Deep Learning with Uncertainty Quantification for Predicting the Segmentation Dice Coefficient of Prostate Cancer Biopsy Images

1st Audrey Xie
Media Arts & Sciences
Massachusetts Institute of Tech.
Cambridge, MA, USA
ahx@mit.edu

2nd Elhoucine Elfatimi
Pathology & Lab Medicine
University of California
Irvine, CA, USA
eelfatim@uci.edu

3nd Sambuddha Ghosal
Media Arts & Sciences
Massachusetts Institute of Tech.
Cambridge, MA, USA
sghosal@mit.edu

4nd Pratik Shah*
Pathology & Lab Medicine
Biomedical Engineering
University of California
Irvine, CA, USA
pratik.shah@uci.edu

2024 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-7488-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMLA61862.2024.00178

***Abstract*—Deep learning models (DLMs) can achieve state-of-the-art performance in histopathology image segmentation and classification, but have limited deployment potential in real-world clinical settings. Uncertainty estimates of DLMs can increase trust by identifying predictions and images that need further review. Dice scores and coefficients (Dice) are benchmarks for evaluation of image segmentation performance, but usually not evaluated with DLM uncertainty quantification. This study reports DLM's trained with uncertainty estimations, using randomly initialized weights and Monte Carlo dropout, to segment tumors from microscopic Hematoxylin and Eosin dye stained prostate core biopsy histology RGB images. Image level maps showed significant correlation [Spearman's rank ($p < 0.05$)] between overall and specific prostate tissue image sub-region uncertainties with model performance estimations by Dice. This study reports that linear models that can predict Dice segmentation scores from multiple clinical sub-region based uncertainties of prostate cancer can be a more comprehensive performance evaluation metric without loss in predictive capability of DLMs with a low root mean square error.**

***Index Terms*—Deep learning, Prostate tumor segmentation, Dice scores, Uncertainty, Real-world deployment**

I. Introduction

Deep neural networks (DNNs) are increasingly used for image detection, classification, and segmentation in medical applications [1], [2]. However, DNNs are considered sub-optimal for clinical settings which require risk analysis and uncertainty estimation in model predictions [3]. Many DNNs use softmax as the final activation function [4], but this leads to incorrect interpretations of confidence levels [5]. Bayesian neural networks offer an alternative by incorporating uncertainty into the model through a distribution over the weights [11], though they are computationally expensive and challenging to scale for high-dimensional images [5]. A more feasible approach is using dropout to approximate Bayesian inference and minimize the Kullback-Leibler divergence between the approximate distribution and the posterior of a deep Gaussian process [5]. Previous research has shown that optimizing small datasets with DNNs for medical image segmentation can lead to promising tool kits for clinical applications [8]. This work reports various uncertainty quantification methods to develop a comprehensive evaluation framework for prostate tumor segmentation models from Hematoxylin and Eosin (HE)-stained histopathology images used in clinical diagnosis [9], [10].

This study leverages Monte Carlo dropout (MCD) and backpropagation for uncertainty quantification to improve the reliability of deep learning model (DLM) predictions in adenocarcinoma prostate biopsy images. It provides detailed uncertainty maps that correlate with segmentation performance metrics, such as Dice coefficients, offering a more clinically relevant evaluation of model accuracy. The approach highlights the importance of tumor region-based uncertainty assessments to potentially help clinicians make better-informed decisions, increasing likelihood of DLM integration into clinical workflows. Additionally, this study introduces a novel algorithm designed to calculate region-specific uncertainties, effectively predicting segmentation performance through linear models. The goal is to promote the adoption of uncertainty quantification methods, ensuring accurate and reliable histopathology image analysis across various clinical workflows.

A. Related Work

Uncertainty estimation techniques and metrics for deep neural networks in non-medical image classification using Monte-Carlo sampling have been reported [5]. However, uncertainty estimations, image processing methods, and Dice coefficient evaluations have not been fully adapted to digital histopathology, such as prostate biopsy image segmentation for tumor and non-tumor regions of interest (ROIs). Another approach for uncertainty quantification uses Normalizing Flows to model data probability distributions without MC dropout, relying on flow-based models instead [7].

In medical applications, MCD has been used for classifying lung adenocarcinoma vs. squamous cell carcinoma without

*Corresponding author: Dr. Pratik Shah Ph.D.(pratik.shah@uci.edu)

1946-0759/24/$31.00 ©2024 IEEE
DOI 10.1109/ICMLA61862.2024.00178

segmentation or Dice score evaluations of ROIs [14]. Other studies trained MCD and Backpropagation-based DLMs using Dice-loss for tumor segmentation in oropharyngeal cancer PET/CT images [15], and developed a Dice-risk uncertainty measure that was outperformed by the coefficient of variation [15]. However these studies did not calculate post hoc Dice scores for different ROIs, which is a focus of this paper. Additionally, a 3D U-Net using MCD generated uncertainty maps for lung cancer segmentation on CT images but did not evaluate Dice scores on prostate HE images [16]. Linear regression models for predicting single Dice scores from grayscale CT scans based on image-level uncertainty have been reported [19]. In a study on skin lesion classification, MCD-based uncertainty estimation led to segmentation errors, primarily due to class imbalances [13], [23]. To our knowledge, no studies have reported the relationship between DLM uncertainties and the segmentation of different clinical ROIs in histopathology images of the prostate or other tissues using individual Dice coefficients. Prior works in tumor classification and segmentation using deep learning [20]–[22] could benefit from correlating uncertainty-based Dice scores with clinical outcomes, as done in this study.

B. Summary of contributions

In this study, a DNN was trained using MCD to calculate pixel-level uncertainty for accurate segmentation of prostate biopsy tumors from HE-stained RGB images. The model's performance was evaluated by five-fold cross-validation and performance metrics (see Methods and Table I-B). Maximum (indicating the highest uncertainty), minimum (reflecting high confidence), mean (providing an overall measure of the model's uncertainty across all pixels), and standard deviation (indicating the variability in uncertainty estimates) together provide a comprehensive view of prediction confidence (Table I-B).

The images were subdivided into three ROIs: tumor, non-tumor tissue, and non-tissue areas. Statistically significant correlations ($p < 0.05$) between Dice scores and uncertainty were found by comparing the model's output with ground truth segmentation masks, and three individual region-based and overall uncertainty of each image were then calculated. A new algorithm (Algorithm 1) was proposed to calculate region-specific uncertainty, and linear models were developed to predict Dice scores using clinical region-specific uncertainties derived from the MCD model. These linear models use constants and uncertainty measures from the clinical ROIs, either individually or combined, to accurately predict model performance (Dice) (Eq. 2 and 3). This study reports that although image-level uncertainty can estimate Dice predictions (Eq.5), sub-clinical region-based uncertainties provide a more precise evaluation without losing predictive capabilities with low Root Mean Square Error (RMSE).

II. Data description

Prostate core biopsy images from Gleason2019 dataset (Grand Challenge for Pathology at MICCAI 2019) [21] contains 244 tissue microarray RGB images (5120 × 5120 resolution) of HE-stained prostate tissue. Expert pathologists annotated each image with Gleason tumor grades (1-5) and generated Ground-truth (GT) segmentation masks. Gleason grades 1 and 2, which are rare, were labeled as non-tumor tissues (black pixels (class 0), while Regions of Gleason Grades 3, 4, and 5 were considered tumors (white pixels (class 1)). The dataset includes 224 images with tumors and 20 without. A detailed description of data pre-processing and training of DLMs with five-fold cross-validation used in this study can be found in [20].

TABLE I
PERFORMANCE OF A DEEP LEARNING MODEL AUGMENTED WITH MONTE-CARLO DROPOUT FOR TUMOR SEGMENTATION FROM PROSTATE CORE BIOPSY MICROSCOPIC H&E IMAGES, INCLUDING AUROC - AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC , TPR – TRUE POSITIVE RATE, FPR – FALSE POSITIVE RATE, STD. DEV. – STANDARD DEVIATION.

Performance			
Dice	0.8809	TPR	0.7730
AUROC	0.8662	FPR	0.0008
Uncertainty			
Maximum	0.0592	Mean	0.0159
Minimum	0.0014	Std. dev.	0.0130

III. Methods

A Bayesian VGG-UNet CNN augmented with MCD was trained on prostate core biopsy images from the Grand Challenge for Pathology at MICCAI Gleason 2019 dataset [21]. The network consists of an encoder-decoder structure [20], where the encoder uses a modified version of the VGG-16 architecture without fully-connected layers [22]. Uncertainty quantification was applied in the decoder to mitigate noise during the encoding process, with MCD used to approximate Bayesian inference by minimizing the posterior in a deep Gaussian Process [5]. Additionally, Bayes-by-backprop (backprop) was employed to quantify uncertainty by evaluating the gradient of the loss function relative to the input images. Both MCD and backprop improved segmentation performance by identifying uncertain regions in the model's predictions. Model training involved hyperparameter optimization, five-fold cross-validation, and advanced performance metrics to ensure a comprehensive evaluation of the segmentation results.

A. Model training

The prostate core biopsy data was split into 80% for training (≈ 197 images) and 20% for testing (≈ 47 images). The model was trained for 50 epochs with a batch size of two, using the Adam optimizer with a learning rate set at $\alpha_{lr} = 0.001$. Input images were resized to 1024 × 1024 (W × H), while output images were set to 512 × 512. Model training and evaluation were performed on an NVIDIA GeForce RTX 3090 Ti with 24GB GPU memory. The workflow for model implementation and processing with uncertainty integration approach is outlined in Fig.1, starting with input images and applying the DLM MCD to obtain outputs.

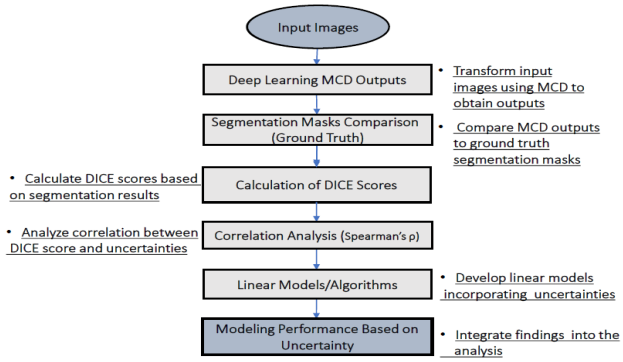

Fig. 1. Integrated analysis flow: Sequential relationships in histopathology image analysis with deep learning, uncertainty maps, Dice score, and Monte Carlo Dropout (MCD).

The outputs were compared to ground truth masks, and Dice scores were calculated. Correlation analysis established the relationship between Dice scores and uncertainty estimations, leading to linear models for predictions. This workflow integrates segmentation performance of clinically relevant regions in histopathology images to assess overall model performance.

B. Model prediction and evaluation

In the process of evaluating the models predictive performance, occurrences of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) across each image output were calculated. Specifically, a TP was recorded each time the model correctly identified a tumor pixel, whereas a TN was noted when it correctly identified pixels as non-tumor or non-tissue. Conversely, an FN occurred when tumor pixels present in the ground truth were overlooked by the model, and an FP was registered when pixels were incorrectly classified as tumor despite being non-tumor or non-tissue in the ground truth. The True Positive Rate (TPR) and False Positive Rate (FPR), alongside the Area Under the Receiver Operating Characteristic (AUROC) and the Dice Score (Dice) formed the basis for calculating key performance metrics. The Dice Score, a critical metric, assesses the congruence between the clinical ground truth labels and the model's segmentation accuracy and was subsequently used to refine linear model estimations. The median values of Dice, AUROC, TPR, and FPR for the ensemble of 47 test images are documented in **(Table I-B)**. Alongside these metrics, detailed analysis of pixel-level uncertainties for each image was perfomed. This involved conducting 50 Monte Carlo (MC) iterations to measure the uncertainty range, defined by the 67th and 33rd percentiles of the sigmoid output values of the model. These uncertainty values were then visualized through heatmaps (Fig.2, and 3) and quantitatively captured in **(Table I-B)**. From the MC data, the central tendencies for the mean uncertainties specific to each image were derived, leading to the calculation of an overall mean uncertainty for all test images. This comprehensive analysis revealed uncertainty levels ranging from a minimum of 0, indicating total certainty in predictions, to a maximum of 1, signifying complete uncertainty for further interpretation and fine-tuning the model's predictive reliability and accuracy.

Mathematically, the model prediction for a given MCD model M and input test image X_{test_i} is calculated by:

$$O^j_{M_p} = \text{argmax}(M_p(X_{\text{test}_i}))$$

for the j^{th} Monte Carlo (MC) iteration, where M_p is the model M in the testing phase. The final model output O_{M_i} is generated by averaging all 50 MC iteration outputs:

$$O_{M_i} = \frac{1}{\alpha}\sum_{j=1}^{\alpha} O^j_{M_p} \tag{1}$$

where $\alpha = 50$ is the number of MC iterations. In the segmentation process, a value-based thresholding method is employed to determine the class to which each pixel belongs. A pixel is assigned to 'class 0' (non-tumor), with a pixel value of '0' in the binary scale, if the mean value across all MC iterations is ≤ 0.95. Conversely, it is assigned to 'class 1' (tumor), with a pixel value of '1', if the mean value is > 0.95. Thus, the final pixel classification in O_{M_i} can be expressed as:

$$\text{if } |p| > 0.95, \text{ then } |p| = 1 \text{ (tumor)},$$

$$\text{if } |p| \leq 0.95, \text{ then } |p| = 0 \text{ (non-tumor)}.$$

To construct the corresponding uncertainty map, Unc_{map_i}, each MC iteration yields an uncertainty value defined by:

$$Unc^j_{M_p} = \max(M_p(X_{\text{test}_i})).$$

The overall uncertainty map is then calculated as the difference between the 67$^{\text{th}}$ and 33$^{\text{rd}}$ percentiles of these values:

$$Unc_{map_i} = P_{67}(\{Unc^j_{M_p}\}^{\alpha}_{j=1}) - P_{33}(\{Unc^j_{M_p}\}^{\alpha}_{j=1}) \tag{2}$$

IV. Results and Discussion

A. Model Performance and Uncertainty Analysis

Five-fold cross-validation was used for segmentation of prostate tumors and estimating segmentation uncertainties of the trained MCD model. Confidence intervals (CI) were calculated using the empirical bootstrap method with n = 5,000 simulations. Values for AUROC for all 47 test images were [0.8938 (95% CI: 0.8582 - 0.9019)], Dice [0.8987 (95% CI: 0.8561 -0.9095)], TPR [0.8382 (95% CI: 0.7792 - 0.8980)] and FPR [0.0279 (95% CI: 0.0296 -0.0829)]). The model in this study demonstrates superior generalization power, with a reported overall accuracy of 93.10% and a loss of 0.2105. These results indicated that the MCD model achieved high performance (Table I-B) for prostate tumor segmentation. The overall uncertainty prediction were on the lower end (less than 1%), with select individual ROI's within an input image showing uncertainties as high as 20%. The segregated uncertainty maps for TFS models, augmented with MCD or backprop, are shown in Fig 2. The uncertainty maps are represented as heatmaps, deeper red indicates high uncertainty, and deeper blue indicates low uncertainty. These maps include binary

segmentation masks and uncertainty estimates for tumor, non-tumor, and non-tissue regions. The overall uncertainty maps also show that tumor-containing images throughout the tissue regions exhibited the least uncertainty, with most of detected uncertainty at the tissue and non-tissue boundary regions Fig. 2. Images without tumors showed higher false positive regions and greater non-tumor tissue uncertainty Fig. 2. These visualizations can identify histopathology images with tumors by interpreting high segmentation uncertainties and the reliability of the model's output to inform clinical decision-making.

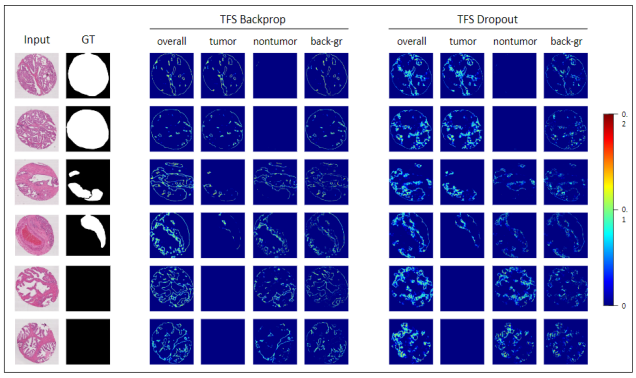

Fig. 2. Segmentation and Uncertainty Maps for Backprop and Monte-Carlo Dropout Augmented Trained-from-Scratch (TFS) models. "GT" – clinical ground-truth, "back-gr" – non-tissue regions.

A comparative analysis of backpropagation vs. dropout uncertainty was modeled with linear regression to predict the Dice Coefficient for histopathology image segmentation. The evaluation used metrics like the coefficient of determination (R^2), significance tests, and RMSE. While backpropagation showed a higher R^2 value (0.502049), indicating a superior overall fit, the uncertainty dropout model had better predictive accuracy with a lower RMSE (0.062141 vs. 0.07486). RMSE was prioritized in this study for its emphasis on accurate predictions and nuanced uncertainty modeling, The RMSE metric identified the linear regression model between uncertainty dropout and Dice scores for its superior predictive accuracy and advanced uncertainty modeling, aligning with the study's focus on precision in predictions.

B. Qualitative evaluation

Figure 3 shows four tumor containing and two non-tumor prostate images from the test data and MCD model output segmentation masks and corresponding uncertainty heatmaps from the trained model. Model predictions O_{M_i} and overall uncertainty map, Unc_{map_i} obtained using step 2 of Algorithm 1 are shown in columns C and D respectively. The region based uncertainty maps generated using step 2 of Algorithm 1, Unc_{iT}, Unc_{iNT} and Unc_{iNTi} are shown in columns E, F and G respectively. Images 1 to 4 contained tumors and images 5 and 6 did not contain ground-truth tumor signatures. Predictions for images 1 and 2 exhibited negligible false positive or false negative regions, images 3 and 4 showed low false positives and good segmentation accuracy (Fig. 3). Images 5 and 6 exhibited the highest false positive regions as related instances were less frequent in the training data compared to images that had tumors in them (Fig. 3). Overall uncertainty maps (column D) showed that images with tumors all throughout the tissue regions (images 1 and 2) exhibited the least uncertainty (Fig. 3). Majority of the detected uncertainty for these images were at the tissue and non-tissue boundary regions. The MCD model showed no uncertainties in the non-tissue regions except for the junction between tissue and non-tissue regions(Fig. 3). This indicated that the model has learnt the non-tissue region well and is uncertain either at the non-tissue and tissue junctions and more frequently at the non-tumor tissue and tumor tissue region intersections.

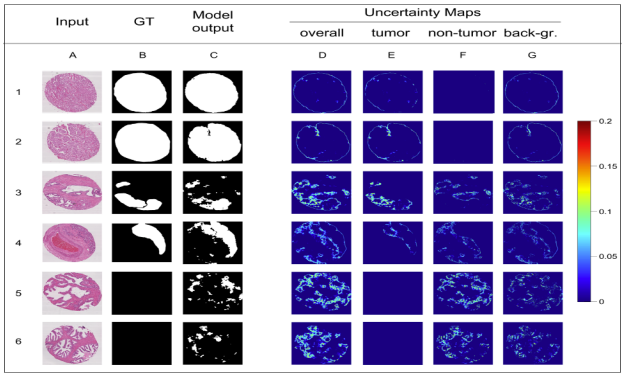

Fig. 3. Visualization of model output segmentation and region-based uncertainty maps for prostate tumor segmentation using a Monte Carlo dropout (MCD) deep learning model trained on Hematoxylin and Eosin (H&E) stained prostate biopsy images. **Columns left to right: A**- Input RGB image; **B**- Clinical ground-truth binary mask; **C**- MCD output segmentation binary mask; **D**- MCD model uncertainty estimate maps; **E**- Tumor tissue uncertainty map; **F**- Non-tumor tissue uncertainty map; **G**- Non-tissue uncertainty map. Color bar shows model uncertainty, with red for higher and blue for lower uncertainty (minimum value 0). "GT" – clinical ground-truth, "back-gr" – non-tissue regions.

This was especially true for regions indicated by the ground-truth as containing tumorous tissues in images 2 and 3 in figure 3. For images 5 and 6 the absence of tumors gives the tissue island like appearance which may lead to a greater degree of confusion for the model and non-tumor tissue uncertainty. Visualizing these individual ROI-based uncertainty maps (Fig. 3), also helped in identifying regions of high uncertainties and make informed decisions on whether to discard a model's output or not based on where the higher uncertainties arose from. For example in image 3 in Fig. 3, high uncertainties were observed in the tumor region indicating that the model may show poor performance within tumor regions in similar images during inference, while similar uncertainties were observed in image 2 for the tumor and non-tissue regions indicating that predictions may be acceptable for cases when differential diagnosis depended on disease pixels.

C. Performance Modeling Through Uncertainty Quantification

This study proposes Algorithm 1 to calculate estimated performance (Dice) of a model using uncertainties for clinical

Step 1: -Model predictions and associated uncertainty maps

Input: Trained MCD model M; Test data X_{test} with n images and ground truth X_{GT}.

Output: Segmentation output, O_{M_i} and uncertainty map, Unc_{map_i} for each X_{test_i}. Thus, $\forall$ i = [1, 2,....., n] images,

$$X_{Dice_i} = Dice(O_{M_i}, X_{GT_i}), List_{Dice} = [X_{Dice_i}]_{i=1}^{n} =$$

$$List_{Unc} = [Unc_{map_i}]_{i=1}^{n}$$

Step 2: Generate ROI uncertainty:

Input: Given a test image, X_{test_i} and a ground-truth mask, X_{GT_i}, the following were generated:

• X_{test_i}-tumor = X_{GT_i}, where 0 represents non-tissue and 1 (255 in grayscale) represents tissue regions with or without tumors.

• X_{test_i}-non-tumor = X_{test_i}-tumor^C, complement of X_{test_i}-bin, a binary image where 1 is non-tissue and 0 is tissue .

Output: Uncertainty maps corresponding to these regions are generated as follows: :

• Unc_{iL} = Unc_{test_i}-tumor = (X_{test_i}-tumor $\odot$ Unc_{map_i}); $List_{T_{unc}} = [mean(Unc_{iL})]_{i=1}^{n}$.

• Unc_{iNL} = Unc_{test_i}-non-tumor = (X_{test_i}-non-tumor $\odot$ Unc_{map_i}); $List_{NT_{unc}} = [mean(Unc_{iNL})]_{i=1}^{n}$. The symbol $\odot$ denotes the Hadamard product.

Step 3: Generate linear regression models:

Input: Dependent variable, $List_{Dice}$ = Y_{Dice} (say) = $[y_{Dice_i}]_{i=1}^{n}$ for n images, and independent variables:

• $List_{Unc}$ = X_0 (say) = $[x_{0i}]_{i=1}^{n}$ (overall uncertainty maps as presented in Fig. 3 column D),

• $List_{T_{unc}}$ = X_1 (say) = $[x_{1i}]_{i=1}^{n}$ (tumor tissue uncertainty maps as presented in Fig. 3 column E) ,

• $List_{NT_{unc}}$ = X_2 (say) = $[x_{2i}]_{i=1}^{n}$ (non-tumor tissue uncertainty maps as presented in Fig. 3 column F) and

• $List_{NT_{unc}}$ = X_3 (say) = $[x_{2i}]_{i=1}^{n}$ (uncertainty maps for non-tissue pixels without tissue as shown in Fig. 3 column G) and that predict Y_{Dice}, where n is the total number of images under consideration.

Output: Model Y_{Dice} in terms of X_1, X_2 and X_3 as follows:

$$Y_{Dice} = \alpha_0 + \alpha_1 * X_1 + \alpha_2 * X_2 + \alpha_3 * X_3 \quad (3)$$

For individual ROIs:

$$Y_{Dice} = \beta_0 + \beta_1 X_1 \quad (4(i))$$

$$Y_{Dice} = \gamma_0 + \gamma_1 X_2 \quad (4(ii))$$

$$Y_{Dice} = \delta_0 + \delta_1 X_3 \quad (4(iii))$$

and for the overall uncertainty, X_0:

$$Y_{Dice} = \theta_a + \theta_b X_0 \quad (5)$$

Algorithm 1: ROI-Unc: Modelling deep learning performance with clinical-class based image region specific uncertainty

ROIs in a medical image. Eq. 3 calculated the Dice of the segmentation model based on a linear combination of tumor-tissue, nontumor tissue and non-tissue uncertainties, and equations 4(i), 4(ii), and 4(iii), calculated the Dice of the segmentation model from the three individual ROIs. Two subsets of test data were used for calculating linear models and for obtaining the values of the coefficients for equations 3 to 5 and are described below:

1) Tumor-containing images: Forty four images that contained tumor labels from the test data set were used as the first subset. From these tumor containing images, we obtained α_0 = 1.0036, α_1 = -12.2397, α_2 = -19.5844 , and α_3 = -6.2364; for tumor uncertainty (Spearman's correlation coefficient (ρ) = -0.4878), β_0 = 0.9264 and β_1 = -7.9295; for non-tumor uncertainty (ρ = -0.5012), γ_0 = 0.9011 and γ_1 = -13.9173, for non-tissue uncertainty (ρ = -0.5969), δ_0 = 0.9702 and δ_1 = -26.9116, and for the overall uncertainty (ρ = -0.8496), θ_a = 1.0074 and θ_b = -14.9116. Low RMSE of ≤ 0.1 was calculated for all four linear models for the tumor containing images.

2) Tumor-free images: Three images from the test data and 17 images from the training data that did not contain any tumor labels were merged into the second subset. From these tumor-free images α_0 = 0.9991, α_1 = 0 (which is expected since the tumor regions did not contribute to the model uncertainties or performance), α_2 = -12.306 and α_3 =7.2247; for non-tumor uncertainty (Spearman's correlation coefficient (ρ = -0.8496), γ_0 = 0.9993 and γ_1 = -8.5907; for non-tissue uncertainty (ρ = -0.8496), δ_0 = 0.9993 and δ_1 = -16.02, and for the overall uncertainty (ρ = -0.8496), θ_a = 0.9994 and θ_b = -5.6296 were calculated. The RMSE values for all four linear models were ≤ 0.16. The RMSE values for the tumor-free images for four linear models were even lower at 0.014. For all derived coefficients, except for α_3 for the tumor-free images, negative coefficient values in the majority of the calculations indicated that higher uncertainty in the MCD model predictions decreased its performance. From the coefficients derived for Eq. 3, the significance of each clinical region in determining overall model performance becomes clear. A more negative coefficient implies a greater reduction in Dice scores. For tumor-containing images, non-tumor tissue region uncertainties had the highest coefficient value, meaning they had the largest negative impact on performance. The MCD model was most affected by non-tumor tissue uncertainty, followed by tumor-tissue regions, with non-tissue pixels contributing about half as much. In tumor-free images, non-tumor tissue uncertainty played a more substantial role than pixels without tissue, as indicated by the positive α_3 coefficient.

V. Discussion

Uncertainties in DLMs can significantly impact predictions, especially in critical ROIs such as tumor, non-tumor, and non-tissue areas. These uncertainties are crucial for model reliability in clinical settings. In this study, MCD and back-propagation were used to quantify uncertainties, with Dice scores evaluating segmentation accuracy (Figs.2 and 3). Tumor regions exhibited lower uncertainties, while higher uncertainties occurred at tissue boundaries. Non-tumor images showed more false-positive uncertainties in non-tumor regions, demonstrating the method's potential for real-world applications, even in scenarios where ground truth may be limited, as outlined in Eq.5). A new approach (Algorithm 1) was developed to calculate region-specific uncertainties and predict

performance. The non-tissue region (X_3) contributed to overall uncertainty but had less effect on the Dice score than non-tumor tissue regions. The prediction of model performance, based on Dice scores, was guided by overall uncertainty, as described in Equation 5. Backpropagation, combined with MCD, offered deeper insights into areas of low prediction confidence, supporting more informed clinical decision-making. Future research could apply these uncertainty quantification techniques to other medical imaging datasets for greater model reliability.

VI. Conclusion and Limitations

This study demonstrates that region-based linear models can predict deep learning performance (Dice) using clinically meaningful uncertainty estimates (Equations 3 -5), Previous studies typically focused on either Dice or overall uncertainty, but this work uniquely combines both, offering a more comprehensive evaluation of model performance, especially for prostate adenocarcinoma segmentation. To our knowledge, this is the first time, three distinct clinical region-based uncertainties—tumor, non-tumor, and non-tissue regions—were decoupled and analyzed; showing that linear models accurately predict performance across these regions. Spearman's rank correlation ($p < 0.05$) revealed a strong negative correlation between increased uncertainty and lower Dice scores. Algorithm 1 was introduced for unsupervised performance estimation, allowing prediction of Dice scores and generation of uncertainty maps, aiding in clinical decision-making.Although the study is currently limited to the prostate biopsy dataset, MCD, and Bayes-by-backprop techniques, it lays the groundwork for applying these uncertainty estimation methods to other datasets and clinical contexts. Future research should expand on these findings, exploring broader applications to enhance the reliability and accuracy of deep learning models in healthcare.

VII. Data, Code and Model Availability

All resources, results, code, models, and data from this study are available on GitHub https://github.com/Prof-Pratik-Shah-Lab-UCI/Prostate_Project_2024 upon request to the corresponding author.

References

[1] J. Lee, S.Jun, Y. W. Cho, H. Lee, G. Bae Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: general overview," Korean journal of radiology, vol. 18, no. 4, pp. 570–584, 2017.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F.Ciompi, M. Ghafoorian, J. A. V. D. Laak, B. V. Ginneken, and C. I S anchez, "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60–88, 2017

[3] G. Choy, O. Khalilzadeh, M. Michalski, S. Do, A. E Samir, O. S Pianykh, J. Raymond Geis, P. V Pandharipande, James A Brink, and K. J Dreyer, "Current applications and future impact of machine learning in radiology," Radiology, vol. 288, no. 2, pp. 318–328, 2018.

[4] J. S Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in Advances in neural information processing systems, 1990, pp. 211–217

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

[6] P. Shah, J. Lester, JG. Deflino ,V. Pai, "Responsible Deep Learning for Software as a Medical Device." arXiv preprint arXiv:2312.13333. 2023 Dec 20.

[7] R. Selvan, F. Faye, J. Middleton, A. Pai, "Uncertainty quantification in medical image segmentation with normalizing flows." In Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11, pp. 80-90. Springer International Publishing, 2020.

[8] S. Ghosal, P. Shah, "A deep-learning toolkit for visualization and interpretation of segmented medical images." Cell Reports Methods. 2021 Nov 22;1(7).

[9] A. Bayat, C. Anderson, P. Shah, "Automated end-to-end deep learning framework for classification and tumor localization from native non-stained pathology images." InMedical Imaging 2021: Image Processing 2021 Feb 15 (Vol. 11596, pp. 43-54). SPIE.

[10] A. Rana, A. Lowe, M. Lithgow, K. Horback , T. Janovitz, Da. A. Silva, H. Tsai, V. Shanmugam, A. Bayat, P. Shah, "Use of deep learning to develop and analyze computational hematoxylin and eosin staining of prostate core biopsy images for tumor diagnosis". JAMA network open. 2020 May 1;3(5):e205111-.

[11] C. M. Bishop, "Bayesian neural networks," Journal of the Brazilian Computer Society, vol. 4, pp. 61–68, 1997.

[12] M. Kampffmeyer, A. B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 1–9

[13] P. V. Molle, T. Verbelen, C. D Boom, B. Vankeirsbilck, J. D. Vylder, B. Diricx, T. Kimpe, P. Simoens, and B. Dhoedt, "Quantifying uncertainty of deep neural networks in skin lesion classification," in Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures, pp. 52–61. Springer, 2019

[14] J. M. Dolezal, A. Srisuwananukorn, D. Karpeyev, S. Ramesh, S., Kochanny, B. Cody, A. T. Pearson. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. Nature communications, 13(1), 6572. 2022

[15] J. Sahlsten, J. Jaskari, KA. Wahid, S. Ahmed, E. Glerean, R. He, BH. Kann , A. Mäkitie , CD. Fuller, MA. Naser , K. Kaski; Application of simultaneous uncertainty quantification for image segmentation with probabilistic deep learning: Performance benchmarking of oropharyngeal cancer target delineation as a use-case. medRxiv. 2023 Feb 24.

[16] FC. Maruccio , WS. Eppinga, MH. Laves, R. Fonolla Navarro, M. Salvi, F. Molinari , Papaconstadopoulos P; "Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation". Physics in Medicine and Biology. 2023.

[17] A. Vasiliuk, D. Frolova,M. Belyaev, B. Shirokikh, " Exploring Structure-Wise Uncertainty for 3D Medical Image Segmentation". arXiv preprint arXiv:2211.00303.Nov 1. 2022.

[18] K. Hoebel, V. Andrearczyk, A. Beers, J. Patel, K. Chang, A. Depeursinge, H. M. uller, and J. K. Cramer, "An exploration of uncertainty information for segmentation quality assessment," in Medical Imaging 2020: Image Processing. International Society .

[19] K. Hoebel, V. Andrearczyk, A. Beers, J. Patel, K. Chang, A. Depeursinge, H. M. uller, and J. K. Cramer, "An exploration of uncertainty information for segmentation quality assessment," in Medical Imaging 2020: Image Processing. International Society for Optics and Photonics, 2020, vol. 11313, p. 113131K

[20] P. F. O. Ronneberger and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, p. 234–241

[21] grand challenge.org, "Gleason 2019 homepage,"URL: https://gleason2019.grandchallenge.org/, 2019

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014., 2014.

[23] M.Combalia; F.Hueto ,S.Puig, Jo. Malvehy, "Uncertainty Estimation in Deep Neural Networks for Dermoscopic Image Classification" IEEEE Explore, 2020.