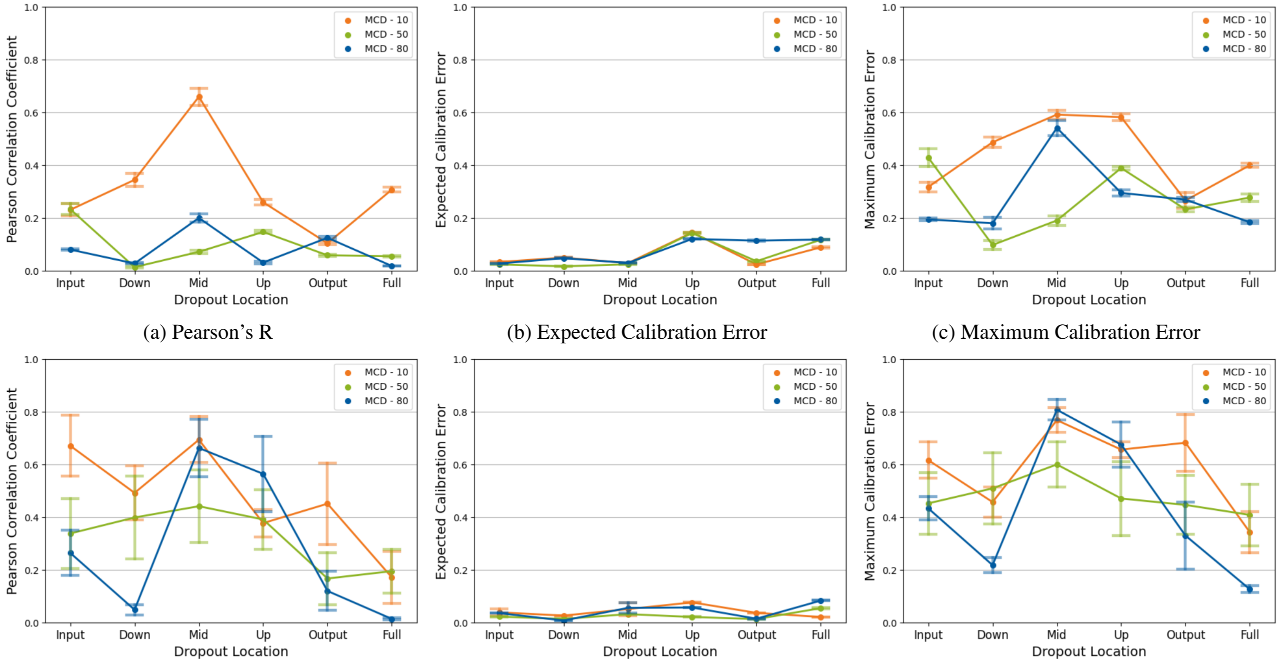
(f) **Full** U-Net

Figure 5. Visualization of varying dropout layer locations within the U-Net block of the StarDist model, with the dropout layers highlighted in green.

G. Uncertainty Visualization

In Figure 9, we showcase a series of example images drawn from three distinct datasets, each serving to visualize uncertainty and gauge the associated prediction certainty. These images provide valuable insights into the model's certainty when predicting diverse instances across various datasets.

In each image of Figure 9, the red polygons represent median cluster predictions $\overline{P}_m \mid m \in \{1, 2, ..., M\}$. The



(a) Pearson's R

(b) Expected Calibration Error

(c) Maximum Calibration Error

(d) Pearson's R

(e) Expected Calibration Error

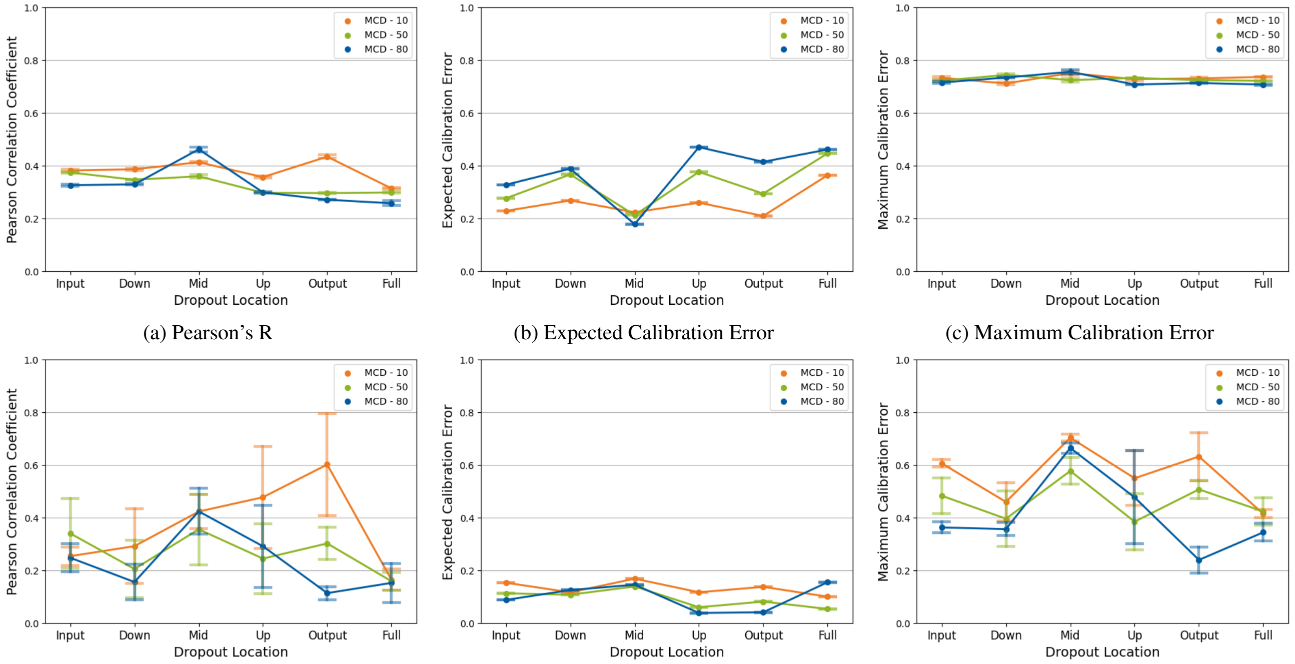
(f) Maximum Calibration Error

Figure 6. Plot showing calibration errors as a function of the dropout layer location for the Bubble dataset using the Monte-Carlo Dropout techniques. Panels (a) to (c) depict certainty estimates using the Pixel Approach, while panels (d) to (f) represent certainty estimates using the Radial Approach. The observed randomness in calibration errors implies a lack of apparent correlation between the dropout layer's placement.

region enclosed by the two yellow polygons delineates the spatial uncertainty peculiar to each specific instance. Additionally, the hybrid certainty score $chyb$ for each bubble prediction is conveniently located in the bottom-right corner.

Bubble and DSB2018 Dataset: Images in Figure 9a and Figure 9b represent instances characterized by star-convex shapes. Our observations reveal a distinct pattern in certainty scores. Correct predictions are accompanied by high certainty scores, while incorrect predictions yield lower certainty scores. This pattern further reinforces the well-calibrated nature of the certainty scores, as corroborated in Figure 5 in the main article and Figure 1.

GlaS Dataset: In this dataset, we explore instances with complex structures. These structures pose a challenge for the StarDist model, resulting in incorrect predictions associated with high certainty scores (Figure 9c). This scenario leads to an unsatisfactory calibration outcome, as evident in Figure 2. This underscores the importance of evaluating the estimated certainties for model reliability and informed decision-making.



(a) Pearson's R

(b) Expected Calibration Error

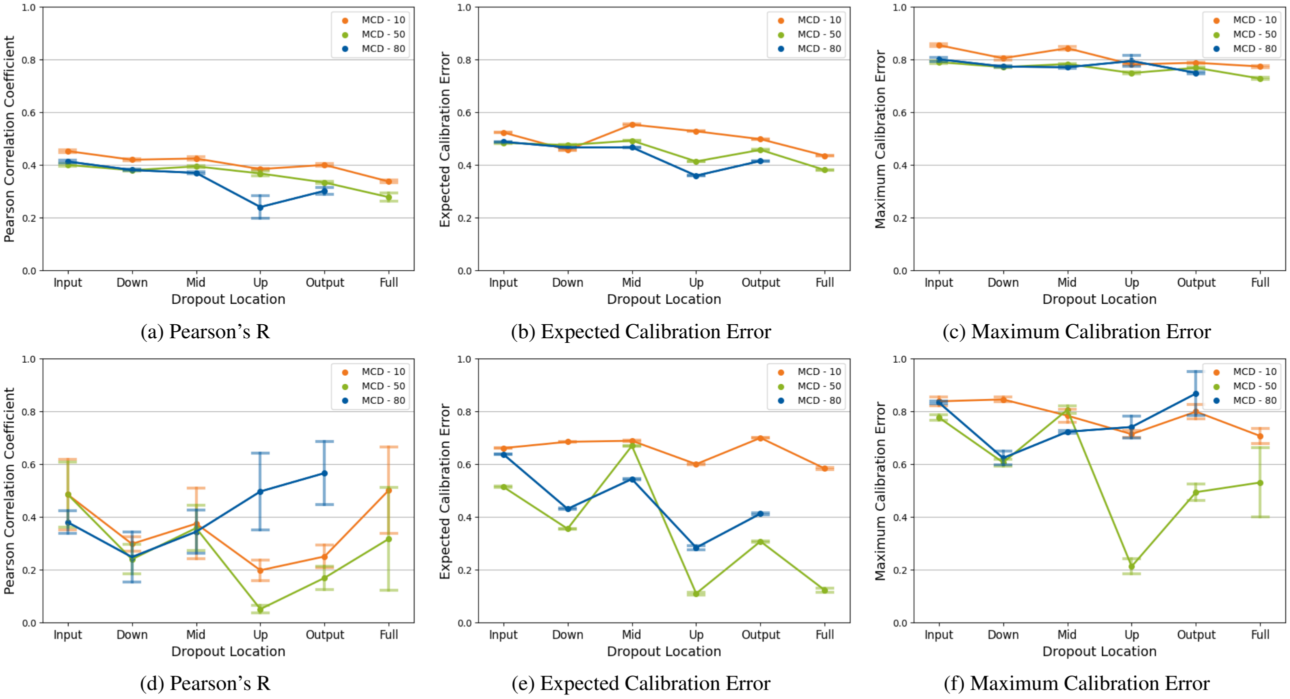
(c) Maximum Calibration Error

(d) Pearson's R

(e) Expected Calibration Error

(f) Maximum Calibration Error

Figure 7. Plot showing calibration errors as a function of the dropout layer location for the DSB2018 dataset using the Monte-Carlo Dropout techniques. Panels (a) to (c) depict certainty estimates using the Pixel Approach, while panels (d) to (f) represent certainty estimates using the Radial Approach. The observed randomness in calibration errors implies a lack of apparent correlation between the dropout layer's placement.



(a) Pearson's R

(b) Expected Calibration Error

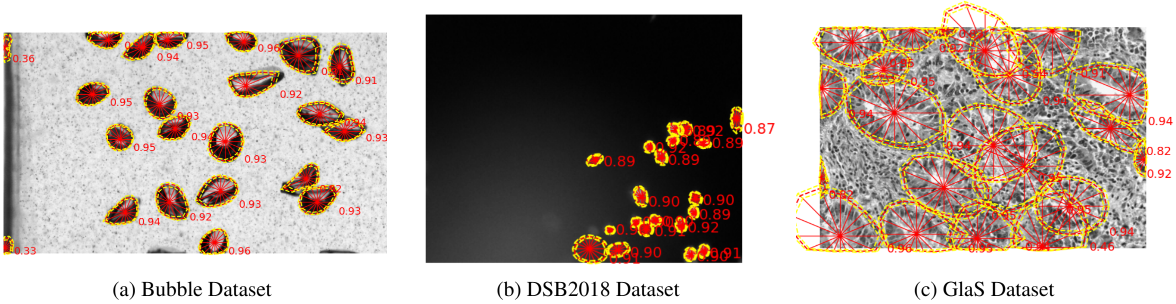
(c) Maximum Calibration Error

(d) Pearson's R

(e) Expected Calibration Error

(f) Maximum Calibration Error

Figure 8. Plot showing calibration errors as a function of the dropout layer location for the GlaS dataset using the Monte-Carlo Dropout techniques. Panels (a) to (c) depict certainty estimates using the Pixel Approach, while panels (d) to (f) represent certainty estimates using the Radial Approach. The observed randomness in calibration errors implies a lack of apparent correlation between the dropout layer's placement. No calibration error is observed when utilizing the Full U-Net 5f configuration with a dropout rate of $d_{rate} = 0.8$. This outcome is due to the model's inability to generate instance predictions, given that it is not optimized for this particular dataset.



(a) Bubble Dataset

(b) DSB2018 Dataset

(c) GlaS Dataset

Figure 9. Visualization of the uncertainty for the three datasets. The red polygons represent median cluster predictions, and the region enclosed by the two yellow polygons delineates the spatial uncertainty peculiar to each specific instance. For (a) Bubble and (b) DSB2018 datasets, correct predictions are accompanied by high certainty scores, while incorrect predictions yield lower certainty scores. In the (c) GlaS dataset, we observe incorrect predictions associated with high certainty scores.