Evidential DualUNet: Single-Pass Uncertainty for Cell Instance Segmentation

David Anglada-Rotger DAVID.ANGLADA@UPC.EDU
Ferran Marques FERRAN.MARQUES@UPC.EDU
Montse Pardàs MONTSE.PARDAS@UPC.EDU
Image Processing Group (GPI), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

Editors: Under Review for MIDL 2026

Abstract

Accurate and trustworthy cell instance segmentation requires models that not only detect and classify nuclei but also communicate how much evidence supports each prediction. DualU-Net is a fast and effective two-head multi-task architecture for this problem, but—like most deterministic models—it provides no principled uncertainty estimates. We introduce *Evidential DualU-Net*, the first evidential framework for multi-task cell instance segmentation. Its segmentation head predicts Dirichlet concentration parameters, enabling single-pass, closed-form aleatoric, epistemic, and vacuity uncertainties at both pixel and instance level, while its centroid decoder is complemented with two lightweight geometric uncertainty cues that quantify localisation reliability without auxiliary models or sampling. Together, these evidential and geometric measures expose complementary failure modes and allow principled filtering of low-confidence nuclei. Across multi-tissue and multi-stain datasets, Evidential DualU-Net matches or surpasses deep ensembles in error separation at a fraction of the cost, maintains or improves calibration over deterministic baselines, and generalises across datasets without retuning. This work provides an interpretable and computationally practical uncertainty formulation for digital pathology. Code and weights are available at: https://github.com/davidanglada/Evidential-DualU-Net.

Keywords: Cell Instance Segmentation, Evidential Deep Learning, Uncertainty Estimation, Multitask Learning, Nuclei Segmentation

1. Introduction

Digital pathology has rapidly expanded with the adoption of whole-slide imaging and large annotated datasets, enabling computational models to support routine diagnostic workflows. Cell-level quantification—including nucleus detection, instance segmentation, and phenotype classification—is central to tasks such as Ki-67 assessment, immune profiling, and tumour microenvironment analysis, yet remains time-consuming and variable when performed manually. Fast and reliable automated systems are therefore essential. Within this context, DualU-Net (Anglada-Rotger et al., 2025), developed inside the DigiPatICS project (Temprana-Salvador et al., 2022) from the Institut Català de la Salut (ICS) of Catalunya, was designed as a lightweight, single-pass architecture capable of accurate and efficient cell instance segmentation.

Beyond accuracy, reliability is equally crucial: *a model should know what it does not know*. Uncertainty estimation is particularly important in digital pathology, where ambiguous morphology, heterogeneous staining, artefacts, and domain shifts routinely lead to failure cases that must be flagged for review. Crucially, uncertainty must be available *at inference time* without the computational burden of ensembles or sampling-based methods. Existing approaches for uncertainty estimation often require multiple forward passes, making them impractical for high-throughput clinical pipelines. Evidential Deep Learning (EDL) offers a promising alternative, providing closed-form aleatoric, epistemic, and vacuity estimates from a single prediction. However, prior work has focused almost exclusively on semantic segmentation; no existing method provides interpretable, instance-level evidential uncertainty or applies EDL to multi-task cell instance segmentation. Furthermore, current uncertainty methods for instance segmentation remain complex and computationally demanding, underscoring the need for simpler and more principled formulations.

Up to the authors knowledge, this work introduces the first evidential framework for multi-task cell instance segmentation. Our key contributions are: i) we extend DualU-Net with a Dirichlet-based evidential segmentation head and a multi-term loss, producing calibrated aleatoric, epistemic, and vacuity estimates at both pixel and instance level, while preserving the previous segmentation and classification performance; ii) we introduce simple, closed-form geometric uncertainty cues for the centroid decoder (peak and mass), enabling reliable detection-oriented uncertainty without auxiliary models or sampling; iii) through extensive evaluation on multi-tissue and multi-stain histopathology datasets, we show that the proposed evidential scheme matches or improves deep ensembles in error separation at a fraction of the computational cost, generalises across datasets without retuning, and yields interpretable uncertainty maps that clearly expose classification and detection issues relevant for digital pathology workflows.

2. Related Work

Cell instance segmentation: HoVer-Net (Graham et al., 2019) established the dominant three-decoder paradigm for nuclear instance segmentation by jointly predicting semantic masks, horizontal/vertical maps, and cell-type labels. Transformer-based adaptations such as CellViT (Hörst et al., 2024) and HistoNext (Chen et al., 2025) retain this multi-head structure while incorporating long-range contextual modeling to refine boundaries and improve classification accuracy. These models highlight the effectiveness of coupling semantic and detection cues for reliable cell delineation. In our previous work DualU-Net (Anglada-Rotger et al., 2025), the architecture is streamlined to only two decoders—one for semantic segmentation and another that regresses Gaussian centroid maps—whose outputs are combined via a watershed reconstruction. This reduction preserves the benefits of multi-task supervision while substantially improving architectural simplicity and computational efficiency, yet it follows the same deterministic formulation as prior models.

Uncertainty estimation and calibration: Predictive uncertainty in deep learning usually decomposes into *aleatoric* uncertainty, arising from intrinsic ambiguity in the data, and *epistemic* uncertainty, reflecting limited model knowledge or out-of-distribution behavior (Kendall and Gal, 2017). Estimating both components simultaneously remains difficult in many tasks. Multi-pass methods such as MC Dropout (Gal and Ghahramani, 2016) or deep ensembles (DE) (Lakshminarayanan et al., 2017) provide good approximations of epistemic uncertainty, with the latter shown to remain robust under distribution shift (Ovadia et al., 2020), but they are computationally expensive for day-to-day diagnostic workflows and do not yield explicit aleatoric estimates. Probabilistic segmentation frameworks such as Probabilistic U-Net (Kohl et al., 2018) or PhiSeg (Baumgartner et al., 2019) introduce latent sampling or generative priors and can capture ambiguity, yet they require multiple stochastic passes and are not well suited to densely packed nuclei. None of these approaches provide simple, closed-form estimates of both uncertainty types. Uncertainty has also been investigated for error prediction and active learning in biomedical imaging (Tan et al., 2025b; Anglada-Rotger et al., 2024), though most efforts remain in semantic or single-task settings.

Calibration is equally important, as cross-entropy-trained models often produce overconfident predictions. Post-hoc techniques such as temperature scaling (Guo et al., 2017) adjust confidence after training, while train-time strategies (e.g., MMCE (Kumar et al., 2018), focal-loss variants (Mukhoti et al., 2020) or BSCE-GRA (Lin et al., 2025)) aim to regularize confidence throughout optimization. Despite these advances, calibrated and instance-aware uncertainty estimation for multi-task cell segmentation remains under-explored.

Evidential Deep Learning (EDL): EDL introduces a probabilistic view of classification in which the network does not output a single categorical distribution, but instead predicts the parameters of a *distribution over* categorical distributions. In a standard setting, a categorical likelihood for an input x with class probabilities $\mathbf{p} = (p_1, \ldots, p_K)$ is $p(y = k \mid \mathbf{p}) = p_k$, with $\mathbf{p}$ typically produced by a softmax layer. EDL generalizes this by placing a Dirichlet prior over $\mathbf{p}$. Following Sensoy et al. (Sensoy et al., 2018), the network outputs non-negative evidence values e_k, which define concentration parameters $\alpha_k = e_k + 1$ of a Dirichlet distribution $D(\mathbf{p} \mid \boldsymbol{\alpha})$. The predictive probabilities are given by the Dirichlet mean (see Section 3). The Dirichlet formulation allows uncertainty to be read directly from the predicted parameters $\boldsymbol{\alpha}$. The total evidence $S = \sum_k \alpha_k$ reflects how strongly the model supports its prediction: when S is small, the Dirichlet distribution is broad, indicating that the model has not accumulated enough evidence to commit to any class. This behaviour is captured by vacuity, which represents uncertainty due purely to a lack of support in the data. In contrast, the spread of the Dirichlet around its mean captures the remaining uncertainty and gives rise to analytic measures of aleatoric and epistemic uncertainty. All these quantities are obtained in closed form, allowing EDL to produce calibrated uncertainty estimates from a single forward pass without sampling or ensembles. Training encourages the model to increase evidence when predictions are correct and suppress it when they are wrong, preventing unwarranted confidence.

EDL has also been explored in semantic segmentation. In (Ancha et al., 2024) evidential models are applied to pixelwise OOD-aware segmentation. EDL has been also used in several biomedical tasks, such as semantic segmentation (Tan et al., 2025a), uncertainty-guided 3D mitochondria segmentation (Shi et al., 2024), interpretable evidential uncertainty supervision (Li et al., 2025), or semi-supervised segmentation via mutual evidential learning (He et al., 2025). These works demonstrate growing interest in evidential segmentation, but they remain limited to single-task semantic settings: none provide interpretable uncertainty at the instance level, nor do they extend evidential modeling to multi-task formulations.

3. Materials and Methods

Datasets. We evaluate on two annotated histopathology datasets. PanNuke (Gamper et al., 2020) provides 7904 H&E patches (256×256) across 19 tissues with 189k nuclei labeled into five classes. We also use a proprietary breast Ki-67 IHC dataset (Anglada-Rotger et al., 2024) with 52 tiles (1024×1024) from four patients, each containing pixel-level nuclei masks and three-class labels (positive, negative, non-epithelial).

Evidential segmentation head and loss. DualU-Net (Anglada-Rotger et al., 2025) contains two decoders: a semantic segmentation head and a centroid-regression head. We keep this architecture but replace the segmentation logits with Dirichlet evidence. For each pixel x, the segmentation decoder outputs non-negative evidence values $e_k(x) \geq 0$, which define the Dirichlet concentration parameters $\alpha_k(x) = e_k(x) + 1$, $\boldsymbol{\alpha}(x) = (\alpha_1(x), \ldots, \alpha_K(x))$, the predictive class probabilities are given by the Dirichlet mean $\hat{p}_k(x) = \frac{\alpha_k(x)}{S(x)}, S(x) = \sum_{j=1}^{K} \alpha_j(x)$. The predictive categorical distribution at pixel x is defined as $\hat{\mathbf{p}}(x) = (\hat{p}_1(x), \cdots, \hat{p}_K(x))$. Following (Sensoy et al., 2018), the evidential loss combines a data-fitting term encouraging $\hat{\mathbf{p}}(x)$ to match the one-hot label $\mathbf{y}(x)$ with a KL regularizer that discourages unwarranted evidence. To penalize evidence for incorrect classes while leaving the correct class unpenalized, we construct the modified Dirichlet parameter vector $\tilde{\boldsymbol{\alpha}}(x) = \big(\tilde{\alpha}_1(x), \ldots, \tilde{\alpha}_K(x)\big)$, where each component is defined as

$$\tilde{\alpha}_k(x) = \begin{cases} 1, & \text{if } k = y(x), \\ \alpha_k(x), & \text{otherwise.} \end{cases} \tag{1}$$

This way, the per-pixel segmentation loss is

$$\mathcal{L}_{\text{EDL}}^{\text{seg}}(x) = \|\mathbf{y}(x) - \hat{\mathbf{p}}(x)\|_2^2 + \lambda_{KL}\,\mathrm{KL}\Big(D(\mathbf{p} \mid \tilde{\boldsymbol{\alpha}}(x)) \;\Big\|\; D(\mathbf{p} \mid \mathbf{1})\Big), \tag{2}$$

As shown in (Tan et al., 2025b), incorporating a Dice term improves the optimization dynamics of evidential semantic segmentation. For this reason, all our experiments include an additional Dice component. In the original DualU-Net (Anglada-Rotger et al., 2025), the Dice was class-weighted to mitigate strong label imbalance; however, such weighting is uncommon in EDL frameworks. We therefore evaluate two variants: (i) standard (unweighted) Dice and (ii) class-weighted Dice. The centroid decoder and its regression objective remain unchanged from the original DualU-Net. The full training objective is

$$\mathcal{L} = \lambda_{seg}\mathcal{L}_{\text{EDL}}^{\text{seg}} + \lambda_{dice}\,\mathcal{L}_{\text{Dice}} + \lambda_{cent}\,\mathcal{L}_{\text{cent}}, \tag{3}$$

Segmentation-head evidential uncertainty. Let $\mathcal{D}$ be the training dataset and $\hat{y}$ the categorical prediction at pixel x, modeled as a random variable $\hat{y} \sim \mathrm{Cat}(\mathbf{p})$ where $\mathbf{p}$ is drawn from the Dirichlet distribution $D(\mathbf{p} \mid \boldsymbol{\alpha}(x))$. For a Bayesian classifier with Dirichlet-distributed class probabilities $\mathbf{p} \sim D(\mathbf{p} \mid \boldsymbol{\alpha}(x))$, as in (Kendall and Gal, 2017; Tan et al., 2025a), we use $u_{\text{ale}}(x) = \mathbb{E}_{\text{Dir}}\big[\mathrm{Var}_{\text{Cat}}\big(\hat{y} \mid \mathbf{p}\big)\big]$, $u_{\text{epi}}(x) = \mathrm{Var}_{\text{Dir}}\big[\mathbb{E}_{\text{Cat}}\big[\hat{y} \mid \mathbf{p}\big]\big]$

For the Dirichlet prior, it admits the following closed forms (see Appendix A):

$$u_{\text{ale}}(x) = \sum_{k=1}^{K} \frac{\alpha_k(x)\big(S(x) - \alpha_k(x)\big)}{S(x)\big(S(x)+1\big)}, \qquad u_{\text{epi}}(x) = \sum_{k=1}^{K} \frac{\alpha_k(x)\big(S(x) - \alpha_k(x)\big)}{S^2(x)\big(S(x)+1\big)}. \tag{4}$$

A third quantity naturally arises in evidential models: vacuity. While aleatoric and epistemic uncertainties separate noise from model uncertainty, vacuity measures the absence of evidence accumulated from the data $u_{\text{vac}}(x) = \frac{K}{S(x)}$.

Cell analysis requires uncertainty not only at the pixel level but also at the instance level, since downstream evaluation (detection F1, classification F1) and clinical interpretation are performed per nucleus rather than per pixel. Instance masks Ω_i are obtained with the same watershed reconstruction as in DualU-Net. For each instance, we aggregate evidential parameters by averaging:

$$\bar{\alpha}_k^{(i)} = \frac{1}{|\Omega_i|} \sum_{x \in \Omega_i} \alpha_k(x), \qquad \bar{S}^{(i)} = \sum_{k=1}^{K} \bar{\alpha}_k^{(i)}. \tag{5}$$

At the pixel level, all Dirichlet parameters—including the background class—contribute to uncertainty because they shape the full predictive distribution. However, for instance-level uncertainty we are interested only in the reliability of the classification of a segmented nucleus. Therefore, when computing instance-level uncertainty, we exclude the background component from $\bar{\boldsymbol{\alpha}}^{(i)}$ and renormalize over the $K-1$ foreground classes. This ensures that $u_{\text{ale}}(\Omega_i)$, $u_{\text{epi}}(\Omega_i)$, and $u_{\text{vac}}(\Omega_i)$ quantify uncertainty about the nucleus class, not about residual background evidence. Substituting the resulting foreground-only $\bar{\boldsymbol{\alpha}}^{(i)}$ into u_{epi}, u_{ale}, u_{vac} yields instance-level $u_{\text{ale}}(\Omega_i)$, $u_{\text{epi}}(\Omega_i)$, and $u_{\text{vac}}(\Omega_i)$. To make all uncertainty quantities directly comparable and easily interpretable, we normalize u_{ale}, u_{epi}, and u_{vac} to the range $[0, 1]$. Each expression admits a closed-form theoretical minimum and maximum determined by the Dirichlet parameters $\boldsymbol{\alpha}$ (see Appendix B). For each uncertainty type, we compute its attainable bounds and apply an affine normalization.

Centroid-head uncertainty. While Heteroscedastic Uncertainty (Kendall and Gal, 2017) provides a standard probabilistic framework for regression by minimizing the Gaussian Negative Log Likelihood (NLL), we explicitly opt for a geometric approach. In sparse centroid regression, the NLL objective is not only prone to optimization instability due to class imbalance, but it also strictly models pixel-intensity noise. In contrast, our proposed geometric reliability measures target structural failures.

Let $g : \mathcal{X} \to [0, \infty)$ denote the Gaussian density map predicted by the centroid decoder, where $g(x)$ is the value at pixel x. For each reconstructed nucleus instance $\Omega_i \subset \mathcal{X}$, assumed to arise from an isotropic Gaussian with standard deviation σ, the ideal density integrates to the analytic mass $G_{\max} = 2\pi\sigma^2$. Departures of g from this template reflect unreliable centroid localisation. We extract two complementary geometric cues: (i) *Peak uncertainty*, which assesses the sharpness of the predicted Gaussian by the maximum value $p_{\max}^{(i)} = \max_{x \in \Omega_i} g(x)$; diffuse or weak responses indicate uncertain detections. We define

$$u_{\text{peak}}(\Omega_i) = 1 - p_{\max}^{(i)}. \tag{6}$$

(ii) *Mass-ratio uncertainty*, which measures energy preservation. Let $m_{\text{pred}}^{(i)} = \sum_{x \in \Omega_i} g(x)$ denote the predicted mass; deviations from $G_{\max}$ are quantified symmetrically as

$$u_{\text{mass}}(\Omega_i) = \frac{|m_{\text{pred}}^{(i)} - G_{\max}|}{G_{\max}}. \tag{7}$$

Values near zero correspond to correct centroid strength, whereas large deviations signal missing, diffuse, or overly dominant Gaussian responses. These two cues provide simple and direct measures of centroid reliability for each nucleus. A single scalar uncertainty value is obtained via a linear combination $u_{\text{cent}}(\Omega_i) = \lambda_{\text{peak}}\, u_{\text{peak}}(\Omega_i) + \lambda_{\text{mass}}\, u_{\text{mass}}(\Omega_i),$.

Table 1: Quantitative uncertainty evaluation on PanNuke and Ki-67. *Left:* segmentation-head uncertainty and calibration results (EDL head) compared with Deep Ensembles (DE) and the deterministic DualU-Net baseline. *Right:* centroid-head uncertainty results. Complete centroid histograms and eCDF plots in Appendix D.

M	**UM**	**ACE ↓**	**MCE ↓**	**A-UCE ↓**	**M-UCE ↓**	**KS ↑**	**AUROC ↑**
				PanNuke			
Ours	u_{ale}	$\mathbf{0.061}_{\pm 0.004}$	$0.289_{\pm 0.010}$	$0.157_{\pm 0.010}$	$0.326_{\pm 0.025}$	$0.392_{\pm 0.003}$	$0.759_{\pm 0.003}$
	u_{epi}	$\mathbf{0.061}_{\pm 0.004}$	$0.289_{\pm 0.010}$	$0.100_{\pm 0.004}$	$0.251_{\pm 0.017}$	$0.392_{\pm 0.003}$	$0.759_{\pm 0.003}$
	u_{vac}	$\mathbf{0.061}_{\pm 0.004}$	$0.289_{\pm 0.010}$	$\mathbf{0.054}_{\pm 0.004}$	$0.246_{\pm 0.016}$	$0.391_{\pm 0.003}$	$0.758_{\pm 0.003}$
Ours w	u_{ale}	$0.095_{\pm 0.003}$	$0.383_{\pm 0.005}$	$0.175_{\pm 0.005}$	$0.382_{\pm 0.008}$	$0.442_{\pm 0.005}$	$0.791_{\pm 0.002}$
	u_{epi}	$0.095_{\pm 0.003}$	$0.383_{\pm 0.005}$	$0.113_{\pm 0.002}$	$0.333_{\pm 0.002}$	$\mathbf{0.442}_{\pm 0.005}$	$\mathbf{0.796}_{\pm 0.003}$
	u_{vac}	$0.095_{\pm 0.003}$	$0.383_{\pm 0.005}$	$0.080_{\pm 0.003}$	$0.321_{\pm 0.003}$	$0.441_{\pm 0.005}$	$0.796_{\pm 0.003}$
Base	u_s	$0.234_{\pm 0.004}$	$0.417_{\pm 0.027}$	$0.198_{\pm 0.004}$	$0.353_{\pm 0.027}$	$0.287_{\pm 0.016}$	$0.692_{\pm 0.010}$
DE		$0.131_{\pm 0.001}$	$\mathbf{0.220}_{\pm 0.019}$	$0.085_{\pm 0.001}$	$\mathbf{0.159}_{\pm 0.013}$	$0.344_{\pm 0.006}$	$0.721_{\pm 0.003}$
				Ki67			
Ours	u_{ale}	$\mathbf{0.106}_{\pm 0.048}$	$\mathbf{0.161}_{\pm 0.040}$	$0.217_{\pm 0.080}$	$0.287_{\pm 0.069}$	$0.452_{\pm 0.147}$	$0.786_{\pm 0.088}$
	u_{epi}	$\mathbf{0.106}_{\pm 0.048}$	$\mathbf{0.161}_{\pm 0.040}$	$\mathbf{0.096}_{\pm 0.046}$	$\mathbf{0.173}_{\pm 0.068}$	$0.450_{\pm 0.146}$	$0.787_{\pm 0.088}$
	u_{vac}	$\mathbf{0.106}_{\pm 0.048}$	$\mathbf{0.161}_{\pm 0.040}$	$0.111_{\pm 0.036}$	$0.175_{\pm 0.027}$	$0.446_{\pm 0.148}$	$0.786_{\pm 0.089}$
Ours w	u_{ale}	$0.132_{\pm 0.053}$	$0.220_{\pm 0.056}$	$0.201_{\pm 0.086}$	$0.258_{\pm 0.078}$	$0.470_{\pm 0.156}$	$0.796_{\pm 0.090}$
	u_{epi}	$0.132_{\pm 0.053}$	$0.220_{\pm 0.056}$	$0.122_{\pm 0.095}$	$0.222_{\pm 0.123}$	$\mathbf{0.471}_{\pm 0.157}$	$\mathbf{0.796}_{\pm 0.090}$
	u_{vac}	$0.132_{\pm 0.053}$	$0.220_{\pm 0.056}$	$0.131_{\pm 0.059}$	$0.195_{\pm 0.086}$	$0.471_{\pm 0.159}$	$0.796_{\pm 0.090}$
Base	u_s	$0.286_{\pm 0.153}$	$0.430_{\pm 0.120}$	$0.207_{\pm 0.144}$	$0.226_{\pm 0.119}$	$0.252_{\pm 0.113}$	$0.663_{\pm 0.071}$
DE		$0.159_{\pm 0.120}$	$0.283_{\pm 0.179}$	$0.112_{\pm 0.065}$	$0.252_{\pm 0.108}$	$0.311_{\pm 0.126}$	$0.690_{\pm 0.076}$

M	**UM**	**KS ↑**	**AUROC ↑**
		PanNuke	
Ours	u_{cent}	$0.429_{\pm 0.006}$	$0.782_{\pm 0.003}$
	u_{mass}	$0.410_{\pm 0.005}$	$0.767_{\pm 0.003}$
	u_{peak}	$0.338_{\pm 0.025}$	$0.712_{\pm 0.016}$
Ours w	u_{cent}	$\mathbf{0.461}_{\pm 0.010}$	$\mathbf{0.801}_{\pm 0.009}$
	u_{mass}	$0.448_{\pm 0.007}$	$0.787_{\pm 0.009}$
	u_{peak}	$0.361_{\pm 0.015}$	$0.723_{\pm 0.008}$
		Ki67	
Ours	u_{cent}	$0.591_{\pm 0.113}$	$0.862_{\pm 0.058}$
	u_{mass}	$0.575_{\pm 0.121}$	$0.851_{\pm 0.063}$
	u_{peak}	$0.520_{\pm 0.096}$	$0.823_{\pm 0.052}$
Ours w	u_{cent}	$\mathbf{0.612}_{\pm 0.092}$	$\mathbf{0.875}_{\pm 0.047}$
	u_{mass}	$0.596_{\pm 0.099}$	$0.863_{\pm 0.056}$
	u_{peak}	$0.543_{\pm 0.058}$	$0.843_{\pm 0.033}$

Two uncertainties for two error types. For each nucleus Ω_i, our method outputs two complementary uncertainty families. Segmentation-head evidential uncertainties $\big(u_{\text{epi}}(\Omega_i)$, $u_{\text{ale}}(\Omega_i)$, $u_{\text{vac}}(\Omega_i)\big)$ reflect ambiguity in the class distribution and are therefore linked to *classification* errors. Centroid-based geometric scores $\big(u_{\text{cent}}(\Omega_i)$, $u_{\text{peak}}(\Omega_i)$, $u_{\text{mass}}(\Omega_i)\big)$ capture the sharpness and stability of the predicted Gaussian response, making them indicative of *detection* errors. Together, they offer complementary, instance-level reliability signals.

4. Results

Experiments and implementation details. We follow PanNuke three-fold cross-validation and Ki-67 leave-one-patient-out cross-validation. Training uses 100 epochs with constant learning rates (2×10^{-4} for PanNuke, 1×10^{-4} for Ki-67) and batch sizes 64 and 8 respectively. Centroid uncertainty uses fixed weights $\lambda_{\text{mass}} = 0.6$ and $\lambda_{\text{peak}} = 0.3$ when forming the final combined score U_{cent}. We include two baselines: the original DualU-Net (Shannon-entropy uncertainty) and a ten-model deep ensemble (entropy on mean predictions). For the *centroid head*, no established baselines exist, and we simply evaluate the proposed geometric cues. We consider two evidential variants: *Ours* with unweighted Dice and loss weights $\lambda_{\text{seg}} = 1$, $\lambda_{\text{dice}} = 0.4$, $\lambda_{\text{cent}} = 0.7$, $\lambda_{\text{kl}} = 0.4$, and *Ours w* with class-weighted Dice and $\lambda_{\text{kl}} = 0.2$, both using a 40-epoch warm-up for λ_{kl}. All hyperparameters have been selected on PanNuke validation folds and reused on Ki-67 without further tuning. *Ours w* preserves DualU-Net accuracy: on PanNuke, Detection F1 remains $0.80 \to 0.81$, mean Class F1 $0.54 \to 0.52$, and Dice $0.76 \to 0.76$; on Ki-67, Detection F1 stays 0.82, mean Class F1 $0.57 \to 0.58$, and Dice $0.86 \to 0.83$.

Evaluation metrics. We evaluate uncertainty quality using Adaptive Calibration Error (ACE) (Nixon et al., 2019) and its maximum (MCE), as well as Adaptive UCE (A-UCE) and its maximum (M-UCE) using quantile-based binning (Laves et al., 2019). Error–uncertainty separability is quantified using the Kolmogorov–Smirnov (KS) statistic (Tan et al., 2025c) and AUROC, computed between continuous uncertainty values and binary correctness indicators. Calibration metrics (ACE, MCE, A-UCE, M-UCE) are reported only for the segmentation head, whose evidential formulation yields probabilistic class predictions. For the centroid head, uncertainty derives from geometric cues rather than calibrated probabilities; accordingly, only KS and AUROC are evaluated, as these measure how well uncertainty ranks correct versus incorrect detections.

Segmentation uncertainty. Table 1 (left) reports calibration and error-separation metrics for all segmentation-head uncertainty measures. Across both datasets, the evidential formulation (*Ours* and *Ours w*) consistently increases the separation between correct and incorrect predictions. On PanNuke, all three evidential uncertainties ($u_{\text{ale}}, u_{\text{epi}}, u_{\text{vac}}$) achieve substantially higher KS and AUROC than *base*; two-sample t-tests yield extremely small p-values ($p < 10^{-6}$). Compared to DE, evidential uncertainties also show higher KS and AUROC on average, but the differences are not statistically significant ($p > 0.05$). The three evidential uncertainties behave similarly, and pairwise tests show no statistically significant differences between them ($p > 0.1$). Both uncertainty distribution histogram and eCDF plots reveal a clearer separation between error and non-error instances for evidential measures than for baselines (Figure 1). The weighted-Dice variant (*Ours w*) further improves error separation on PanNuke, raising KS from $0.392 \rightarrow 0.442$ and AUROC from $0.759 \rightarrow 0.796$; two-sample t-tests indicate that these improvements are statistically significant ($p < 0.05$). On Ki-67, evidential uncertainties (*Ours* and *Ours w*) again outperform both baselines in KS and AUROC, and additional histogram and eCDF analyses for this dataset are provided in Appendix C. As on PanNuke, the three uncertainty types remain statistically indistinguishable ($p > 0.1$), and the weighted-Dice variant shows a consistent but not significant improvement relative to the unweighted version.

Segmentation calibration. Across both datasets, evidential models clearly improve calibration over the deterministic baseline while essentially matching the behaviour of deep ensembles. On PanNuke, ACE is reduced from 0.236 (Base) to 0.095 (*Ours w*) and to 0.061 (*Ours*), with all improvements over the baseline strongly significant ($p < 10^{-4}$). Differences between EDL and DE are not statistically significant ($p > 0.1$), indicating that evidential training attains ensemble-level calibration. For MCE, *Ours* significantly improves over the baseline ($p < 10^{-3}$), whereas *Ours w* shows a smaller but still meaningful reduction ($p \approx 0.03$); DE remains the best in this metric (see Figure 2). A-UCE and M-UCE follow the same pattern. On PanNuke, EDL notably reduces A-UCE relative to the baseline (e.g. vacuity: $0.198 \rightarrow 0.054$, $p < 10^{-3}$), achieving values comparable to DE ($p > 0.1$). On Ki-67, higher variance prevents statistical separation between methods ($p > 0.15$), but evidential calibration remains at least as good as DE and better than the deterministic baseline.

Centroid uncertainty. Table 1 (right) reports KS and AUROC for the centroid-head uncertainties ($u_{\text{peak}}, u_{\text{mass}}, u_{\text{cent}}$). On PanNuke, centroid cues show strong separation between correct and incorrect detections: the combined score u_{cent} reaches $\text{KS} = 0.429 \pm 0.006$ and $\text{AUROC}_{\text{err}} = 0.782 \pm 0.003$ for *Ours*, increasing to $\text{KS} = 0.461 \pm 0.010$ and $\text{AUROC}_{\text{err}} = 0.801 \pm 0.009$ for *Ours w*. The KS improvement is statistically significant ($p \approx 0.03$), whereas AUROC differences fall within fold variability ($p > 0.1$). Among the individual cues, u_{mass} is the most informative ($\text{KS} \approx 0.41 \rightarrow 0.45$), u_{peak} performs slightly worse ($\text{KS} \approx 0.34 \rightarrow 0.36$),

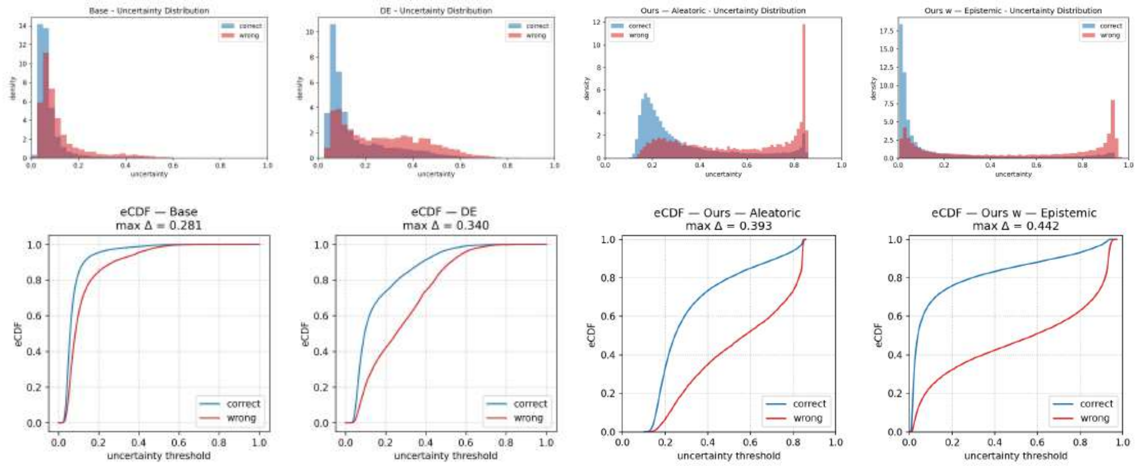

Figure 1: Segmentation-head uncertainty histograms (top) and eCDFs (bottom). Errors in red, correct instances in blue. Columns: Base, DE, *Ours*, *Ours w*. For evidential models we plot the best separator (u_{ale} for *Ours*, u_{epi} for *Ours w*). See additional histogram and eCDF analyses in Appendix C

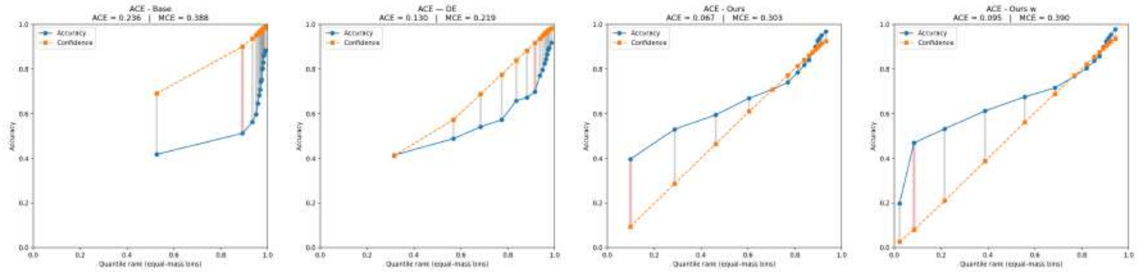

Figure 2: ACE plots for the segmentation head. Left to right: Base, DE, *Ours*, *Ours w*.

and their combination u_{cent} yields the strongest signal. On Ki-67, centroid uncertainties are even more discriminative. Both variants achieve high separation (u_{cent}: KS $= 0.591 \pm 0.113$ and 0.612 ± 0.092; AUROC$_{\text{err}}$ $= 0.862 \pm 0.058$ and 0.875 ± 0.047 for *Ours* and *Ours w*, respectively), but the two models are statistically indistinguishable across all centroid metrics ($p > 0.1$). As in PanNuke, u_{mass} retains the strongest single-feature performance, with u_{peak} slightly lower but still clearly informative.

Qualitative results. Figure 3 illustrates qualitative examples from a representative PanNuke fold using the *Ours w* configuration. Across the examples, nuclei highlighted with high segmentation-head or centroid-head uncertainty consistently correspond to meaningful failure modes: clear classification mistakes, missed or imprecise detections, or instances that, despite being labeled as correct, exhibit ambiguous morphology or borderline staining and could warrant ground-truth revision.

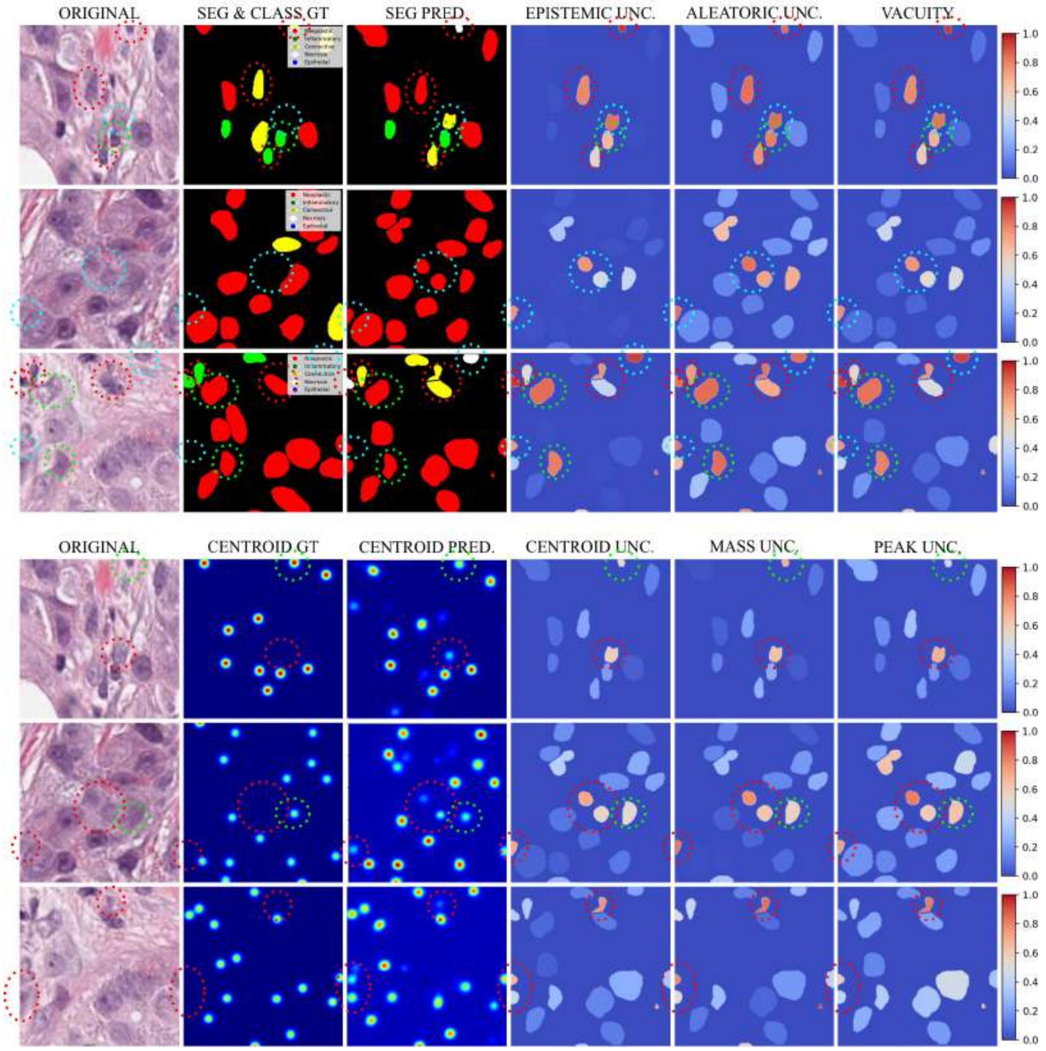

Figure 3: Qualitative uncertainty examples on PanNuke using the *Ours w* configuration. We show the original patch, ground-truth labels, predictions, and the three uncertainty measures for each head (u_{epi}, u_{ale}, u_{vac} for segmentation and u_{cent}, u_{mass}, u_{peak} for detection). For segmentation, red circles mark class-mismatch errors, blue false-positive nuclei, and green correctly predicted but ambiguous cases. For centroids, red circles highlight missed or imprecise detections, and green circles indicate correct detections with residual uncertainty. These examples illustrate how segmentation- and centroid-based uncertainties jointly identify unreliable instances.

5. Discussion and Conclusions

We introduced an evidential formulation of DualU-Net that provides, in a single forward pass, two complementary uncertainty families: segmentation-driven evidential uncertainty (aleatoric, epistemic, vacuity) targeting classification errors, and centroid-derived geometric uncertainty (peak and mass) targeting detection and localisation errors. Together, they offer a unified decomposition of instance-level reliability that aligns with the two dominant failure modes in cell instance segmentation.

Across PanNuke and Ki-67, the evidential scheme consistently outperforms the deterministic baseline and matches or surpasses DE in error separation, while being substantially more efficient. Although the three segmentation-head uncertainties differ qualitatively—aleatoric tending to produce higher intensities, epistemic and vacuity spanning wider dynamic ranges (Figure 1 and Appendix C)—their quantitative behaviour is statistically indistinguishable in terms of error discrimination (Table 1). In PanNuke, this alignment is visually evident (Figure 3): all three uncertainties assign high values to the same problematic nuclei, highlighting classification mistakes, false positives, or low-confidence predictions that merit inspection.

In addition, EDL improves mean calibration, aligning confidence with correctness (Table 1). Under-confidence at low predicted probabilities arises naturally: when evidence is small, vacuity dominates and the Dirichlet mean drifts toward the uniform prior, reducing confidence even for correct predictions (Figure 2). The weighted Dice variant improves error separation by sharpening gradients on difficult pixels, but this also amplifies the KL penalty in low-evidence regions, yielding more diffuse Dirichlet outputs and thus poorer calibration in the lowest-confidence bins.

For the centroid head, the proposed Gaussian-based uncertainty measures are simple, interpretable, and computationally free at inference. Mass-based uncertainty is consistently the strongest cue, while peak uncertainty provides complementary information ; their combination yields the best KS and AUROC values (Table 1). Qualitative examples confirm that high-uncertainty instances correspond to misdetections, poor localisations, or ambiguous annotations, demonstrating the practical interpretability of these geometric cues (Figure 3).

Importantly, all hyperparameters optimised on PanNuke transfer directly to Ki-67 without re-tuning, highlighting the cross-dataset generalisation of the evidential framework and its robustness under domain shift. The ability to surface uncertainty at inference time enables model introspection for pathologists and supports downstream applications such as active learning, quality control of annotations, and uncertainty-aware dataset curation.

To our knowledge, this is the first evidential instance segmentation model in a multi-task setting for digital pathology, demonstrating both methodological and practical value. Future work includes extending evidential modelling to centroid regression via Normal–Inverse–Gamma uncertainty (Amini et al., 2019), enabling a fully evidential DualU-Net architecture.

Acknowledgments

This publication is part of the R&D&I project PID2023-148614OB-I00, funded by MICIU/AEI/10.13039/501100011033/ and by FEDER, EU. This research has also been funded by European Development Funds Regional, Programa operatiu FEDER de Catalunya 2014-2020 through the project DigiPatICS.

References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *CoRR*, abs/1910.02600, 2019. URL http://arxiv.org/abs/1910.02600.

Siddharth Ancha, Philip R. Osteen, and Nicholas Roy. Deep evidential uncertainty estimation for semantic segmentation under out-of-distribution obstacles. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6943–6951, 2024. doi: 10.1109/ICRA57147.2024.10611342.

David Anglada-Rotger, Julia Sala, Ferran Marques, Philippe Salembier, and Montse Pardàs. Enhancing ki-67 cell segmentation with dual u-net models: A step towards uncertainty-informed active learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5026–5035, 2024.

David Anglada-Rotger, Berta Jansat, Ferran Marques, and Montse Pardàs. Two heads are enough: Dualu-net, a fast and efficient architecture for nuclei instance segmentation. In *Medical Imaging with Deep Learning*, 2025. URL https://openreview.net/forum?id=1KOCklgxQd.

Christian F. Baumgartner, Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation, 2019. URL https://arxiv.org/abs/1906.04045.

J. Chen, R. Wang, W. Dong, et al. Histonext: dual-mechanism feature pyramid network for cell nuclear segmentation and classification. *BMC Medical Imaging*, 25(9), 2025. doi: 10.1186/s12880-025-01550-2. URL https://doi.org/10.1186/s12880-025-01550-2.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. 33rd International Conf. on Machine Learning*, pages 1050–1059, 2016.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020.

Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58, 101563, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

Yuanpeng He, Yali Bi, Lijian Li, Chi-Man Pun, Wenpin Jiao, and Zhi Jin. Mutual evidential deep learning for medical image segmentation, 2025. URL https://arxiv.org/abs/2505.12418.

F. Hörst, M. Rempe, L. Heine, C. Seibold, J. Keyl, G. Baldini, S. Ugurel, Je. Siveke, B. Grünwald, Jan E., and J. Kleesiek. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103143. URL https://www.sciencedirect.com/science/article/pii/S1361841524000689.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *CoRR*, abs/1806.05034, 2018. URL http://arxiv.org/abs/1806.05034.

A. Kumar, S. Sarawagi, and U. Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the International Conference on Machine Learning*, pages 2805–2814, 2018.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *CoRR*, abs/1909.13550, 2019. URL http://arxiv.org/abs/1909.13550.

Yuzhu Li, An Sui, Fuping Wu, and Xiahai Zhuang. *Uncertainty-Supervised Interpretable and Robust Evidential Segmentation*, page 649–658. Springer Nature Switzerland, September 2025. ISBN 9783032051851. doi: 10.1007/978-3-032-05185-1_62. URL http://dx.doi.org/10.1007/978-3-032-05185-1_62.

Jinxu Lin, Linwei Tao, Minjing Dong, and Chang Xu. Uncertainty–weighted gradients for model calibration. *CVPR*, 2025.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. Calibrating deep neural networks using focal loss. *CoRR*, abs/2002.09437, 2020. URL https://arxiv.org/abs/2002.09437.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Yaniv Ovadia, Emily Fertig, and Daisy et al. Ren. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 33, 2020.

Murat Sensoy, Melih Kandemir, and Lance M. Kaplan. Evidential deep learning to quantify classification uncertainty. *CoRR*, abs/1806.01768, 2018. URL http://arxiv.org/abs/1806.01768.

Ruohua Shi, Lingyu Duan, Tiejun Huang, and Tingting Jiang. Evidential uncertainty-guided mitochondria segmentation for 3d em images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4847–4855, Mar. 2024. doi: 10.1609/aaai.v38i5.28287. URL https://ojs.aaai.org/index.php/AAAI/article/view/28287.

Hai Siong Tan, Kwancheng Wang, and Rafe McBeth. Uncertainty-error correlations in evidential deep learning models for biomedical segmentation. In Wei-Ta Chu, Chih-Ya Shen, and Hong-Han Shuai, editors, *Technologies and Applications of Artificial Intelligence*, pages 91–105, Singapore, 2025a. Springer Nature Singapore. ISBN 978-981-96-4589-3.

Hai Siong Tan, Kwancheng Wang, and Rafe McBeth. Uncertainty-error correlations in evidential deep learning models for biomedical segmentation. In *Technologies and Applications of Artificial Intelligence*, pages 91–105. Springer, 2025b.

Hai Siong Tan, Kwancheng Wang, and Rafe McBeth. Uncertainty-error correlations in evidential deep learning models for biomedical segmentation. In *Technologies and Applications of Artificial Intelligence*, pages 91–105, Singapore, 2025c. Springer Nature Singapore. ISBN 978-981-96-4589-3.

J Temprana-Salvador, P López-García, J Castellví, et al. Digipatics: Digital pathology transformation of the catalan health institute network of 8 hospitals-planification, implementation, and preliminary results. *Diagnostics (Basel)*, 12(4):852, 2022. doi: 10.3390/diagnostics12040852.

Appendix A. Closed-Form Evidential Uncertainty for the Segmentation Head

Dirichlet–Categorical evidential model. For each pixel x, the segmentation head predicts Dirichlet concentration parameters $\boldsymbol{\alpha}(x) = (\alpha_1(x), \ldots, \alpha_K(x))$ with total evidence $S(x) = \sum_k \alpha_k(x)$. These induce a Dirichlet distribution $\mathbf{p}(x) \sim D(\mathbf{p} \mid \boldsymbol{\alpha}(x))$ over class probabilities, and a categorical prediction $\hat{y}(x) \sim \mathrm{Cat}(\mathbf{p}(x))$. Aleatoric and epistemic uncertainties are defined as

$$u_{\mathrm{ale}}(x) = \mathbb{E}_{\mathrm{Dir}}\left[\mathrm{Var}_{\mathrm{Cat}}\left(\hat{y} \mid \mathbf{p}\right)\right], \qquad u_{\mathrm{epi}}(x) = \mathrm{Var}_{\mathrm{Dir}}\left[\mathbb{E}_{\mathrm{Cat}}\left[\hat{y} \mid \mathbf{p}\right]\right]. \tag{8}$$

Closed-form expressions. For completeness, we derive the closed forms in Eq. (11) starting from the definitions in Section 3. Let $\boldsymbol{\alpha}(x) = (\alpha_1(x), \ldots, \alpha_K(x))$ and $S(x) = \sum_{k=1}^{K} \alpha_k(x)$, and denote p_k the k-th component of $\mathbf{p}$.

- *Aleatoric uncertainty.* For a categorical variable with one-hot encoding, the conditional variance given $\mathbf{p}$ is

  $$\mathrm{Var}_{\mathrm{Cat}}(\hat{y} \mid \mathbf{p}) = \sum_{k=1}^{K} p_k(1-p_k) = \sum_{k=1}^{K} \big(p_k - p_k^2\big).$$

  Taking the expectation under the Dirichlet prior,

  $$u_{\mathrm{ale}}(x) = \mathbb{E}_{\mathrm{Dir}}\Big[\mathrm{Var}_{\mathrm{Cat}}(\hat{y} \mid \mathbf{p})\Big] = \sum_{k=1}^{K} \Big(\mathbb{E}[p_k] - \mathbb{E}[p_k^2]\Big).$$

  Using standard Dirichlet moments,

  $$\mathbb{E}[p_k] = \frac{\alpha_k(x)}{S(x)}, \qquad \mathbb{E}[p_k^2] = \frac{\alpha_k(x)\big(\alpha_k(x)+1\big)}{S(x)\big(S(x)+1\big)},$$

  we obtain

  $$\mathbb{E}[p_k] - \mathbb{E}[p_k^2] = \frac{\alpha_k(x)}{S(x)} - \frac{\alpha_k(x)\big(\alpha_k(x)+1\big)}{S(x)\big(S(x)+1\big)} = \frac{\alpha_k(x)\big(S(x)-\alpha_k(x)\big)}{S(x)\big(S(x)+1\big)}.$$

  Summing over k gives

  $$u_{\mathrm{ale}}(x) = \sum_{k=1}^{K} \frac{\alpha_k(x)\big(S(x)-\alpha_k(x)\big)}{S(x)\big(S(x)+1\big)}. \tag{9}$$

- *Epistemic uncertainty.* The epistemic term is defined as the variability (under the Dirichlet prior) of the mean categorical prediction:

  $$u_{\mathrm{epi}}(x) = \mathrm{Var}_{\mathrm{Dir}}\Big[\mathbb{E}_{\mathrm{Cat}}[\hat{y} \mid \mathbf{p}]\Big] = \sum_{k=1}^{K} \mathrm{Var}_{\mathrm{Dir}}(p_k),$$

  where we sum the component-wise variances of p_k. For the Dirichlet,

  $$\mathrm{Var}_{\mathrm{Dir}}(p_k) = \mathbb{E}[p_k^2] - \mathbb{E}[p_k]^2 = \frac{\alpha_k(x)\big(\alpha_k(x)+1\big)}{S(x)\big(S(x)+1\big)} - \left(\frac{\alpha_k(x)}{S(x)}\right)^2.$$

  Bringing to a common denominator $S(x)^2\big(S(x)+1\big)$,

  $$\mathrm{Var}_{\mathrm{Dir}}(p_k) = \frac{\alpha_k(x)\big(\alpha_k(x)+1\big)S(x) - \alpha_k^2(x)\big(S(x)+1\big)}{S(x)^2\big(S(x)+1\big)} = \frac{\alpha_k(x)\big(S(x)-\alpha_k(x)\big)}{S(x)^2\big(S(x)+1\big)}.$$

  Summing over k yields

  $$u_{\mathrm{epi}}(x) = \sum_{k=1}^{K} \frac{\alpha_k(x)\big(S(x)-\alpha_k(x)\big)}{S(x)^2\big(S(x)+1\big)}. \tag{10}$$

Putting both together, we recover the compact expressions:

$$u_{\text{ale}}(x) = \sum_{k=1}^{K} \frac{\alpha_k(x)\,(S(x) - \alpha_k(x))}{S(x)\,(S(x)+1)}, \qquad u_{\text{epi}}(x) = \sum_{k=1}^{K} \frac{\alpha_k(x)\,(S(x) - \alpha_k(x))}{S(x)^2\,(S(x)+1)}. \tag{11}$$

Appendix B. Theoretical Bounds and Normalization for Evidential Uncertainties

We derive here the analytic extrema used to normalize all uncertainties to $[0,1]$. Let $\boldsymbol{\alpha}(x)$ be a K-class Dirichlet with total evidence $S(x) = \sum_k \alpha_k(x)$. To obtain meaningful and comparable bounds, we consider the extreme configurations of $\boldsymbol{\alpha}$ that maximise (or minimise) each uncertainty while remaining consistent with the evidential interpretation of the Dirichlet.

- *Bounds for aleatoric uncertainty.* Aleatoric uncertainty

  $$u_{\text{ale}}(x) = \sum_{k=1}^{K} \frac{\alpha_k(S - \alpha_k)}{S(S+1)}$$

  quantifies intrinsic class ambiguity *given fixed evidence.* It is maximised when the classifier assigns equal expected class probabilities, i.e. when the Dirichlet is symmetric:

  $$\alpha_1 = \cdots = \alpha_K = c, \qquad S = Kc.$$

  Substituting,

  $$u_{\text{ale}} = \sum_{k=1}^{K} \frac{c(Kc - c)}{Kc(Kc+1)} = \frac{Kc(K-1)c}{Kc(Kc+1)} = \frac{K-1}{K} \cdot \frac{c}{c + \frac{1}{K}}.$$

  As $c \to \infty$ (high-evidence but fully ambiguous), this converges to

  $$u_{\text{ale}}^{\max} = \frac{K-1}{K}.$$

  Regarding its minimum, it is 0, achieved when one class dominates ($\alpha_j \to S$, others $\to 0$).

- *Bounds for epistemic uncertainty.* Epistemic uncertainty

  $$u_{\text{epi}}(x) = \sum_{k=1}^{K} \frac{\alpha_k(S - \alpha_k)}{S^2(S+1)}$$

  captures the variability of the Dirichlet mean under uncertain evidence. It is maximised when the model expresses *complete ignorance*, i.e. the Dirichlet concentration is at its weakest:

  $$\alpha_1 = \cdots = \alpha_K = 1, \qquad S = K.$$

Substituting,

$$u_{\text{epi}} = \sum_{k=1}^{K} \frac{1(K-1)}{K^2(K+1)} = K\,\frac{K-1}{K^2(K+1)} = \frac{K-1}{K(K+1)}.$$

Any increase in evidence (larger α_k) monotonically decreases u_{epi}, hence this is the theoretical maximum. Regarding its minimum, it is 0, achieved when one class dominates ($\alpha_j \to S$, others $\to 0$).

- *Bounds for vacuity.* Vacuity

$$u_{\text{vac}}(x) = \frac{K}{S(x)}$$

reflects the *absence of evidence.* Its minimum occurs when evidence is arbitrarily large ($S \to \infty$):

$$u_{\text{vac}}^{\min} = 0.$$

Its maximum occurs when the evidence is minimal, i.e. $\alpha_k = 1$ for all classes, $S = K$:

$$u_{\text{vac}}^{\max} = \frac{K}{K} = 1.$$

Given any uncertainty value $u(x)$ with theoretical interval $[u_{\min}, u_{\max}]$, we map it to a common $[0,1]$ scale via

$$\tilde{u}(x) = \frac{u(x) - u_{\min}}{u_{\max} - u_{\min}}.$$

This yields aligned and interpretable uncertainty scores across pixels, instances, uncertainty types, and datasets.

Appendix C. Additional segmentation-head uncertainty plots.

We provide full histogram and eCDF visualisations for all segmentation-head uncertainties ($u_{\text{ale}}, u_{\text{epi}}, u_{\text{vac}}$) at the instance level, separately for PanNuke and Ki-67 and for both evidential variants. Specifically, Figures 4 and 5 show the results for *Ours* on PanNuke and Ki-67 respectively, while Figures 6 and 7 report the corresponding plots for *Ours w.* These visualisations complement the main paper results and consistently show strong separation between correct and incorrect nuclei across datasets and uncertainty types.

Appendix D. Additional Centroid Uncertainty Plots

We provide complete histogram and eCDF visualisations for all centroid-head uncertainties (u_{peak}, u_{mass}, and their linear combination u_{cent}), separately for PanNuke and Ki-67 and for both evidential variants. Figures 8 and 9 correspond to *Ours*, and Figures 10 and 11 show the same plots for *Ours w.* Because centroid uncertainty arises from geometric cues rather than class probabilities, all evaluations are conducted at the *instance level.* Across datasets, errors consistently appear in the high-uncertainty tail, while correctly detected nuclei cluster at lower values.

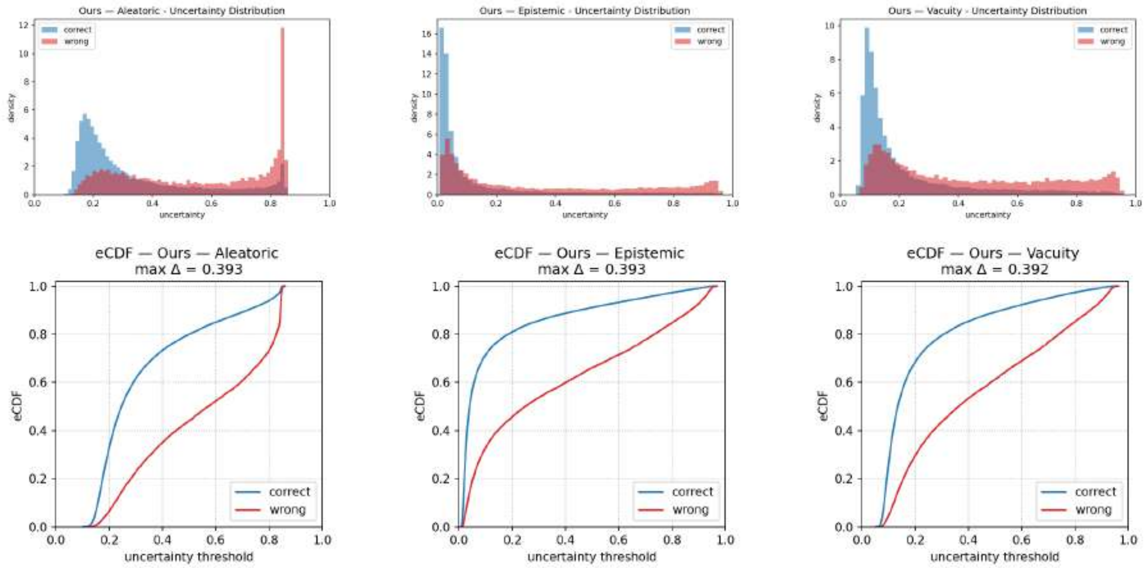

Figure 4: Instance-level histograms (top) and eCDFs (bottom) for segmentation uncertainties (u_{ale}, u_{epi}, u_{vac}) on PanNuke, Ours. Errors in red, correct nuclei in blue.

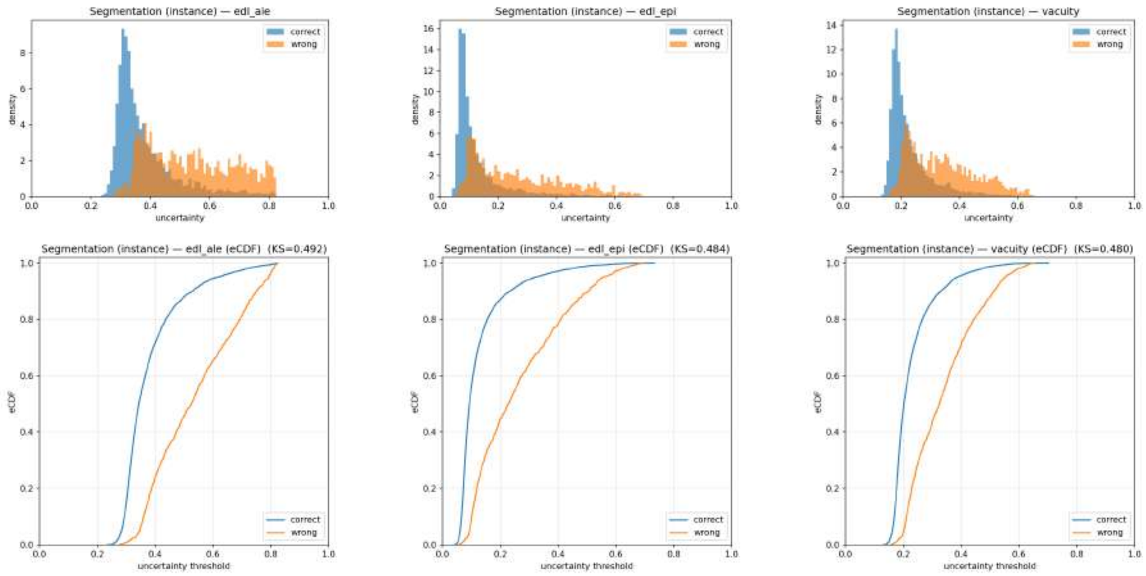

Figure 5: Instance-level segmentation uncertainty histograms and eCDFs for Ki-67, Ours. Errors in red, correct nuclei in blue.

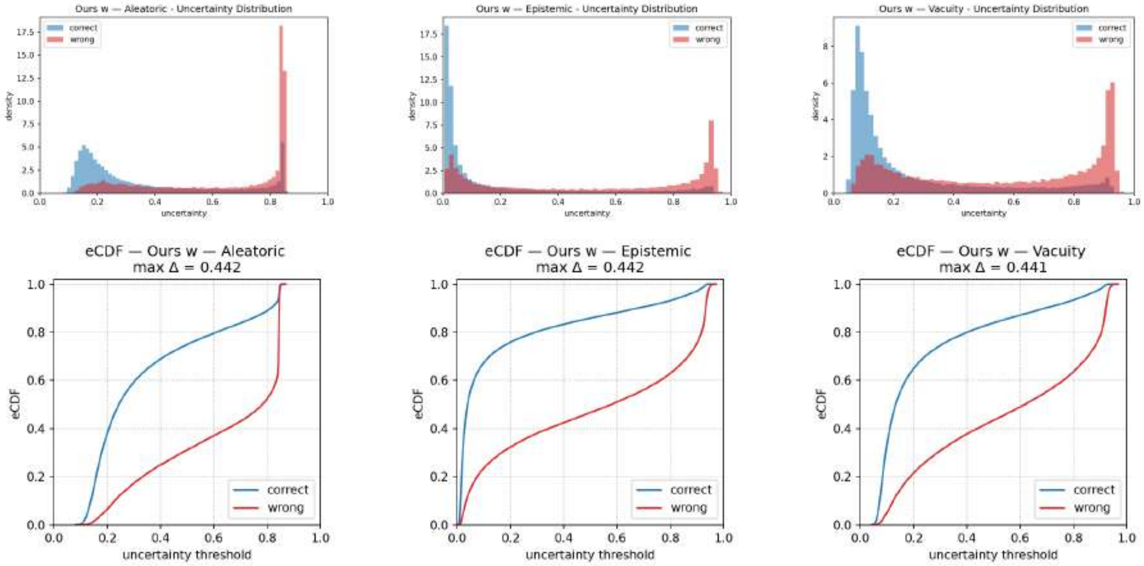

Figure 6: Instance-level segmentation uncertainty histograms and eCDFs for PanNuke, Ours w. Errors in red, correct nuclei in blue.

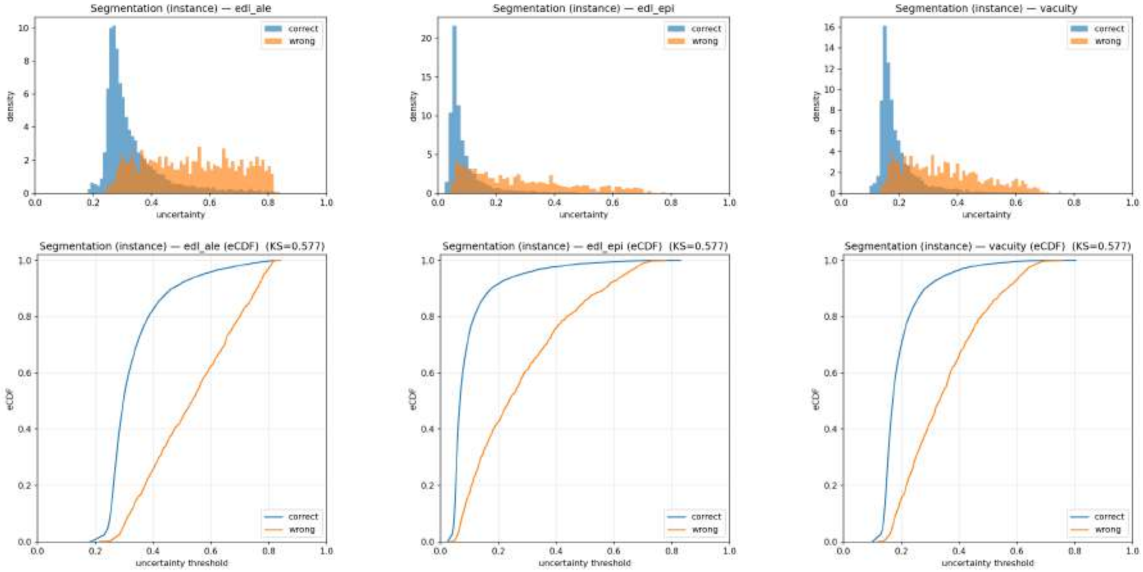

Figure 7: Instance-level segmentation uncertainty histograms and eCDFs for Ki-67, Ours w. Errors in red, correct nuclei in blue.

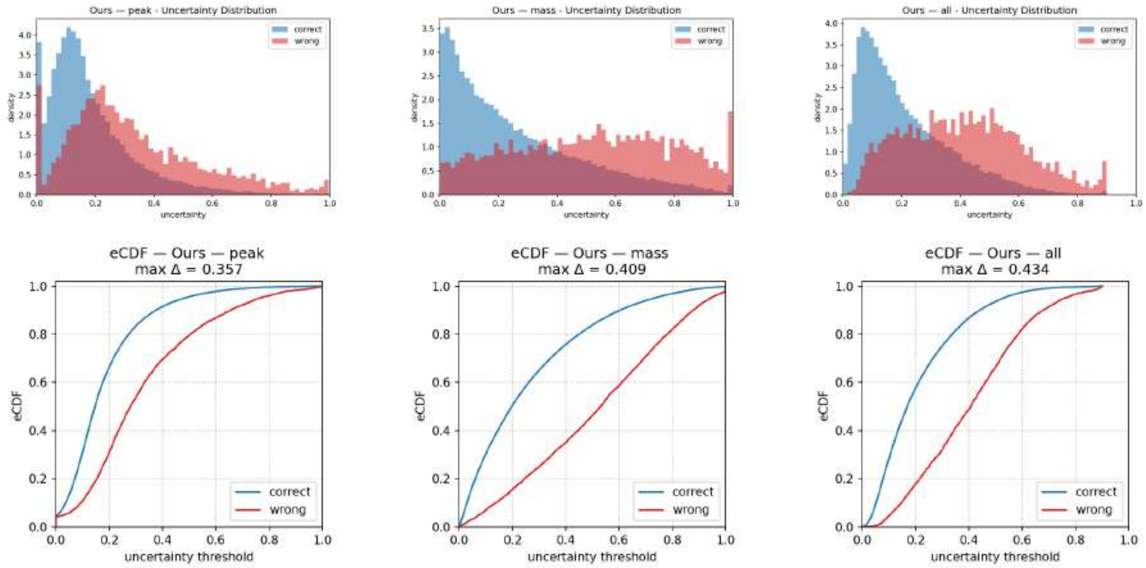

Figure 8: Instance-level centroid-head uncertainties on PanNuke for *Ours*. Top: histograms for $u_{\text{peak}}, u_{\text{mass}}, u_{\text{cent}}$. Bottom: eCDFs. Errors in red, correct nuclei in blue.

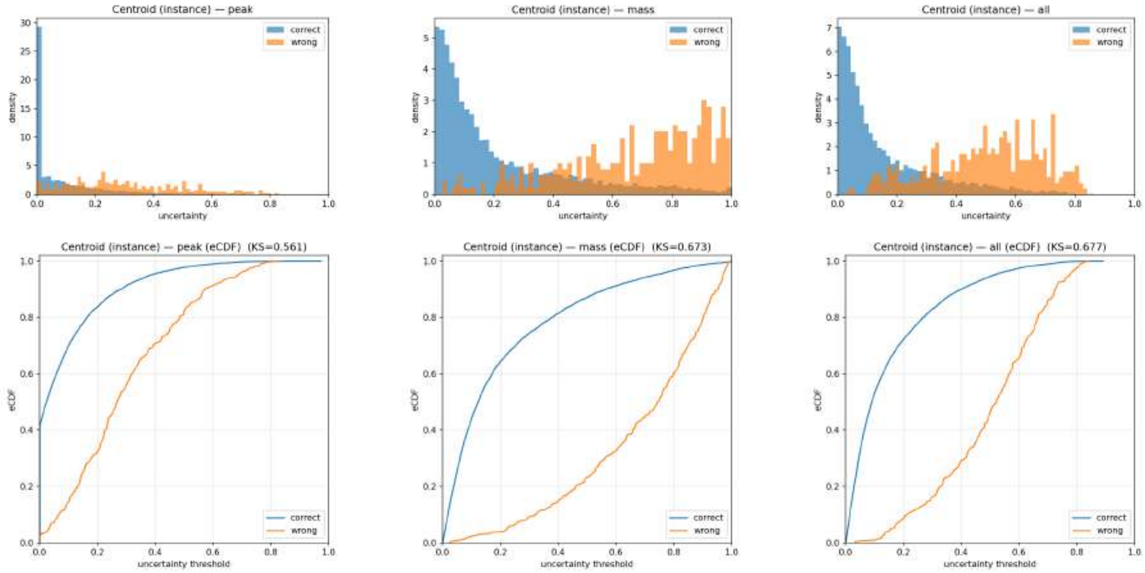

Figure 9: Instance-level centroid-head uncertainties on Ki-67 for *Ours*. Top: histograms. Bottom: eCDFs. Errors in red, correct nuclei in blue.

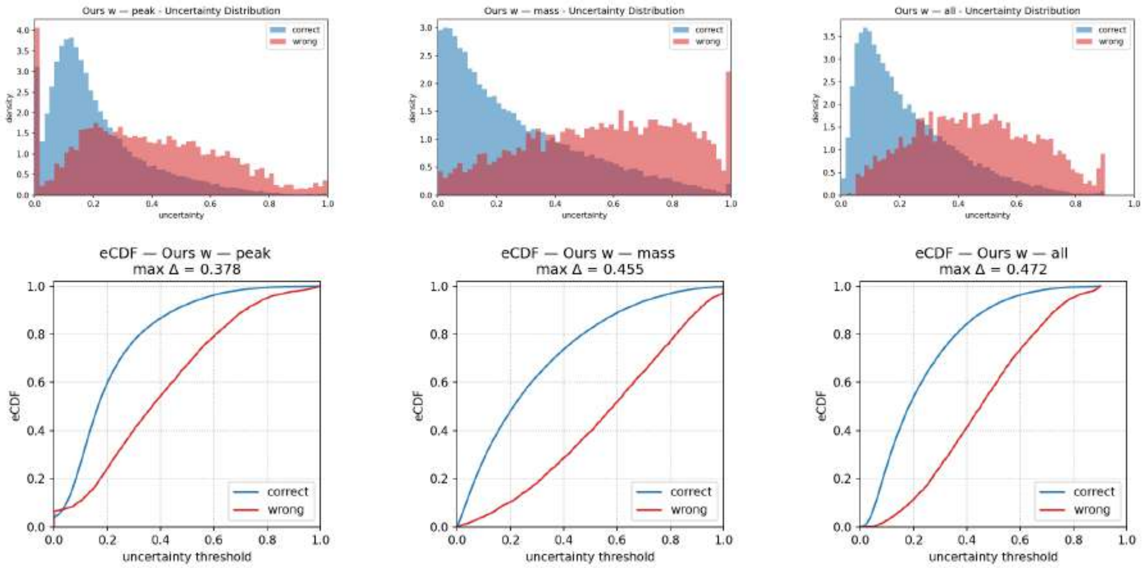

Figure 10: Instance-level centroid-head uncertainties on PanNuke for *Ours w*. Top: histograms. Bottom: eCDFs. Errors in red, correct nuclei in blue.

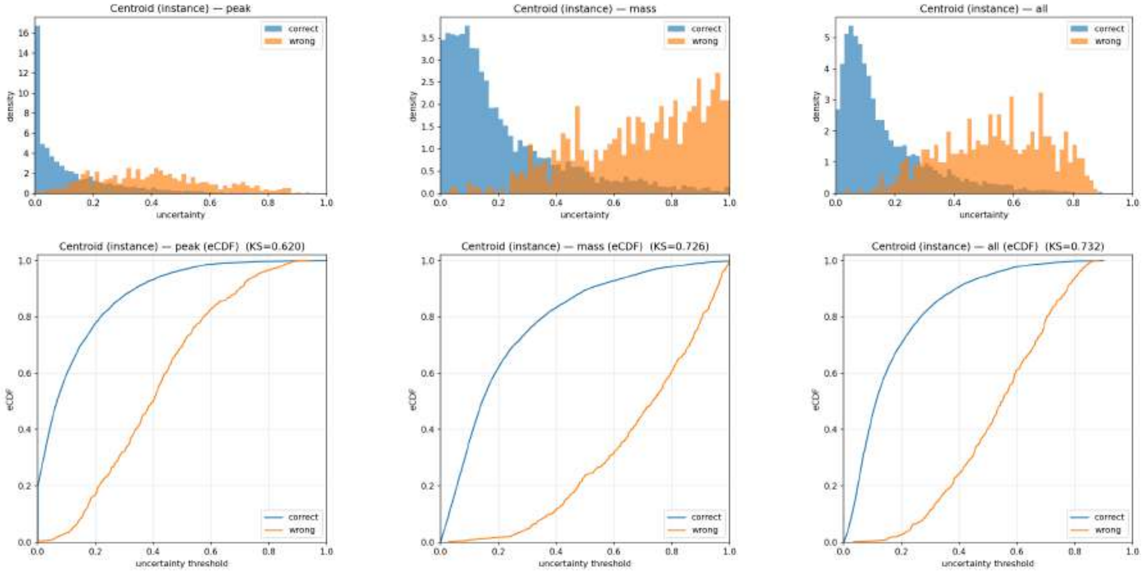

Figure 11: Instance-level centroid-head uncertainties on Ki-67 for *Ours w*. Top: histograms. Bottom: eCDFs. Errors in red, correct nuclei in blue.