Enhancing Ki-67 Cell Segmentation with Dual U-Net Models: A Step Towards Uncertainty-Informed Active Learning

David Anglada-Rotger, Julia Sala, Ferran Marques, Philippe Salembier, Montse Pardàs
Image Processing Group - Universitat Politècnica de Catalunya (UPC)
{david.anglada, ferran.marques, philippe.salembier, montse.pardas}@upc.edu

Abstract

The diagnosis and prognosis of breast cancer relies on histopathology image analysis where markers such as Ki-67 are increasingly important. The diagnosis using this marker is based on quantification of proliferation, which implies counting of Ki-67 positive and negative tumoral cells, excluding stromal cells. A common problem for automatic quantification of these images derives from overlapping and clustering of cells. We propose in this paper an automatic segmentation and classification system that overcomes this problem using two Convolutional Neural Networks (Dual U-Net), whose results are combined with a watershed algorithm. Taking into account that a major issue for the development of reliable neural networks is the availability of labeled databases, we also introduce an approach for epistemic uncertainty estimation that can be used for active learning in instance segmentation applications. We use Monte Carlo Dropout within our networks to quantify the model's confidence across its predictions, offering insights into areas of high uncertainty. Our results show how the postprocessed uncertainty maps can be used to refine ground truth annotations and to generate new labeled data with reduced annotation effort. To initialize the labeling and further reduce this effort, we propose a tool for groundtruth generation which is based on candidate generation with maxtree. Candidates are filtered based on extracted features which can be adjusted for the specific image typology, thereby facilitating precise model training and evaluation.

1. Introduction

Breast cancer is the most common type of cancer for women worldwide, and an early detection and diagnosis are crucial for improving the survival rate. One of the primary methods for diagnosis are immunohistochemical (IHC) tests of biopsies. Pathologists first analyze the tissue obtained through the common Hematoxylin-Eosin (H&E) staining, to detect tumoral areas, and then apply additional stains for further classification of the tumor and for patient risk stratification. To predict prognosis and therapeutic response, usually quantification of cell proliferation is required, which can be assessed with the stain produced by the Ki-67 biomarker [7]. The Ki-67 index is obtained by counting the percentage of positively stained tumoral cells over all the malignant cells (positive and negative). Usually, this involves manually counting between 500 and 1000 cells in three randomly selected high-power fields or estimating by eyeballing the Ki-67 Index, without formally counting. As expected, these methods, although very labor-intensive, result in important variability and low reproducibility depending on the selected zones and the used method [16]. During the last decade, digital pathology is being deployed in an increasing number of pathology departments [22]. Digital pathology involves high-resolution digital images (Whole-Slide Images, WSI) obtained from biopsy samples captured with a scanning device. WSI can contain up to 40 Gb uncompressed data, thus substituting traditional light microscopes. At the same time, digital image analysis (DIA) techniques are emerging for automatic quantification of the most common stains (H&E, Ki-67, ER (estrogen receptor), PR (progesterone receptor) and HER2 for breast cancer) [27].

In order to automatically compute the required indices for the previous stains, nuclear segmentation and classification is required. Recent advances in Computer Vision based on Convolutional Neural Networks (CNNs) produce outstanding results in sematic segmentation. Semantic segmentation identifies each pixel in an image as belonging to one of the predefined classes, generating a mask for each of these classes, but does not separate connected samples of the same class. Instance segmentation and classification can be achieved with more complex CNNs or by combination of semantic segmentation with instance detection methods.

Uncertainty estimation is important for interpreting the trustworthinesss of deep learning models. It can enhance explainability, which is paramount in medical AI as it increasingly influences crucial healthcare decisions [9, 23]. Since uncertain predictions are usually the most informa-

tive, uncertainty estimation is also used in active learning frameworks for selecting the most informative unlabeled samples in order to achieve a high accuracy. Epistemic uncertainty, related to the model and often due to lack of training data, has been shown to be the most appropriate for active learning [15].

In this paper we propose an efficient and accurate algorithm to segment and classify tumor cell nuclei on breast cancer histology based on semantic segmentation combined with nuclei center estimation. The algorithm has been developed for Ki-67 stained images, and succeeds on the major challenges of this kind of images: separation of overlapping cells and classification for proliferation index computation with consistent discrimination of the stroma cells. Furthermore, we complement our system with epistemic uncertainty estimation through Monte Carlo Dropout techniques in the U-Net models. This addition allows to assess the confidence level of the predictions. Combined with a graphical annotation tool it can be used in an active learning framework for correcting errors in the groundtruth and for labeling new data with reduced effort.

The paper is organized as follows. Sec. 2 describes related works, Sec. 3 presents the tools used to generate the groundtruth without an extensive manual labelling effort, Sec. 4 describes the pipeline and the networks used for segmentation, Sec. 5 introduces our approach to uncertainty estimation, Sec. 6 presents the results obtained and Sec. 7 discusses the results and concludes the paper.

2. Related work

2.1. Ki-67 proliferation index computation

In March 2010, the 'International Ki-67 in Breast Cancer Working Group' agreed that the Ki-67 measurement was a key point for tumor proliferation studies and developed guidelines for its analysis based on computing the so-called Ki-67 proliferation score. This score was defined as the percentage of positively stained cells among the total amount of evaluated tumor cells [5]. Manual procedures to obtain this index are time consuming and present a high variability, mainly due to limitations of the human eye and randomness in choice of the regions for cell counting.Thus, several works have focused on the automation of this process.

A recent study [24] has validated the usage of computer assisted image analysis for Ki-67 stained images. Their results confirm that there is a significant benefit of automated image analysis as part of daily pathologists' workflow, both in the consistency of the automated results and in the time savings for pathologists. The work of [4] compares commercial applications that have been developed for semiautomated Ki-67 quantification, many of which rely on measurements in user-defined regions of interest (ROIs). They observed that results depend on the size of the ROI and that a common rejection cause of the software results was due to the confusion between tumor and stroma cells. This caused a rejection of 23% of the samples.

The most common approach taken for DIA systems is to rely on ROIs defined by the user in order to avoid stromal areas. In [1], an automatic approach for Ki-67 index estimation is presented. The process is applied to hot-spot regions (area of higher density of positive tumor cells for Ki-67) where stromal cells are not observed. The system relies on color processing techniques to segment nuclei, which are then classified as positive or negative based on color and shape features. A recent work [2] proposes a pipeline for accurate automatic counting of Ki-67 cells, using UNet for nuclei segmentation, combined with a watershed algorithm to separate overlapped regions, and a final classification into positive and negative nuclei by a random forest classifier using deep features extracted from each nucleus patch. They recognize as the biggest challenge the separation of overlapped cells in clustered areas. The analysis is also performed on manually selected hot-spots of small size, with little presence of stromal cells.

2.2. Semantic segmentation

Semantic segmentation approaches the image segmentation problem by performing pixel-level classifications. CNN-based techniques in end-to-end architectures have become mainstream to approach this task when annotated data is available. The most successful model for biomedical image segmentation has been the one proposed by Ronneberger, U-Net [18]. It follows an encoder-decoder architecture, where the encoder gradually reduces the spatial dimension with pooling layers and the decoder gradually recovers the object details and spatial dimension. Although other semantic segmentation networks have succeeded in different tasks [3, 10, 20, 28] U-Net is still the state of the art for biological images. Semantic segmentation with U-Net combined with connected component analysis has been used for Ki-67 index computation in bladder cancer [12]. However, cases of clustering with severe overlapping were not considered.

2.3. Cell counting

In [13] a supervised learning framework was proposed for visual object counting tasks. It required as annotation a dot for each object, and the goal was to accurately estimate the count, evading the task of learning to detect and localize individual object instances. The problem is approached as mapping learning between an image and an image density map whose integral over any image region gives the count of objects within that region. Learning to infer such density was formulated as a minimization of a regularized risk quadratic cost function. In [26] this mapping problem was approached with CNNs, using a fully convolutional regression network to estimate the mapping from image to image

density. The application in this case was microscopy cell counting and detection. Counting cells in histopathology images is useful in cases of overlapped cells, but it does not give information of the cell morphology or class, which is crucial for applications such as PI calculation.

2.4. HoVer-Net

In [8] a deep learning approach for simultaneous segmentation and classification of nuclear instances in histology images was presented. The network is based on the prediction of horizontal and vertical distances of nuclear pixels to their centres of mass, which are leveraged to separate clustered nuclei and deliver instance segmentation. This is combined with a semantic segmentation network in order to achieve both segmentation and classification of the nuclei. This architecture was applied to H&E stained images corresponding to different types of cancer. The authors proved, for the H&E databases , superior results to those obtained with instance segmentation methods such as Mask-RCNN or SegNet combined with watershed, mainly due to the ability to separate nuclear instances which were overlapping.

2.5. Epistemic uncertainty estimation

Uncertainty estimation can be categorized into two types: aleatoric uncertainty, which deals with the inherent noise within the data, and epistemic uncertainty, which pertains to the model's knowledge about the data used for training. The theoretical foundation for estimating epistemic uncertainty is well-articulated in the pioneering work [6] through the Monte Carlo Dropout technique. This approach leverages dropout layers not only during training but also crucially during inference. By doing so, it approximates Bayesian posterior distributions, generating multiple predictions for a single input. The variability among these predictions provides a measure of epistemic uncertainty, reflecting the model's confidence in its learned representations.

Building on this foundation, [11] elaborated on the nuances of uncertainty in deep learning. In semantic segmentation tasks, this delineation allows models to not only predict with accuracy but also to highlight regions of low confidence. Such regions might require further scrutiny, additional data collection, or expert intervention, thereby elevating the utility and interpretability of the model's outputs. Its benefits in active learning are shown in [14], proving that uncertainty estimation can be used to select annotation regions, producing models with higher accuracy and less labeling effort than annotation of full images.

3. Database construction

After looking for publicly available databases of Ki-67 images with annotated ground-truth, we decided to create our own database and its corresponding annotation. This strategy was selected because we did not find any database with all the characteristics we were looking for in terms of variability of cases and of type of annotation (the database annotation had to define the mask and the class of each cell).

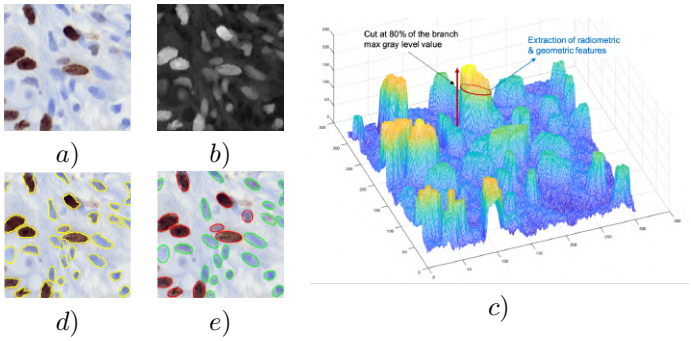

Figure 1. Preliminary ground-truth generation with morphological tools: a) Original image. b) Gray level image. c) Maxtree representation and feature extraction. d) Cell segmentation. e) Cell classification and elliptical approximation

The database was constructed extracting tiles from WSIs from 4 different patients of invasive breast carcinomas. WSIs came from the Vall d'Hebron hospital (Barcelona) and were scanned with a 3DHISTECH Pannoramic 1000 slide scanner with an objective magnification of 40x and a resolution of 0.25 × 0.25 μm/pixel. Starting from Ki-67 WSIs, we have obtained 52 tiles of size 1024x1024 pixels, corresponding to a field of view of 256x256μm. The patients had different levels of proliferation, marked by the Ki-67 stain, and the tiles reflect a wide variety of cellular structures. To reduce the computational load and memory usage, images were downsampled to 512x512 pixels. According to our experiments, this resolution is enough for cell detection and nucleus segmentation.

In these original images, cells have to be classified into three possible classes: positive, negative or non-epithelial (e.g. stroma). The principal characteristics of cells of each class differ both in their nucleus color and shape.

In order to create the ground-truth, we had to rely on the expertise of pathologists from various hospitals (this work was performed in the context of a project with a national health institute). However, it was unrealistic to ask pathologists to annotate the images at the cell level from scratch as each image may involve between 300 and 800 cells. As a result, we created an initial preliminary annotation to try to minimize pathologists' workload.

This preliminary annotation was created by a mathematical morphology approach, as illustrated in Fig. 1. The original images were first converted into gray level images and inverted, so nuclei appeared as maxima (see Fig. 1.b). These images were represented by a maxtree [19], where

the tree leaves are the image maxima, the root node the entire image support and the tree branches highlight the morphological structure around maxima. The image represented as a 3D relief can be seen in Fig. 1.c. For each maximum that had sufficient height, we considered as cell candidate the binary connected component made by a cut at 80% of the maximum value (see the illustration in Fig. 1.c). This way, for each maximum, it was possible to extract a specific shape (cell candidate) as well as its radiometric and geometric features. In practice, for each cell candidate we computed its mean RGB color values, its size measured as number of pixels, and its elongation. The classification was done by manually adjusting the thresholds on the previous features. An example of resulting cell detection can be seen in Fig. 1.d.

As this preliminary annotation had to be corrected by pathologists, it was necessary to transform the detection in a representation that could be easily manipulated by means of a graphical annotation tool. To this end, we approximate each cell by its best fitting ellipse and assigned a specific color to each cell class (see Fig. 1.e.). If necessary, the pathologists could easily modify the control points or the orientation of the ellipse to improve the matching with the actual data. The class could also be easily corrected.

The preliminary ground-truth was given to a set of expert pathologists from various hospitals in Catalonia. When necessary, they modified the cell segmentation results, corrected the classification and accepted the resulting images. To address the scarcity of data, we adopted a prudent approach by implementing a patient-fold cross-validation strategy during the training of our novel deep learning model. Within each fold, we allocated the images of three of the four available patients for the training set and reserved the images of the fourth patient for the test set. This approach not only served to evaluate the model's ability to generalize across new, unseen patient data, mirroring real-world application scenarios, but also aimed to effectively leverage the data from all four patients. Despite the dataset's constrained size (52 images), each image encompasses a range of 500 to 1000 cells, providing a rich dataset for our model's learning and evaluation processes.

The final ground-truth consists of label images where pixels from a given cell are identified with a unique label, and of a list assigning each cell to one possible classes.

4. Dual U-Net segmentation and classification

The algorithm consists of a semantic segmentation network based on a U-Net network that classifies each pixel into one of the four classes for Ki-67 images (positive, negative, stroma and background) given the original color images of size 512x512x3. The resulting pixel class determines the shape, position and class of the predicted cells. However, as mentioned before, this algorithm does not behave accurately when cells are close or overlapping, because it merges them as a single cell. Hence, the need of a second network, which runs in parallel, with the objective to detect the center of each cell. A U-Net model is also used and, to train it, a density map is created for each image, and the task of the network is to regress this density surface from the original images. To predict the cell centers, the local maxima of the predicted density map have to be detected. To do so, a contrast filter is applied which detects local maxima which exceed a given contrast threshold.

Finally, the results of the two networks are merged (see Fig. 2). A watershed is applied on each predicted connected component from the *segment* output to split them into as many cells as centers the *count* network has predicted. In some cases we can find cells without a detected center or vice versa (centers without cell). Experiments have demonstrated that the best results are achieved by utilizing all segmented cells and only the centers associated with at least one cell. For cells that lack a detected center, their centers of mass are computed. Finally, each resulting connected component corresponds to a cell, and a single class has to be assigned to it. Semantic segmentation assigns a class to each individual pixel and, after the cell shape has been defined, majority voting of the pixel class within each connected component is used for class assignment.

4.1. Semantic segmentation

Semantic segmentation aims at classifying the individual pixels of the image into one of the four defined classes (three cell types plus the background). From this pixel classification the cell shape, position and class are defined.

We use as semantic segmentation model the U-Net model [18], which has proved to produce very good results for many biomedical image segmentation tasks.

The network is trained using the ground truth multiclass masks, where each pixel is identified by a label representing each class. A Dice loss is used to train the network and weighted F-score, precision and recall at pixel level for the three cell classes are used as metrics to evaluate the results.

The ResNeXt-50-32x4d [25] is used as encoder. Due to the reduced size of the available database, transfer learning is applied by initializing the weights of the encoder with the "ImageNet" weights. We used Adam as optimizer and a batch-size of 4 images with a learning rate of 0.0005. To avoid overfitting, data augmentation was applied with the following parameters: shift factor range of (-0.1, 0.1) and flipping with probability 0.5. Batch-size and learning rate were not specifically adjusted for this experiment; we used the same values that we use for semantic segmentation in histopathological images of other stains with databases of similar size. The model was trained for 200 epochs, and the best-performing checkpoint was selected. Final results can be seen in Sec. 6.

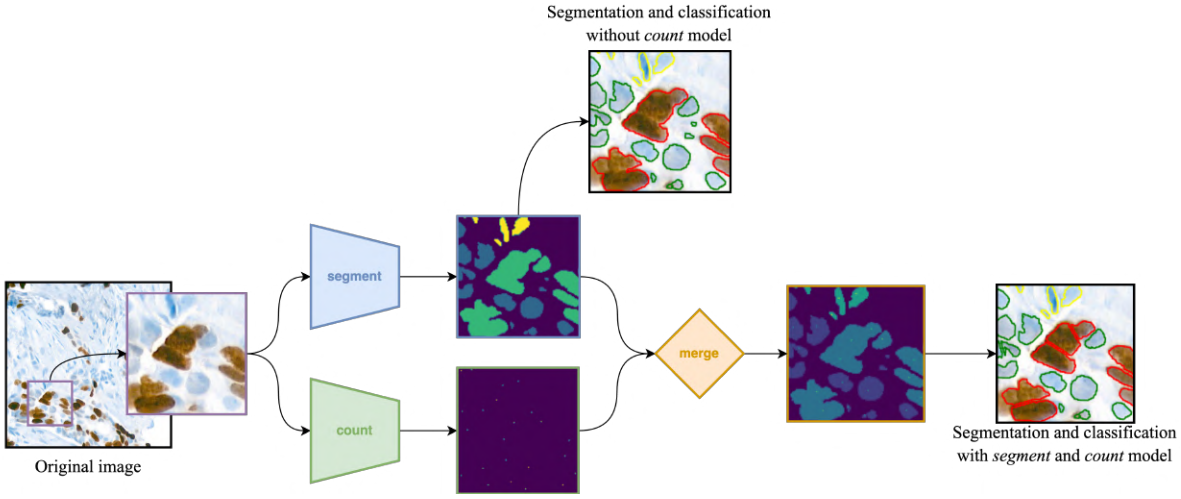

Figure 2. Schematic representation of the application of the segmentation model (named *segment*) and the cell center detection model (named *count*). As can be observed, there is considerable differences between the segmentation and classification without the application of the *count* model (image at the top) and with the application of this network (image at the right).

4.2. Cell center detection

The semantic segmentation can merge different cells in a unique connected component when cells are close. In order to detect the cell centers we use a system which predicts the cell density map from the original image. To do so, a specific Fully Convolutional Regression Network was designed in [26]. In this work, we propose to use for this task the same architecture based on U-Net as for semantic segmentation. Here, in order to regress the density maps, the loss function is the Mean Square Error.

For training the network, density maps for the Ki-67 dataset have to be provided. This can easily be generated from an initial dot annotation. This dot annotation is computed as the mass center of each cell identified in the ground truth. Afterwards, each dot is represented by a Gaussian, and the density map is formed by the superposition of these Gaussians. The central task of the network is to regress this density map from the corresponding original cell image. We are interested in cell center detection, which can be achieved by local maxima detection on the density map.

The standard deviation of the Gaussian functions used to generate the ground truth has to be adjusted according to the spatial dimensions of the cells. Our experiments showed that a standard deviation between 3 and 7 provides accurate results, and we used 7 for further experiments.

For the *count* final step, the local maxima in the predicted image are retrieved. As the prediction is noisy, a contrast filter [21] is applied to only select those local maxima with a contrast greater than a given threshold h. This allows to separate cells whose representative density function overlap. See Fig. 3, where the used contrast value is 0.15.

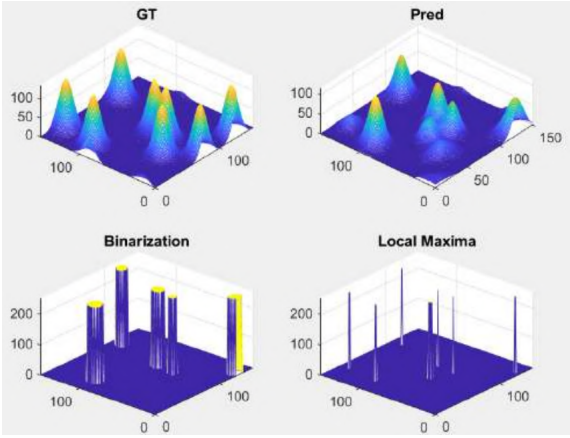

Figure 3. 150x150 square section of an image: Ground truth with Gaussian filter of standard deviation 7, prediction obtained by the network, binarization and local maxima extraction with $h = 0.15$.

As we are interested in measuring the performance in terms of cell center prediction, a matching between the centers of the ground truth and of the prediction is performed. Given each center of the prediction the closest center in the ground truth is searched and vice versa, discarding the correspondences with higher distance than a certain threshold and the correspondences that are not one to one.

The final network used the U-Net architecture, with ResNeXt-50-32x4d, Adam optimizer, batch-size 4 and learning rate 0.0005. The model was trained for 100 epochs, and the best-performing checkpoint was selected. The final results can be seen in Sec. 6.

4.3. Result combination

Analyzing the relation between the predictions of the two networks, we observed that while most connected components detected by the *segment* network had a single center detected by the *count* network, approximately 14% of the components had more than one center, while 15% of the detected components had no center. Moreover, some centers detected by the *count* network did not have any associated component. Although the watershed segmentation separates the connected components with more than one center, the best strategy had to be decided regarding the situations where there is no correspondence between the two networks. Experiments were performed considering all possible combinations of predictions and the decision was to consider all the predictions of the *segment* network, but only the centers of the *count* network that corresponded to a segmented cell.

Once the semantic segmentation is completed and the centers have been detected, we can split the connected components of the semantic segmentation into individual cells using the watershed algorithm. In this case, the predicted centers would be the markers that identify the starting point of each basin that the watershed needs to flood. If there is a cell predicted by the *segment* network that does not have an associated predicted center, the center of mass of the predicted cell is computed as cell center. Basins are defined by the inverse distance transform of the binarized output of the *segment* network. The distance transform shows the distance from each foreground pixel to its closest background pixel. Thus, the inverse distance transform has higher values in the borders and lower values in the centers of the cells, creating basins at each cell. The predicted components are also useful to define the zones that the flooding cannot exceed, called the mask. This method is able to split large connected components in as many markers as it contains because every marker will start a flood if a basin has been defined (see Fig. 4)

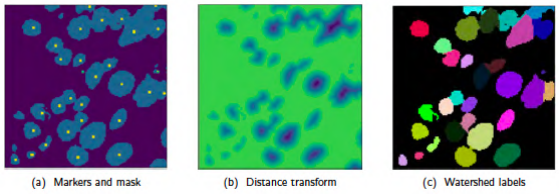

Figure 4. Watershed process. a) The markers showing the center cells that have to be retrieved and the mask showing the regions that need to be labeled, b) The inverse distance transform of the segmentation prediction (in green high values, in blue low values), and c) The watershed labels with a different color for each individual cell.

4.4. Homogenization of individual cells

Since semantic segmentation is pixel based, a predicted cell could have pixels of different classes. So, as a last step, for each one of the connected components produced after the watershed, the median of the values from the multiclass prediction is computed and assigned to each pixel. Thus, each predicted cell will only have one associated class.

5. Epistemic uncertainty estimation

In this study, we introduce an advanced methodology for estimating epistemic uncertainty within our semantic segmentation model. Drawing upon the technique proposed in [11], and using Monte Carlo Dropout (MCD) [6] as a Bayesian approximation, we quantify the uncertainty in our model predictions. To integrate epistemic uncertainty estimation, we augmented the decoder block of our *segment* network with a dropout layer after each convolutional layer. This facilitates the use of MCD during the inference phase, allowing the network to simulate Bayesian inference processes. By conducting inference repetitively (30 times in our study), we capture the variability in the predicted class for each pixel. The mean of the per class variance is interpreted as pixel uncertainty.

The resultant uncertainty heatmaps highlight the borders of segmented cells as areas of high uncertainty. Recognizing that border uncertainties do not significantly contribute to our analysis, we refine these heatmaps by filtering out the border uncertainties in order to focus on the uncertainties within the cell interiors. A subsequent rescaling of heatmap values ensures that these refined maps accurately represent meaningful uncertainties for analysis (Fig. 5). These refined uncertainty heatmaps serve a dual purpose. First, they enable the identification of regions within the ground truth labels that may be inaccurate or uncertain. Such insights are useful for iterative model refinement, allowing us to augment our dataset with revised annotations that address previously undetected ambiguities. Second, heatmaps facilitate an active learning loop, where the model uncertainty directly informs the selection of samples for reannotation and retraining.

While the active learning process itself has not been explicitly developed in this study, we posit that leveraging model uncertainty to inform sample selection for reannotation and retraining could significantly enhance the efficiency and efficacy of data labeling efforts. However, it is worth noting that the envisioned active learning cycle represents a prospective application of our uncertainty estimation technique. We believe that utilizing epistemic uncertainty as a strategic tool for dataset enhancement and model development holds substantial promise, particularly in aligning with the principles of active learning to prioritize data labeling efforts effectively.

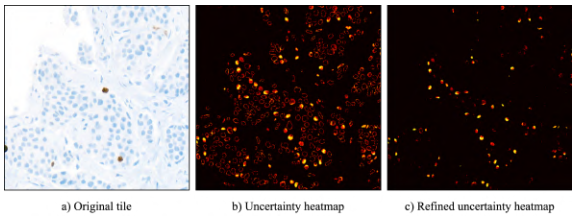

Figure 5. Example of the uncertainty heatmap refinement process. a) The original image. b) The output uncertainty heatmap, being the brigher regions the ones with higher uncertainty. c) The refined heatmap after eliminating the cell borders uncertainties.

6. Results

As explained in Sec. 3, due to the low number of images with ground truth supervised by the pathologists, we have created 4 different folds, each one containing the images of three of the four available patients.

In evaluating our proposed method for cell segmentation and classification, we assessed the weighted F1-Score at the cell level as the primary metric, as it considers the balance between precision and recall in identifying true positive cell predictions. Additionally, recognizing the practical implications of our work in the clinical setting, we introduced a problem-specific metric, the Mean Absolute Error (MAE) of the Proliferation Index. The MAE serves as a crucial tool for evaluating the estimated Ki-67 index. The correct estimation of this index ensures that our method aligns with the specific clinical requirements, providing a more direct and interpretable measure of the model's impact on diagnostic processes. By employing these metrics, we aim to establish the robustness of our approach in both a technical and clinically relevant context. We have selected as baseline for our result comparison the state-of-the-art architecture HoVer-Net (see Sec. 2). For completeness, we also provide F1 Score at cell level (without weights), which we have called Macro-F1 and the F1 score at pixel level, referred to as Micro-F1.

The results obtained in each one of the folds with the networks hyperparameters chosen as described in Sec. 4 are provided in Sec. 6. We present a final cell-based averaged weighted F1-score of 0.68 $\pm$0.07 among the four folds, being the positive class the one with the higher score of 0.80 $\pm$0.06. The non-epithelial class (e.g. stroma class), though, only achieves 0.58 $\pm$0.06. However, note that miss-detection or false alarms of stroma cells does not effect the calculation of the Ki-67 score. This score is only affected when there is a confusion between stroma and non-stroma class. An improvement of the results could be obtained by computing stroma-masks with a separate network [17]. In this case, cells would only need to be classified in positive and negative class, but an additional network would be needed, thus increasing the computation time.

The final mean absolute error between the ground truth and the predicted Ki-67 score is 3.89% $\pm$0.04. We can see that our Dual U-Net model, in average, improves the results of the HoVer-Net model in both weighted F1-Score and Mean Absolute Error (MAE) (see examples in Fig. 6). Additionally, we assess the quality of the uncertainty maps generated through our approach. Notably, cells with low staining intensity exhibit higher uncertainty, prompting further investigation. We showcase examples where, despite both ground truth and prediction identifying cells as positive, the high uncertainty and low staining intensity merit further review by pathologists. Additionally, we highlight the necessity of revisiting cases with ambiguous identification between non-epithelial cells and either negative or positive cells. These doubtful cases, which also significantly impact the Ki-67 score, are also candidates for the ground truth revision. This insight underscores the potential of employing uncertainty maps as a tool for active learning, refining ground truth annotations, and fostering the development of more robust models.

7. Conclusions

We have presented a system for quantification of Ki-67 images which overcomes the common problem of overlapping cells by using two parallel CNNs whose results are merged with a watershed algorithm. The task of the first network is semantic segmentation to classify pixels as belonging to negative, positive, stromal class or background, while the second network detects the nuclei centers by means of density estimation.

Since there was no annotated database available we have also developed a tool for groundtruth generation. This tool, based on connected component extraction using maxtree, has proven effective and easy to adapt to the specific features of the cells in Ki-67 staining. It is a versatile tool that can easily be adapted to compute groundtruth for other histopathological images (different stains or tumors).

Our approach not only surpasses the performance of the state-of-the-art HoVer-Net in cell detection and classification across all metrics but also includes a novel method for epistemic uncertainty estimation using Monte Carlo Dropout. This methodology allows us to quantify the model's prediction confidence, offering invaluable insights into potential inaccuracies or uncertainties within the ground truth annotations. By leveraging these insights, we propose an active learning strategy aimed at enhancing the quality and robustness of our model through iterative ground truth refinement.

The results show the improvement achieved with respect to the state of the art network for cell detection and classification HoVer-Net. It consistently outperforms HoVer-Net in each fold across various evaluation metrics, including

FOLD	Model	W - F1	M - F1	m - F1	N - F1	P - F1	S - F1	MAE
1	Prop	0.76	0.73	0.68	0.81	0.88	0.49	0.82%
	HNet	0.75	0.68	0.65	0.84	0.85	0.34	0.77%
2	Prop	0.69	0.71	0.60	0.71	0.80	0.61	3.36%
	HNet	0.62	0.65	0.52	0.67	0.80	0.49	5.47%
3	Prop	0.66	0.69	0.58	0.67	0.77	0.62	9.36%
	HNet	0.54	0.61	0.47	0.67	0.75	0.40	10.71%
4	Prop	0.60	0.63	0.52	0.55	0.74	0.61	1.99%
	HNet	0.56	0.61	0.45	0.54	0.75	0.56	2.64%
AVG	Prop	$\mathbf{0.68 \pm 0.07}$	$\mathbf{0.69 \pm 0.04}$	$\mathbf{0.60 \pm 0.07}$	$\mathbf{0.68 \pm 0.11}$	$\mathbf{0.80 \pm 0.06}$	$\mathbf{0.58 \pm 0.06}$	$\mathbf{3.89\% \pm 0.04}$
	HNet	0.62 ± 0.10	0.64 ± 0.03	0.52 ± 0.09	0.68 ± 0.12	0.79 ± 0.05	0.45 ± 0.09	$4.89\% \pm 0.04$

Table 1. Metrics obtained for each one of the folds and the average among all the five folds. We present the Macro F1-Score (M-F1), the Weighted F1-Score (W-F1), the Micro F1-Score (m-F1), the F1-Score of the negative class (N-F1), the F1-Score of the positive class (P-F1) and the F1-Score of the non-epithelial class (S-F1). We also present the Mean Absolute Error of the predicted Ki-67 index (MAE).

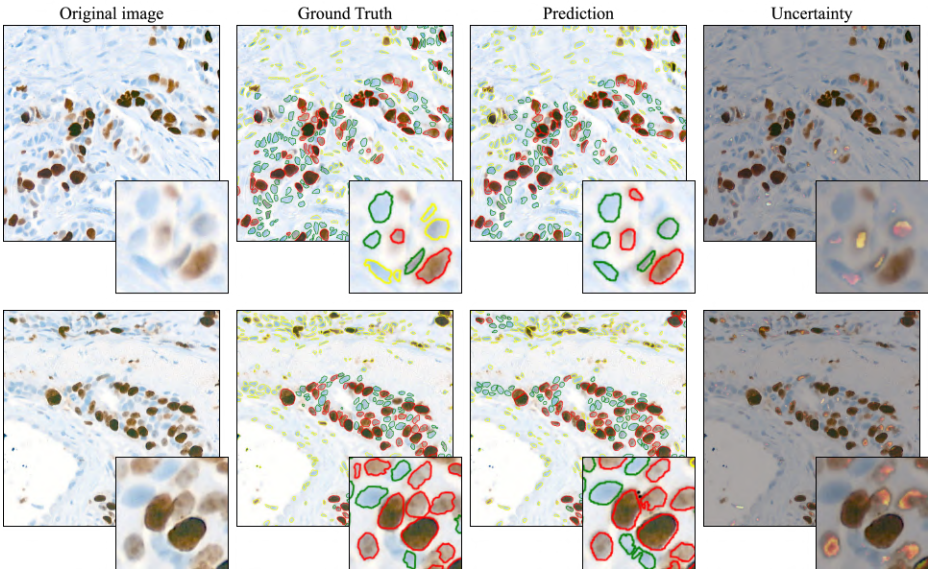

Figure 6. Visualization of the predicted results. We can see the original tile (left column), the ground truth (left-middle column), the predicted output (right-middle column) and epistemic uncertainty heatmap (right column). Regarding the segmentation and classification, in red we have the positive cells, in green we have the negative cells and, in yellow, the non-epithelial cells. Regarding the epistemic uncertainty heatmap, the brighter the pixel is, the higher uncertainty it has.

Macro F1, Weighted F1, Micro F1, Negative F1, Positive F1, and Stroma F1. We have observed that HoVer-Net performs consistent segmentation but fails more often in the classification between positive and stromal cells.

The performance of both models exhibits some variability across the different folds, reflecting the inherent challenges in medical image analysis. This variance underscores the importance of evaluating models across multiple data subsets to ensure their robustness and generalizability. The averaged results across folds demonstrate that our system maintains its superiority in terms of F1 scores, with consistently higher scores compared to HoVer-Net. The lower Mean Absolute Error (MAE) further confirms its accuracy in predicting Ki-67 index, which is actually the metric that the pathologists use to diagnose.

Currently, we are also applying this system for ER and PR staining, obtaining good quality results, which show the capability of generalization of the algorithms proposed.

8. Acknowledgements

This reseach has been funded by the project PID2020-116907RB-100 [AIMING] and PID2020-117142GB-I00 [DeeLight] by MCIN/ AEI /10.13039/ 501100011033.

References

[1] Barbara Rita Barricelli, Elena Casiraghi, Jessica Gliozzo, Veronica Huber, Biagio Eugenio Leone, Alessandro Rizzi, and Barbara Vergani. ki67 nuclei detection and ki67-index estimation: a novel automatic approach based on human vision modeling. *BMC bioinformatics*, 20(1):1–14, 2019. 2

[2] Khaled Benaggoune, Zeina Al Masry, Jian Ma, Christine Devalland, Leila Hayet Mouss, and Noureddine Zerhouni. A deep learning pipeline for breast cancer ki-67 proliferation index scoring. *arXiv preprint arXiv:2203.07452*, 2022. 2

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[4] Matthias Christgen, Sabrina von Ahsen, Henriette Christgen, Florian Länger, and Hans Kreipe. The region-of-interest size impacts on ki67 quantification by computer-assisted image analysis in breast cancer. *Human pathology*, 46(9):1341–1349, 2015. 2

[5] Mitch Dowsett, Torsten O Nielsen, Roger A'Hern, John Bartlett, R Charles Coombes, Jack Cuzick, Matthew Ellis, N Lynn Henry, Judith C Hugh, Tracy Lively, et al. Assessment of ki67 in breast cancer: recommendations from the international ki67 in breast cancer working group. *Journal of the National cancer Institute*, 103(22):1656–1664, 2011. 2

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059, New York, New York, USA, 2016. PMLR. 3, 6

[7] Johannes Gerdes, Hilmar Lemke, HEINZ Baisch, HANS-H Wacker, U Schwab, and H Stein. Cell cycle analysis of a cell proliferation-associated human nuclear antigen defined by the monoclonal antibody ki-67. *The journal of immunology*, 133(4):1710–1715, 1984. 1

[8] 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. 3

[9] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable ai systems for the medical domain?, 2017. 1

[10] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 2

[11] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision?, 2017. 3, 6

[12] S Lakshmi, Deepu Vijayasenan, David S Sumam, Saraswathy Sreeram, and Pooja K Suresh. An integrated deep learning approach towards automatic evaluation of ki-67 labeling index. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pages 2310–2314. IEEE, 2019. 2

[13] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. *Advances in neural information processing systems*, 23, 2010. 2

[14] Bo Li and Tommy Sonne Alstrøm. On uncertainty estimation in active learning for image segmentation. *arXiv preprint arXiv:2007.06364*, 2020. 3

[15] Vu-Linh Nguyen, Sébastien Destercke, and Eyke Hüllermeier. Epistemic uncertainty sampling. In *Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, October 28–30, 2019, Proceedings 22*, pages 72–86. Springer, 2019. 2

[16] Muhammad Khalid Khan Niazi, Caglar Senaras, Michael Pennell, Vidya Arole, Gary Tozbikian, and Metin N Gurcan. Relationship between the ki67 index and its area based approximation in breast cancer. *BMC cancer*, 18(1):1–9, 2018. 1

[17] Montse Pardàs, David Anglada-Rotger, Maria Espina, Ferran Marqués, and Philippe Salembier. Stromal tissue segmentation in ki67 histology images based on cytokeratin-19 stain translation. *Journal of Medical Imaging*, 10(3):037502, 2023. 7

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 4

[19] P. Salembier, A. Oliveras, and L. Garrido. Anti-extensive connected operators for image and sequence processing. *IEEE Transactions on Image Processing*, 7(4):555–570, 1998. 3

[20] Natalia Salpea, Paraskevi Tzouveli, and Dimitrios Kollias. Medical image segmentation: A review of modern architectures. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022. 2

[21] Pierre Soille et al. *Morphological image analysis: principles and applications*. Springer, 1999. 5

[22] Jordi Temprana-Salvador, Pablo López-García, Josep Castellví Vives, Lluís de Haro, Eudald Ballesta, Matias Rojas Abusleme, Miquel Arrufat, Ferran Marques, Josep R Casas, Carlos Gallego, et al. Digipatics: Digital pathology transformation of the catalan health institute network of 8 hospitals—planification, implementation, and preliminary results. *Diagnostics*, 12(4):852, 2022. 1

[23] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11): 4793–4813, 2021. 1

[24] Zoya Volynskaya, Ozgur Mete, Sara Pakbaz, Doaa Al-Ghamdi, and Sylvia L Asa. Ki67 quantitative interpretation: insights using image analysis. *Journal of pathology informatics*, 10(1):8, 2019. 2

[25] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 4

[26] Weidi Xie, J Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, 6(3): 283–292, 2018. 2, 5

[27] Mustafa Yousif, Paul J van Diest, Arvydas Laurinavicius, David Rimm, Jeroen van der Laak, Anant Madabhushi, Stuart Schnitt, and Liron Pantanowitz. Artificial intelligence applied to breast pathology. *Virchows Archiv*, pages 1–19, 2022. 1

[28] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2