Uncertainty Estimation for Tumor Prediction with Unlabeled Data

Juyoung Yun[1], Shahira Abousamra[1], Chen Li[1], Rajarsi Gupta[2], Tahsin Kurc[2],
Dimitris Samaras[1], Alison Van Dyke[3], Joel Saltz[2] and Chao Chen[2]

[1]Stony Brook University, Department of Computer Science, USA

[2]Stony Brook University, Department of Biomedical Informatics, USA

[3]National Cancer Institute, Maryland, USA

Abstract

Estimating uncertainty of a neural network is crucial in providing transparency and trustworthiness. In this paper, we focus on uncertainty estimation for digital pathology prediction models. To explore the large amount of unlabeled data in digital pathology, we propose to adopt novel learning method that can fully exploit unlabeled data. The proposed method achieves superior performance compared with different baselines including the celebrated Monte-Carlo Dropout. Closeup inspection of uncertain regions reveal insight into the model and improves the trustworthiness of the models.

1. Introduction

Modern digital pathology has witnessed significant progress in recent years. Harnessing the learning power of deep neural networks, researchers have developed advanced analysis methods for histopathology images [6, 17, 20]. However, despite the strong prediction power of deep learning models, doctors and caregivers remain concerned when deploying them in clinical settings for diagnosis and prognosis. The lack of trust is due to various factors. The black-box deep neural networks, with millions or even billions of parameters, can potentially overfit and make over-confident predictions even when they are wrong. Furthermore, annotating these large histology images with fine-grained semantic labels such as tumor, tumor infiltrating lymphocytes (TIL), etc., is extremely time-consuming and error-prone. Thus, the models are often trained with limited and potentially noisy labels. To use these models safely, we need information beyond the prediction.

In this paper, we propose to compute model uncertainty for histopathology images. In recent years, uncertainty estimation has caught attention in the machine learning community [12]. Model uncertainty empowers different downstream analysis tools. In an active learning pipeline, uncertainty allows one to select uncertain data for expert verification. In semi-supervised learning, predictions on unlabeled data with low uncertainty can be used as pseudo-labels and be included in the training set. Finally, being able to visualize uncertainty and visualize the data on which the model is confused can significantly increase transparency and thus trustworthiness of AI models, especially in healthcare [5].

To estimate uncertainty, earlier works use an ensemble of models and aggregate their prediction. For example, Monte-Carlo Dropout (MCDropout) method obtains an ensemble of models by randomly knocking out neurons [11]. Alternatively, one may use a surrogate function to approximate the uncertainty, and then train the neural network to fit the surrogate function. For example, Moon et al. [22] define the surrogate function as the probability of correctness of each training datum. Despite the strong performance, the correctness surrogate function, however, requires every training data to be labeled, and thus cannot be used to learn from large amount of unlabeled data, as in the case of digital pathology. Li et al. [21] proposed to use consistency as a surrogate function for uncertainty calibration. The method uses the consistency of the model's prediction on a datum through the training process. It does not require labels, and thus can fully exploit the large amount of unlabeled data.

In this paper, we propose a novel method for uncertainty estimation in digital pathology. To fully utilize both the labeled and unlabeled data, we propose to combine both the correctness surrogate function and the consistency surrogate function for uncertainty estimation. We demonstrate the power of the proposed method in the task of tumor prediction, which can be used to measure Tumor-TIL spatial relationships, and provide important information for diagnosis and prognosis [20]. We show that the proposed method is superior than previous methods including the celebrated MCDropout. Furthermore, we carry out a thorough quantitative and qualitative analysis. Closeup inspection of uncertain regions reveals insight into the model and improves the trustworthiness of the models.

2. Related Works

Patch-wise Cancer Prediction. In the field of digital pathology, especially with a focus on breast cancer histopathology, the evolution from traditional handcrafted feature extraction to the adoption of Convolutional Neural Networks marks a significant advancement. Veta et al. [26] utilized fast radial symmetry and marker-controlled watershed segmentation for nuclei extraction, while Basavanhally et al. [1] employed a geodesic-based active contour model focusing on both morphological and textural features of segmented nuclei. Spanhol et al. [25] explored the BreaKHis dataset to classify histopathological images as benign or malignant, indicating the challenge of maintaining accuracy at higher magnifications. Cruz-Roa et al. [6] proposed a CNN model for automatic classification of invasive ductal carcinoma in whole slide images, which represented a significant step towards automating the differentiation between invasive and non-invasive images. A detailed study by Le et al. [20] focuses on utilizing CNNs for analyzing breast cancer whole slide images (WSIs), emphasizing the significance of spatial relationships between tumor regions and tumor-infiltrating lymphocytes (TILs).

Uncertainty Estimation. Uncertainty estimation in digital pathology can be approached through confidence calibration and ordinal ranking. Confidence calibration methods, as discussed in seminal works by Platt [24], Guo et al. [16], and others, focus on aligning a model's confidence with the actual probability of correct predictions. Ordinal ranking, on the other hand, prioritizes the order of confidence levels among predictions, as explored by Geifman and El-Yaniv [13] and Lakshminarayanan et al. [19], to ensure the model's predictions are consistently reliable. For reducing uncertainty of training model, there are several approaches. MC-Dropout, as outlined by Gal and Ghahramani [11], offers a practical framework for uncertainty estimation by simulating Bayesian inference, allowing for dynamic uncertainty evaluation. Correctness Ranking Loss [22] emphasizes the model's accuracy on the available labeled dataset, making it an essential method for enhancing the quality of predictions in scenarios where direct supervision is limited. Consistency Ranking Loss, introduced by Li et al. [21], is particularly effective in semi-supervised settings, where it utilizes both labeled and unlabeled data to improve the model's confidence estimation.

Our work integrates these methodologies to advance the field of digital pathology, aiming to improve model accuracy and the reliability of automated diagnostic systems. By exploring these cutting-edge approaches to uncertainty estimation, we contribute to the ongoing development of more precise and trustworthy AI tools in medical imaging.

3. Method

We propose to train uncertainty calibration using the consistency ranking method [21] paired with the tumor patch classification method [20]. This enables us to take advantage of the large amounts of freely available unlabeled data to get better calibrated uncertainty.

3.1. Semi-supervised Learning with Uncertainty Calibration

We train our tumor prediction model on patches extracted from WSIs. The model learns to predict whether a patch is tumor positive or negative. During inference, a WSI is tiled into non-overlapping patches and the predicted patch-wise probabilities form the tumor predicted probability map.

The training dataset is comprised of n labeled patches and p unlabeled patches and is represented as $D = (X, Y)$, and U, respectively, such that, $X = \{x_1, ..., x_n\}$ is the set of labeled image patches, $Y = \{y_1, ..., y_n\}$ is the set of corresponding labels, $y_i \in \{0, 1\}$, and $U = \{x_{n+1}, ..., x_{n+p}\}$ is the set of unlabeled image patches.

Consistency Ranking Loss. In this paper, we explore the use of the Consistency Ranking Loss [21] to train our tumor prediction model and calibrate the model's uncertainty. The consistency ranking loss uses the consistency of the model's predictions on the training patches to estimate the uncertainty of the model. For each patch, highly consistent prediction across training epochs indicates higher confidence and thus lower uncertainty. Similarly, less consistency or more fluctuations in the model's prediction for a patch indicates more uncertainty. Hence this relative measure of consistency can act as a surrogate function of the model's uncertainty. The consistency loss tries to make the predicted probability values reflect this uncertainty surrogate function. Since the consistency measure does not rely on data labels, it can be applied on unlabeled patches, making it suitable for semi-supervised learning.

More formally, we train our model $f(x; W) : X \rightarrow [0, 1]$ using both labeled (X, Y) and unlabeled data U. For every data point x_i, its predicted classification $\hat{y}_i^t = \arg\max_{y \in \{0,1\}} f(x_i; W^t)$, where W^t denotes the model's weights at the t-th epoch.

We define a sample's training consistency as the frequency of obtaining the same prediction in consecutive epochs throughout the training:

$$c_i = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{1}\{\hat{y}_i^t = \hat{y}_i^{t+1}\} \tag{1}$$

For each patch x_i from the collective set $X \cup U$, we use κ_i to denote the model's maximum softmax output for x_i. If the training consistency of one data point, c_i, is less than that of another, c_s, then the κ_i should also be less than κ_s, thereby

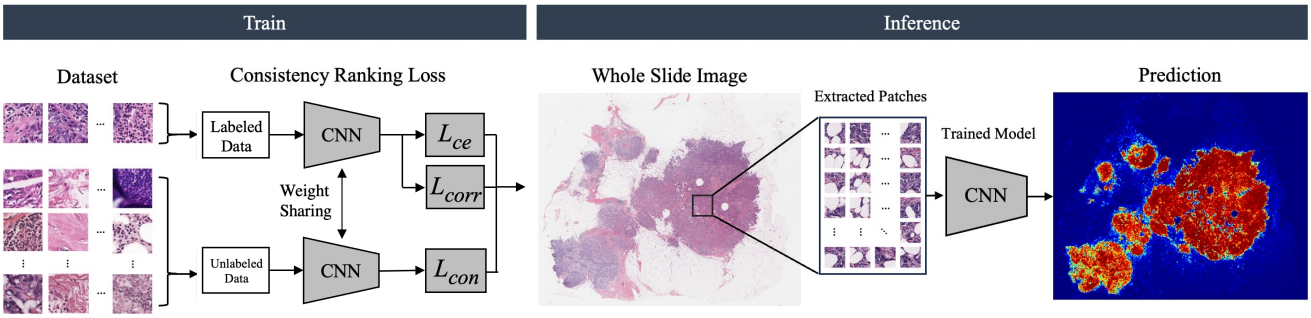

Figure 1. Schematic representation of the patch-wise cancer prediction workflow. The diagram illustrates the sequential phases involved in processing and classifying histopathological images for cancer detection. L_{ce}, L_{corr}, and L_{con} each represent Cross Entropy Loss, Correctness Ranking Loss [22], and Consistency Ranking Loss [21], respectively.

maintaining a consistent ranking between the training consistency and the model's confidence in its predictions. The formula of the Consistency Ranking Loss is:

$$\mathcal{L}_{cons}(f) = \sum_{s=1}^{n+p} \sum_{i=1, c_i<c_s}^{n+p} max\{0, (c_s - c_i) - (\kappa_s - \kappa_i)\} \tag{2}$$

The goal during training is to adjust the confidence estimator κ so that the difference in confidence closely mirrors the difference in consistency, ensuring that $\kappa_s - \kappa_i \geq c_s - c_i$ for all cases where $c_s > c_i$. In this way, it is possible that the difference $\kappa_s - \kappa_i$ is greater than $c_s - c_i$ because the loss only enforces a ranking rather than an exact value difference.

Correctness Ranking Loss. To make the most of the labeled data for confidence estimation, a measure of prediction correctness is utilized as a separate ranking loss, $\mathcal{L}_{corr}$, which is applied only to labeled samples. Correctness of a sample represents the frequency of correct predictions for that sample $\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{\hat{y}_i^t = y_i\}$. The total loss, $\mathcal{L}$, combines these elements to optimize the model's performance:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{corr} + \lambda_2 \mathcal{L}_{cons} \tag{3}$$

Here $\mathcal{L}_{CE}$ is the cross entropy loss on labeled samples. λ_1 and λ_2 are the weights of the correctness and consistency losses, respectively.

3.2. Post-Processing and Masking

In the context of breast cancer, regions affected by invasive cancer are often found close to one another [20]. This means the likelihood of a region being cancerous is influenced by the cancer status of neighboring regions. We construct a comprehensive WSI probability map, $\mathcal{H}$, from the patch-wise model predictions. $\mathcal{H}$ is then refined into an aggregated probability map, named $\mathcal{A}$. In this aggregated map, the probability score for a patch is an aggregation of the scores for that patch and neighboring patches within a certain distance range. The aggregation function used is the Max function. This can be formulated as follow [20]:

$$\mathcal{A}(i,j) = Max(\{\mathcal{H}(m,n) | m,n \in [\frac{i}{w}w, (\frac{i}{w}+1)w]\}) \tag{4}$$

$\mathcal{H}(m,n)$ represents the probability score for a specific patch located at (m,n) on the map $\mathcal{H}$. Similarly, $\mathcal{A}(i,j)$ refers to the aggregated probability score of a patch at location (i,j), and the aggregation window of size w is defined as:

$$[[\frac{i}{w}w, ([\frac{i}{w}+1]w)] \times [[\frac{i}{w}]w, (\frac{i}{w}]+1)w] \tag{5}$$

After postprocessing, all patches with probability greater than or equal to 0.5 are considered tumor positive, and are considered tumor negative otherwise.

3.3. Implementation Details

Following the data augmentation and training settings outlined by Le et al. [20], we adopt the ResNet-34 architecture [18] initialized with pre-training on the ImageNet dataset [8] and customized with the last fully connected layer containing 512 neurons tailored for binary classification. The patch size used for training and testing is 350×350 pixels at 40x magnification. The patches are normalized and resized to 224×224 pixels before feeding to the model.

4. Experiments and Results

4.1. Experimental Setup

Our experimental design leverages the ResNet-34 architecture to implement various classification methods, including baseline (Cross-entropy loss), MC-dropout, Correctness, and Consistency methods. For the MC-dropout method, we incorporated an additional dropout layer in the ResNet-34 architecture, situated before the final fully connected layer, with a dropout rate of 0.5 to effectively model uncertainty.

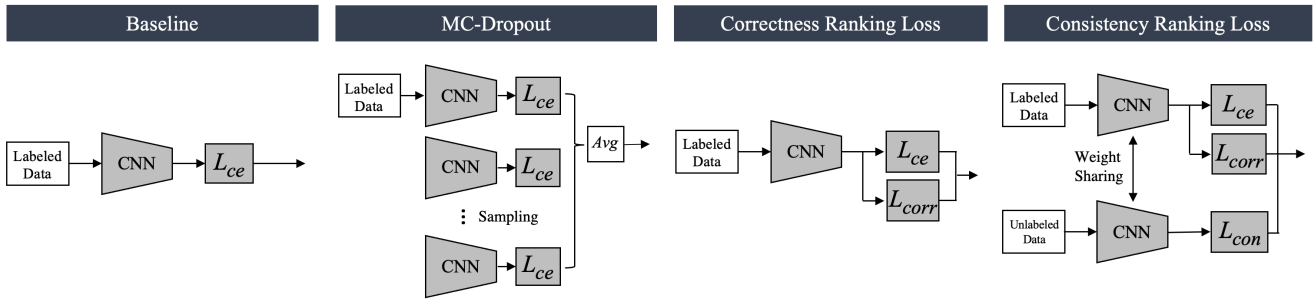

Figure 2. Model architectures for performance evaluation in our application: Baseline (Cross-entropy Loss), MC-Dropout [11], Correctness Ranking Loss [22], and the employed Consistency Ranking Loss [21] structure. L_{ce}, L_{corr}, and L_{con} each represent Cross Entropy Loss, Correctness Ranking Loss [22], and Consistency Ranking Loss [21], respectively.

This model variant underwent 50 stochastic forward passes during inference to robustly estimate the prediction uncertainty. The Correctness model is trained on labeled data and utilized the correctness ranking loss. The Consistency model is trained on both labeled and unlabeled data, and the training process utilized both the correctness and the consistency ranking losses. Fig. 2 shows a comparative visualization of the training process across the different methods.

Each model was trained over 20 epochs using a mini-batch size of 256, a momentum of 0.9, and a weight decay of 0.0001, starting with an initial learning rate of 0.01, which was decreased by a factor of 10 after the 8th and 16th epochs to fine-tune the learning process. We use stochastic gradient descent [2] to optimize each method. For the weighted loss in Eq. 3, we set $\lambda_1 = \lambda_2 = 0.5$

Data augmentation operations applied include: random patch rotation by up to 22.5 degrees, random vertical and horizontal flipping, and random adjustments or perturbations to their brightness, contrast, and saturation levels. During the testing phase, the only preprocessing applied was normalizing the color channels, with zero mean and one standard deviation.

4.2. Datasets

The datasets used for training and validation were compiled from image patches sourced from 102 and 7 breast cancer WSIs, respectively, that are part of the Surveillance, Epidemiology, and End Results (SEER)-Linked Virtual Tissue Repository (VTR) Pilot Breast Cancer Genomics Study [10]. We assess the efficacy and generalizability of our deep learning models with 195 TCGA WSIs, which had been previously manually labeled by Cruz-Roa et al. [7].

In this study, we explore how uncertainty calibration can enable us to train tumor prediction model with a limited set of annotations and yet get higher quality predictions by taking advantage of the vast amount of the unlabeled data available. To achieve this, we conduct experiments using only 10% and 20% random samplings of the training data labels and treat the rest of the dataset as unlabeled. We refer to these training datasets as $SEER_{10}$ and $SEER_{20}$, respectively. The training and test datasets statistics are shown in Table 1. Earlier studies have demonstrated the advantage of incorporating a larger proportion of negative samples compared to positive ones within training with digital pathology datasets [3, 20]. Subsequently, we chose a 1:2 ratio of tumor-positive to tumor-negative patches for training.

Table 1. Dataset Distribution for Training, Validation, and Testing in Breast Cancer Detection for showing the composition of datasets derived from SEER and TCGA sources, detailing the training and validation sets with specified percentages of labeled data.

Source	Purpose	ID	WSIs (N)	Labeled (N)	Cancer-Positive (N)	Cancer-Negative (N)
$SEER_{10}$	Training	D_{tr10}	102	33,000	11,000	22,000
	Validation	D_{val}	7	10,224	4,953	5,271
$SEER_{20}$	Training	D_{tr20}	102	66,000	22,000	44,000
	Validation	D_{val}	7	10,224	4,953	5,271
TCGA	Testing	T_{tcga}	195	-	-	-

4.3. Evaluation Metrics.

In assessing the performance and uncertainty of our models, we utilize a comprehensive suite of metrics:

F1 score. Evaluates the accuracy of the binary prediction, that is after applying a threshold=0.5 to the predicted probability to classify patches as tumor positive or negative.

AURC and E-AURC. The Area Under the Risk-Coverage Curve (AURC) and its normalized version Excess-AURC (E-AURC) [14] compute the risk or the error at different confidence thresholds. They measure how well the true and false predictions are separated by their uncertainty.

FPR-95%. The False Positive Rate (FPR) at 95% True Positive Rate (TPR) measures how often the model incorrectly labels a patch as tumor when TPR=95%.

Table 2. Validation of trained model with SEER_{10} (D_{tr10}) and SEER_{20} (D_{tr20}) datasets, adjusted according to the size of labeled training data. Superior results are accentuated in bold for quick identification. To simplify the data presentation, we adjusted AURC and E-AURC figures by a factor of 10^3, FPR figures by 10^2, and NLL figures by 10. SEER_{10} and SEER_{20} indicate subsets comprising 10% and 20% of the complete dataset, which were specifically annotated for the purpose of training. The evaluation was conducted using a validation dataset (D_{val}).

Model	Training Dataset (% of SEER)	Labeled SEER (N) (Pos / Neg)	Unlabeled SEER (N) (Pos / Neg)	Method	F1↑	AURC↓	E-AURC↓	FPR-95↓	ECE↓	NLL↓	Brier↓
ResNet 34	SEER_{10} 10%	33,000 (11,000/22,000)	300,724 (88,889/211,715)	Baseline	0.846	59.46	51.57	78.70	2.38	3.76	23.61
				MCdropout [11]	0.850	66.21	51.57	78.56	3.09	4.15	24.59
				Correctness [22]	0.838	67.14	50.24	79.24	4.05	4.11	25.19
				Consistency [21]	**0.853**	**55.18**	**41.74**	**76.07**	**1.97**	**3.69**	**22.57**
	SEER_{20} 20%	66,000 (22,000/44,000)	267,724 (77,889/189,715)	Baseline	0.840	66.32	49.79	79.41	5.02	4.14	25.31
				MCdropout [11]	0.843	63.76	47.99	78.14	5.65	4.33	25.37
				Correctness [22]	0.853	60.07	46.49	79.00	4.71	4.03	24.01
				Consistency [21]	**0.854**	**55.84**	**42.56**	**77.35**	**1.39**	**3.62**	**22.46**

Table 3. Performance comparison of patch-wise tumor prediction models on TCGA Whole Slide Images. This table presents the evaluation results of various methods on the 195 TCGA dataset T_{tcga}, measured across a range of metrics that assess both accuracy and uncertainty. Each method's performance is quantified by F1 score, AURC, E-AURC, FPR-95, ECE, NLL, Brier score, and Hausdorff Dist. The models have undergone post-processing to assign probabilistic predictions to each patch, which are then compared against the ground truth annotations to assess their effectiveness in tumor detection. To simplify the data presentation, we adjusted AURC and E-AURC figures by a factor of 10^3, FPR figures by 10^2, and NLL figures by 10.

Model	Training Dataset (% of SEER)	Test WSIs (N) (TCGA)	Method	F1↑	AURC↓	E-AURC↓	FPR-95↓	ECE↓	NLL↓	Brier↓	Hausdorff Dist↓
ResNet 34	SEER_{10} 10%	195	Baseline	0.772	139.03	134.80	5.76	4.70	9.11	58.34	190
			MCdropout [11]	0.784	134.76	130.77	6.22	5.54	14.53	62.66	209
			Correctness [22]	0.783	131.22	127.35	5.74	4.79	11.16	57.56	183
			Consistency [21]	**0.784**	**125.51**	**122.13**	**4.20**	**4.53**	**8.92**	**55.61**	**156**
	SEER_{20} 20%	195	Baseline	0.749	155.13	150.02	8.22	8.21	14.05	74.59	217
			MCdropout [11]	0.773	132.94	129.10	5.50	6.02	14.14	61.30	208
			Correctness [22]	0.760	145.26	140.58	6.03	6.03	12.94	65.27	212
			Consistency [21]	**0.774**	**129.19**	**125.38**	**4.82**	**4.82**	**7.41**	**53.59**	**164**

Brier Score. The Brier Score [4] quantifies the accuracy of probabilistic predictions, penalizing more the predictions that are confident but incorrect, making it a useful tool for calibration assessment.

ECE. The Expected Calibration Error (ECE) [23] groups the predicted probabilities into bins and aggregates the mean difference between the prediction confidence and the prediction accuracy in each bin, offering a summary of the model's calibration.

NLL. The Negative Log Likelihood (NLL) [15] evaluates the model's predicted probabilities against the actual class labels, emphasizing the cost of being confidently wrong.

Hausdorff Distance. The Hausdorff distance [9] is used to quantify how similar the boundaries of the predicted and ground truth tumor masks. It computes the maximum distance between any point in one image and the nearest in the other and vice versa. It evaluates how well the model is performing in terms of accurately delineating the tumor regions.

To clarify, high F1 score is better, while low AURC, E-AURC, FPR-95, Brier, ECE, NLL, and Hausdorff Distance are better. These metrics enable a nuanced evaluation of our model's predictive performance and its capability to handle uncertainty, essential for the reliable detection of tumor in pathology images.

4.4. Results

We evaluate and compare the trained models, both quantitatively and qualitatively. We first present the quantitative evaluation performed on patch-wise predictions from the SEER validation set and on WSIs from the TCGA-BRCA test dataset. We then present the qualitative results from predictions on WSIs.

Validation Patch-wise Evaluation. We evaluate the trained models on patch-wise prediction using validation patches from the SEER dataset. The results in Table 2 show that the consistency ranking loss has the best performance across all metrics. More importantly, it achieves the lowest score on the uncertainty metrics by a large margin indicating the higher quality of the prediction in terms of the model's uncertainty estimation.

Test WSI Evaluation. We test the models trained on SEER data on TCGA WSIs. The test slides were first patch-wise processed to generate probability maps and then postpro-

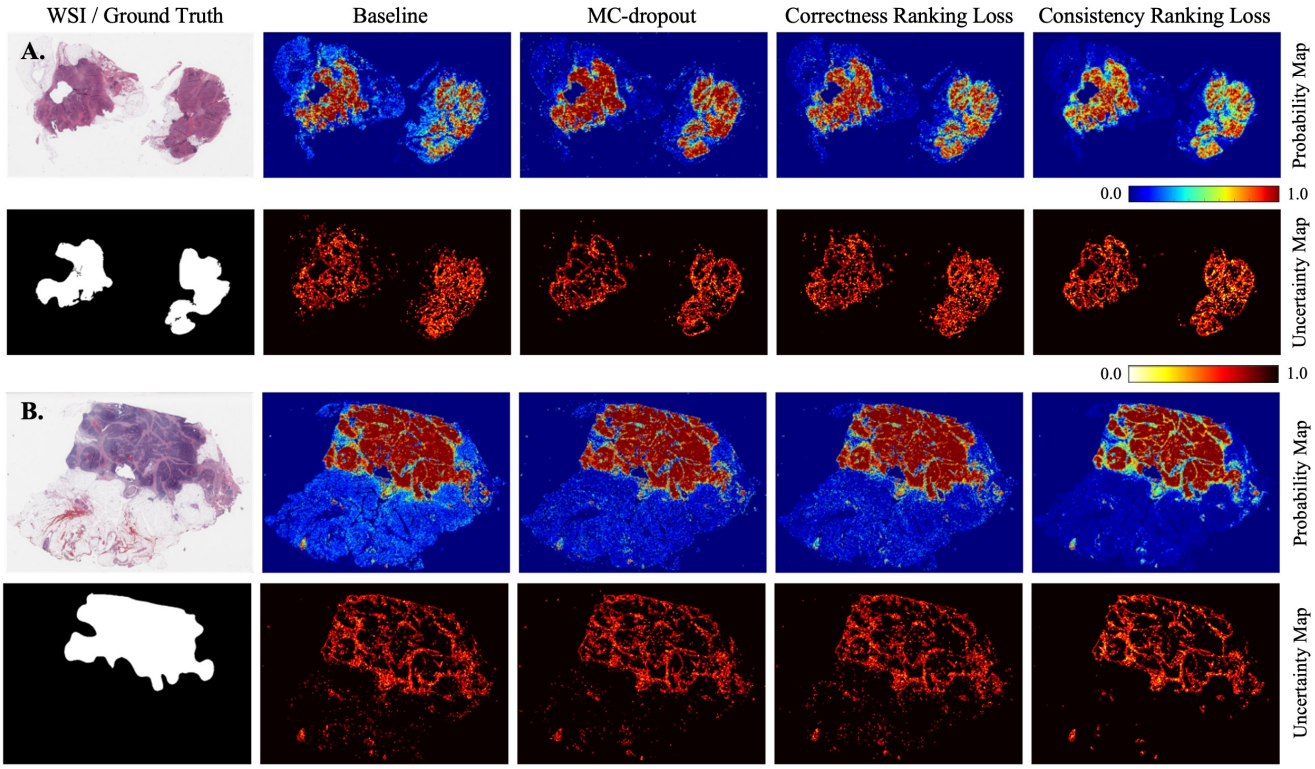

Figure 3. Comparative visualization of tumor prediction in The Cancer Genome Atlas whole slide images (WSIs) using various methodologies based on ResNet-34 trained with D_{tr20}. The top parts of each element represent the Probability Map. The bottom parts are the corresponding Uncertainty Map. Columns show the output of the baseline method, MC-dropout, Correctness Ranking Loss, and Consistency Ranking Loss, respectively, with heatmaps reflecting the probability of tumor presence. Probability maps exhibit the model's confidence in tumor presence, with warmer tones (red) indicating higher likelihood of tumor (probabilities closer to 1), and cooler tones (blue) suggesting lower probabilities (closer to 0). The Uncertainty Map indicates that the color becomes more yellow as the values approach 1, which signifies higher uncertainty. Conversely, the color turns blacker as the values approach 0, indicating higher certainty.

cessed as outlined in Section 3.2. The aggregated probability maps and the generated binary masks are compared against the ground truth binary maps. The results in Table 3 reflect a comprehensive evaluation of the methods' performance on the test slides. Similar to the patch-wise evaluation, the model trained with the consistency ranking loss achieves best performance in all categories, and especially in the uncertainty evaluation metrics. This indicates the higher reliability of the consistency model, which is an essential factor in medical tasks. Moreover, the evaluation of the predicted masks of tumor regions using Hausdorff distance confirms that the consistency model can better handle the ambiguity that often occurs at the tumor boundary.

Qualitative Results. Fig. 3 and Fig. 4 present a qualitative comparison of four different methods for predicting tumor in WSIs. The figures illustrate both the probability maps (odd rows) and the uncertainty maps (even rows) generated by the models. The uncertainty maps are computed using the following mathematical formula:

$$p_u = 2 \times (0.5 - |p_{pr} - 0.5|) \tag{6}$$

where p_{pr} and p_u are the predicted probability and uncertainty for patch p in the WSI, respectively. The closer p_u is to 1, the greater the uncertainty of the prediction, indicating that the model is less confident of the prediction, and conversely, a p_u value closer to 0 signifies higher certainty in the prediction, implying that the model has a more confidence in its assessment of the patch being tumor or not. Fig. 5, shows sample zoomed in regions from the results of the consistency prediction. It illustrates regions with various morphological characteristics and how the consistency model react to these different regions.

A high level inspection of Fig. 3 and Fig. 4, allows us to make the following observations:

- The consistency model has less noise in its probability and uncertainty maps. This is depicted by the better structure visibility in it's maps and the cleaner delineation of the tumor regions.

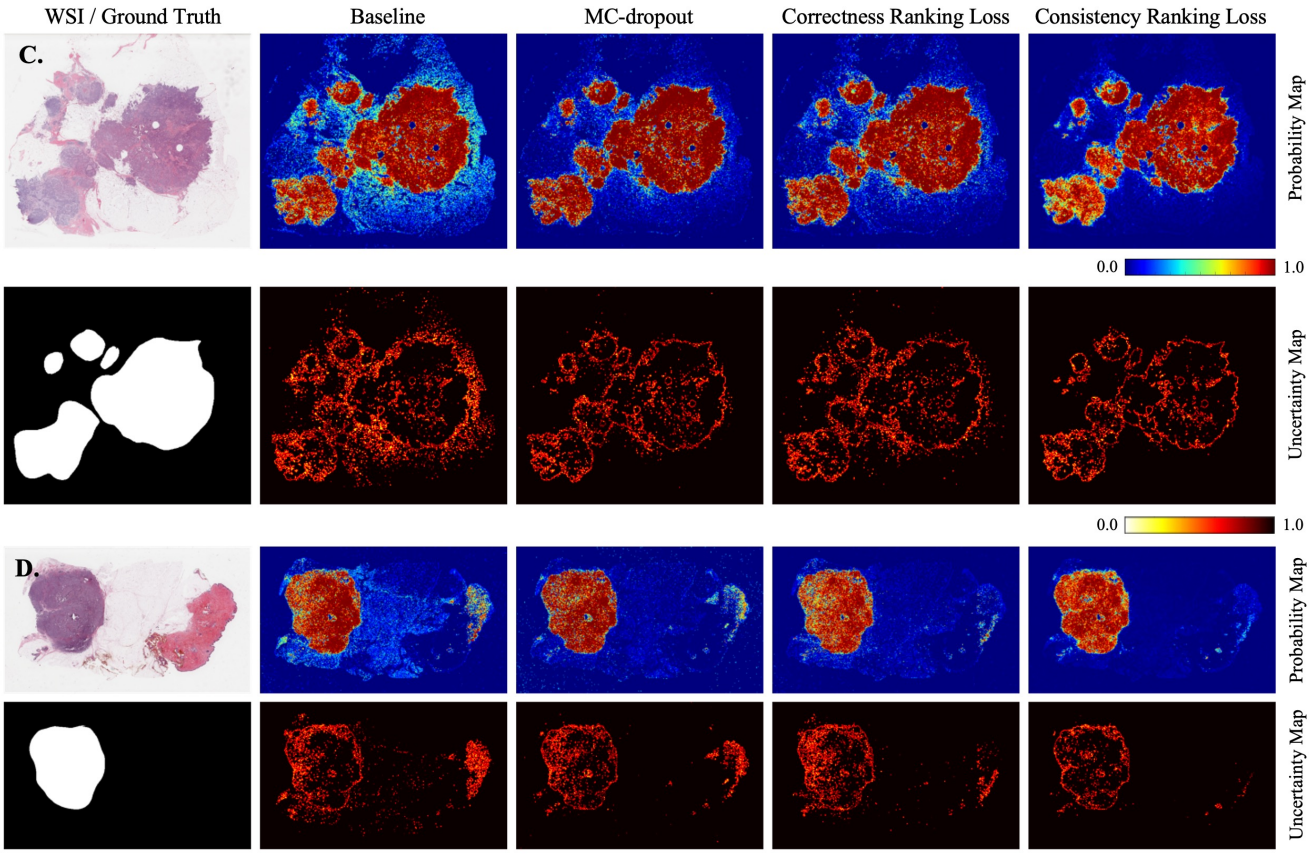

Figure 4. Comparative visualization of tumor prediction in The Cancer Genome Atlas whole slide images (WSIs) using various methodologies based on ResNet-34 trained with D_{tr20}. The top parts of each element represent the Probability Map. The bottom parts are the corresponding Uncertainty Map. Columns show the output of the baseline method, MC-dropout, Correctness Ranking Loss, and Consistency Ranking Loss, respectively, with probability map reflecting the probability of tumor presence. Probability maps exhibit the model's confidence in tumor presence, with warmer tones (red) indicating higher likelihood of tumor (probabilities closer to 1), and cooler tones (blue) suggesting lower probabilities (closer to 0). The Uncertainty Map indicates that the color becomes more yellow as the values approach 1, which signifies higher uncertainty. Conversely, the color turns blacker as the values approach 0, indicating higher certainty.

- The uncertainty maps from the consistency model show higher uncertainty around the tumor boundary and more confidence positive prediction inside the tumor. This agrees with how pathologists perceive the tumor regions during annotation. On the contrary, all the other methods show more uncertainty inside the tumor regions.

We hypothesize that the highly structured uncertainty corresponds to specific morphology in the tumor microenvironment. Taking a closer look at the probability maps in Fig 5, we observe the following characteristics of the consistency method predictions:

- In Fig 5 (A), the uncertain and low probability patches within the tumor region depict tumor infiltrating lymphocytes and stroma.
- In Fig 5 (B), we see two lines of stroma passing through the tumor regions and the probability map clearly and sharply captures them with medium to low probability values.
- In Fig 5 (C), we observe high confidence inside the tumor, and more uncertainty as the tumor meets the tissue edge. Similarly in Fig 5 (D), we observe lower probability at the forefront of the tumor region and more confidence as we step inside the tumor.

The previous observations show that inspecting the probability and uncertainty maps can reveal more the insights about the tumor microenvironment beyond simply finding the tumor regions. This confirms the value of training with uncertainty calibration applied on labeled as well as unlabeled data. It emphasizes the boost in performance that can be achieved when we have limited supervision, which is often the case in medical tasks. Moreover, it proves the reliability and trustworthiness of the models, which are essential in clinical applications and research.

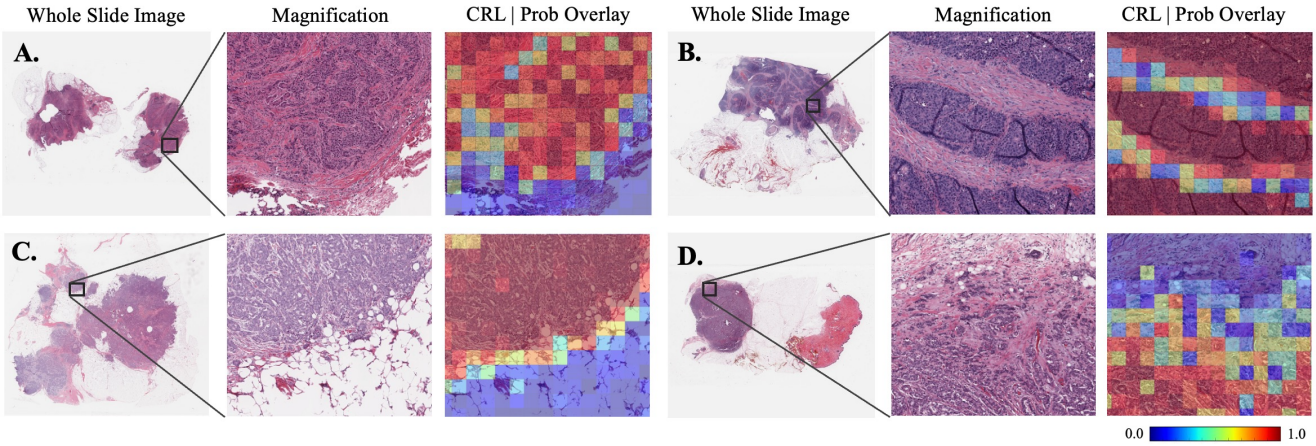

Figure 5. Comparative analysis of tumor prediction using Consistency Ranking Loss (CRL) demonstrating the transition from Whole Slide Images to magnified views and the probability map overlays. These overlays provide visual representation of prediction certainty and uncertainty, with the color spectrum indicating the likelihood of tumor presence—ranging from cooler tones for lower probabilities to warmer tones for higher probabilities.

5. Conclusion

Our research presents a significant leap in patch-wise tumor prediction by integrating consistency ranking loss and correctness ranking loss, allowing us to train with uncertainty calibration on both labeled and unlabeled data. The proposed approach has shown to bolster the accuracy of predictions as well as improves the expression of model's uncertainty in it's prediction probability. By adeptly utilizing both labeled and unlabeled data, our approach directly addresses the pivotal challenge of sparse data availability in the field. The resulting method not only deliver more accurate tumor detection but also enhance their trustworthiness, an issue of great importance to pathologists. The proposed approach can be extended to other medical tasks for more reliable models with fewer annotation cost.

6. Acknowledgements

We would like to acknowledge that the SEER WSIs used in this analysis were obtained through the support and conducted by the Surveillance Research Program of the Division of Cancer Control and Population Sciences at the National Cancer Institute of the National Institutes of Health. This research was supported by the National Science Foundation (NSF) grants IIS-2123920 and CCF-2144901, the National Institute of General Medical Sciences (NIGMS) grant R01GM148970, and the National Cancer Institute (NCI) grant 5UH3CA225021.

References

[1] Ajay Basavanhally, Shridar Ganesan, Michael Feldman, Natalie Shih, Carolyn Mies, John Tomaszewski, and Anant Madabhushi. Multi-field-of-view framework for distinguishing tumor grade in er+ breast cancer from entire histopathology slides. *IEEE Transactions on Biomedical Engineering*, 60(8):2089–2099, 2013. 2

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013. 4

[3] Yeman Brhane Hagos, Albert Gubern Mérida, and Jonas Teuwen. Improving breast cancer detection using symmetry information with deep learning. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 90–97, Cham, 2018. Springer International Publishing. 4

[4] GLENN W. BRIER. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950. 5

[5] Melvin Chua, Donghwan Kim, Jeonghwan Choi, et al. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, 7:711–718, 2023. 1

[6] Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie N.C. Shih, John Tomaszewski, Anant Madabhushi, and Fabio González. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific Reports*, 7:46450, 2017. 1, 2

[7] Angel Cruz-Roa, Hannah Gilmore, Anant Madabhushi, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie Shih, John E. Tomaszewski, and Fabio A. González. High-throughput adaptive sampling for whole-slide histopathology image analysis (hashi) via convolutional neural networks: Application to invasive breast cancer detection. *PLoS ONE*, 13(5):e0196828, 2018. 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[9] M.-P. Dubuisson and A.K. Jain. A modified hausdorff distance for object matching. In *Proceedings of 12th International Conference on Pattern Recognition*, pages 566–568 vol.1, 1994. 5

[10] Máire A Duggan, William F Anderson, Sean Altekruse, Lynne Penberthy, and Mark E Sherman. The surveillance, epidemiology, and end results (seer) program and pathology: toward strengthening the critical relationship. *The American journal of surgical pathology*, 40(12):e94–e102, 2016. 4

[11] 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. 1, 2, 4, 5

[12] J. Gawlikowski, C.R.N. Tassi, M. Ali, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023. 1

[13] Yonatan Geifman and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. *ICLR*, 2019. 2

[14] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019. 4

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. *MIT Press*, 2016. 5

[16] 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-Volume 70*, pages 1321–1330. JMLR. org, 2017. 2

[17] Zobia Hameed, Begonya Garcia-Zapirain, J. J. Aguirre, et al. Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network. *Sci Rep*, 12:15600, 2022. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[20] Han Le, Rajarsi Gupta, Le Hou, Shahira Abousamra, Danielle Fassler, Luke Torre-Healy, Richard A. Moffitt, Tahsin Kurc, Dimitris Samaras, Rebecca Batiste, Tianhao Zhao, Arvind Rao, Alison L. Van Dyke, Ashish Sharma, Erich Bremer, Jonas S. Almeida, and Joel Saltz. Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer. *The American Journal of Pathology*, 190(7):1491–1504, 2020. 1, 2, 3, 4

[21] Chen Li, Xiaoling Hu, and Chao Chen. Confidence estimation using unlabeled data, 2023. 1, 2, 3, 4, 5

[22] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning*, 2020. 1, 2, 3, 4, 5

[23] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015. 5

[24] John C. Platt. Probabilities for sv machines. In *Advances in Large Margin Classifiers*, 2000. 2

[25] Fabio Alexandre Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 2560–2567, 2016. 2

[26] Mitko Veta, Paul J. van Diest, Robert Kornegoor, Andre Huisman, Max A. Viergever, and Josien P. W. Pluim. Automatic nuclei segmentation in h&e stained breast cancer histopathology images. *PLOS ONE*, 8:e70221, 2013. 2