GROUP 2

Quantifying Uncertainty in a Tumor Segmentation Model

Claire Chu, Sara Colando, Dhruba Nandi, Xavier Serrano

Transparency as Explainability

Transparency **exposes** a model's properties to various stakeholders to **better understand**, **improve**, and **contest** model predictions.

Uncertainty Quantification in models **communicates** to stakeholders:

(a) if and when they should **trust** model predictions
(b) assess how **fair** these predictions are on sample-wide and patient-specific cases

So, **Uncertainty is Transparency** and **Uncertainty is Explainability**

How Does Uncertainty Enhance Explainability?

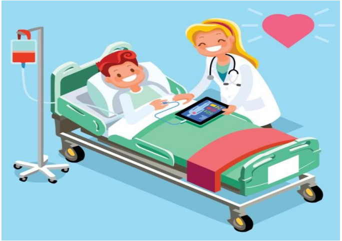
Explainable to Clinicians:

- Allowing physicians to **more confidently** segment tumors
- **Clarity in review processes** leading up to implementation of models in a clinical setting

Explainable to Model Designers:

- Help model designers **understand weaknesses**
- Collaboration with domain experts can clarify various types of errors and their implications

Explainable to Patients:



- Encourage **trust** between clinician and patient
- Help patients understand strengths and limitations of models without an overload of technical information

Central Goal:

Quantify model uncertainty by using a **partially bayesian neural network** (pBNN) to communicate where the model is uncertain of its prediction.

Research Questions:

1. Where is this model failing, and **how is it failing** to properly segment the tumor?
2. In what cases is the model **certain but still makes a mistake** in tumor segmentation?

Outline of Methods

Deterministic **U-NET** Model

MRI Scans

Ground Truth Labels

Determine **Most Sensitive Layer** of U-NET

Train **Partially Bayesian** Neural Network

Generate 100 Predictions for Each Test Patient

Analyze Discrepancy Between Prediction and Ground Truth

Compare Uncertainty for Different Discrepancy Values

Prediction as Thresholded Mean

Uncertainty as Standard Deviation of Predictions

Discrepancy as (Prediction - Truth)

Outline of Methods

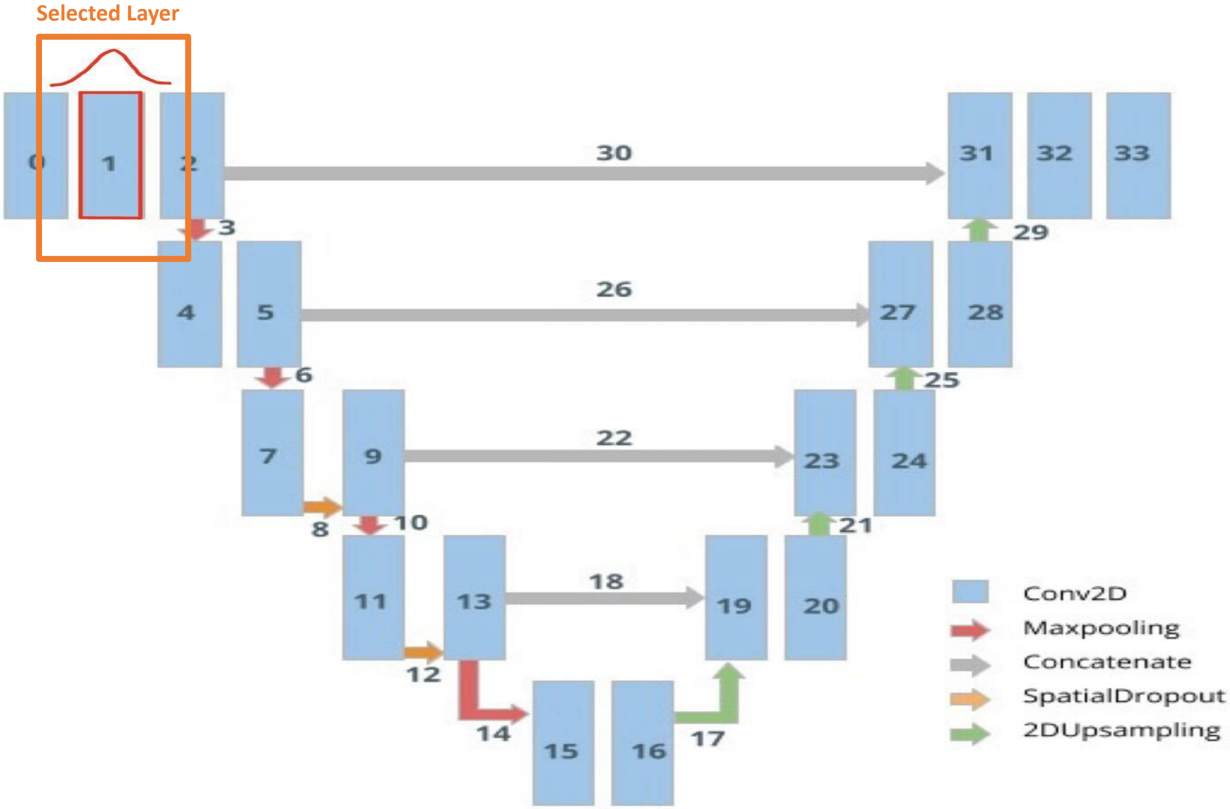
Deterministic **U-NET** Model

MRI Scans

Ground Truth Labels

Determine **Most Sensitive Layer** of U-NET

U-NET Architecture



(Snehal Prabhudesai 2022)

Outline of Methods

Deterministic **U-NET** Model

MRI Scans

Ground Truth Labels

Determine **Most Sensitive Layer** of U-NET

Train **Partially Bayesian** Neural Network

Bayesian Inference

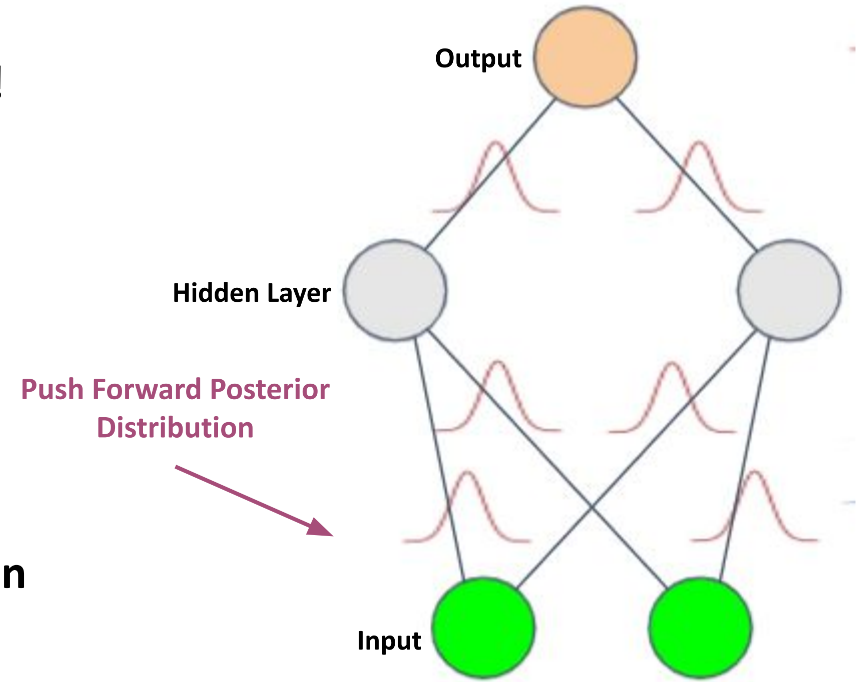
Allows us to **update** the probability of a hypothesis as more data becomes available!

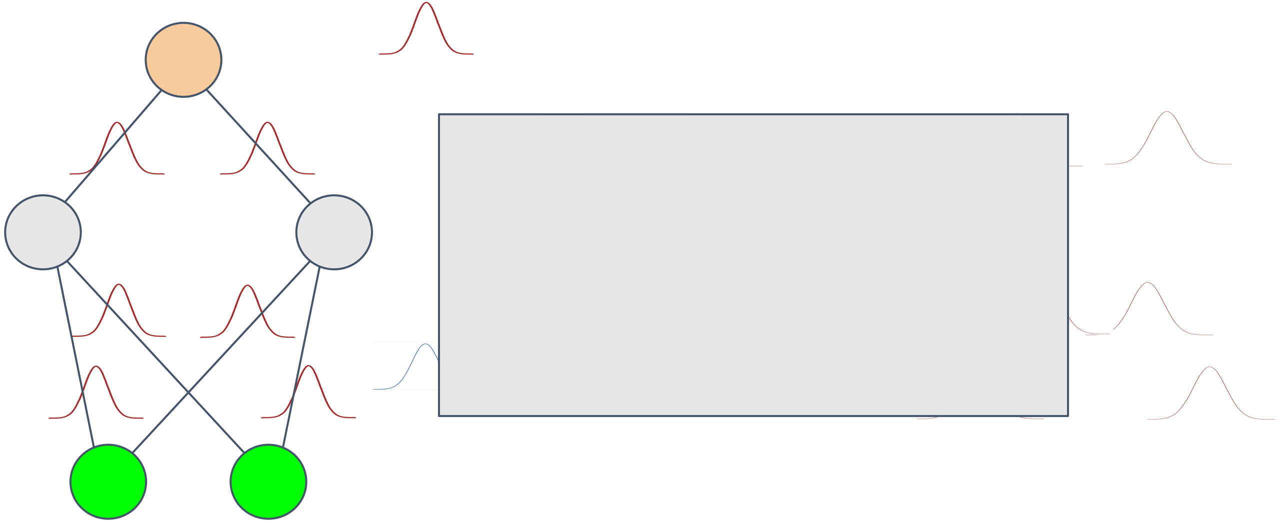
$$p(w|x_T, y_T) = \frac{p(y_T|x_T, w)p(w)}{p(y_T|x_T)}$$

In neural net:

Using bayesian inference, the weights are **sampled** push-forward **posterior distribution** generated during training.

Example: Full Bayesian Neural Net





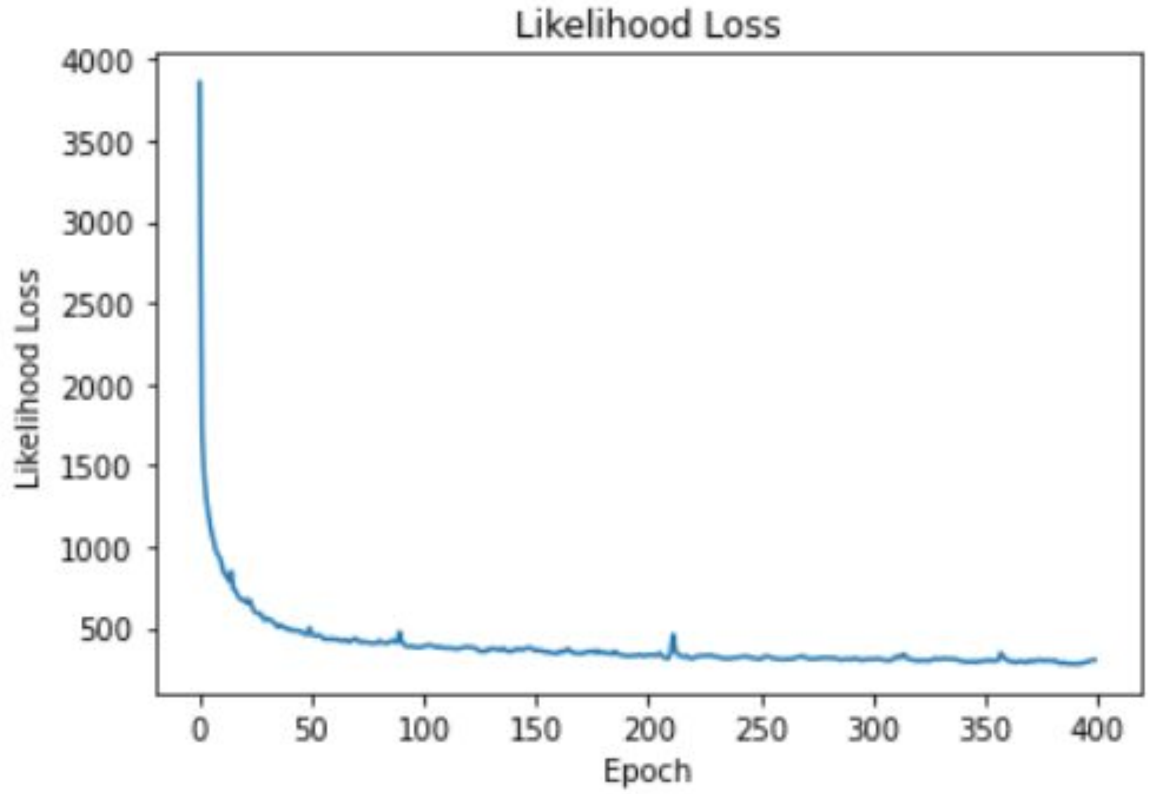
Why Use a Partially Bayesian Neural Net?

Targeted Bayesian inference on a small, strategically chosen **single layer** of the Deep Neural Network while training the rest of the network using less-expensive deterministic methods.

Promises of using a pBNN:

- ❏ **Less Computationally Expensive** than using a complete bayesian neural networks.
- ❏ Outputs a predicted value for each pixel between 0 (no tumor) and 1 (tumor) that **serves as a probability** for pixel classification**.**
- ❏ Standard Deviation of sampled predictions **can quantify model uncertainty** → **which increases explainability.**

Tuning the Hyperparameters



Training Summary:

Epochs = 400
Batch Size/Epoch: 256
Parameters: 7.8 million
Training Time: 11 hours

Outline of Methods

Deterministic **U-NET** Model

MRI Scans

Ground Truth Labels

Determine **Most Sensitive Layer** of U-NET

Train **Partially Bayesian** Neural Network

Generate 100 Predictions for Each Test Patient

Analyze Discrepancy Between Prediction and Ground Truth

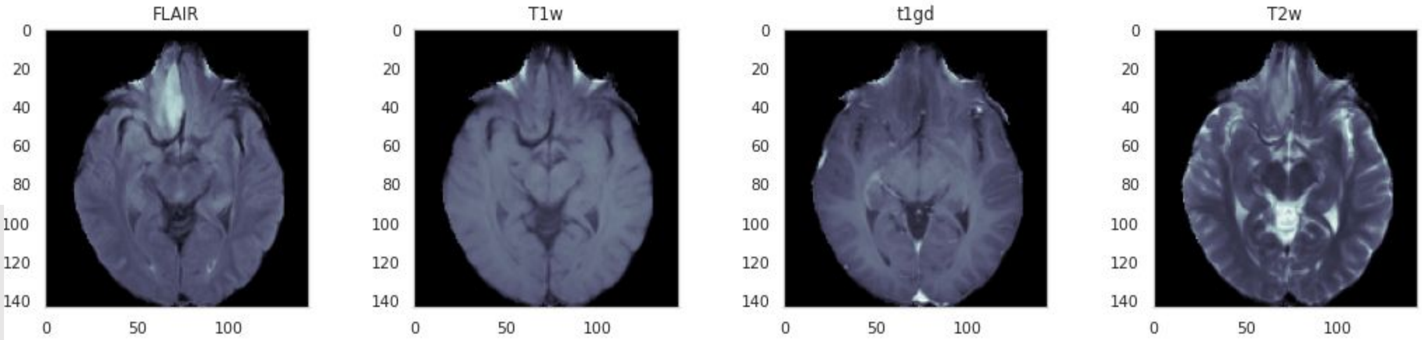
Compare Uncertainty for Different Discrepancy Values

Prediction as Thresholded Mean

Uncertainty as Standard Deviation of Predictions

Discrepancy as (Prediction - Truth)

INPUTS



Female, age 41
37.13 month survival time

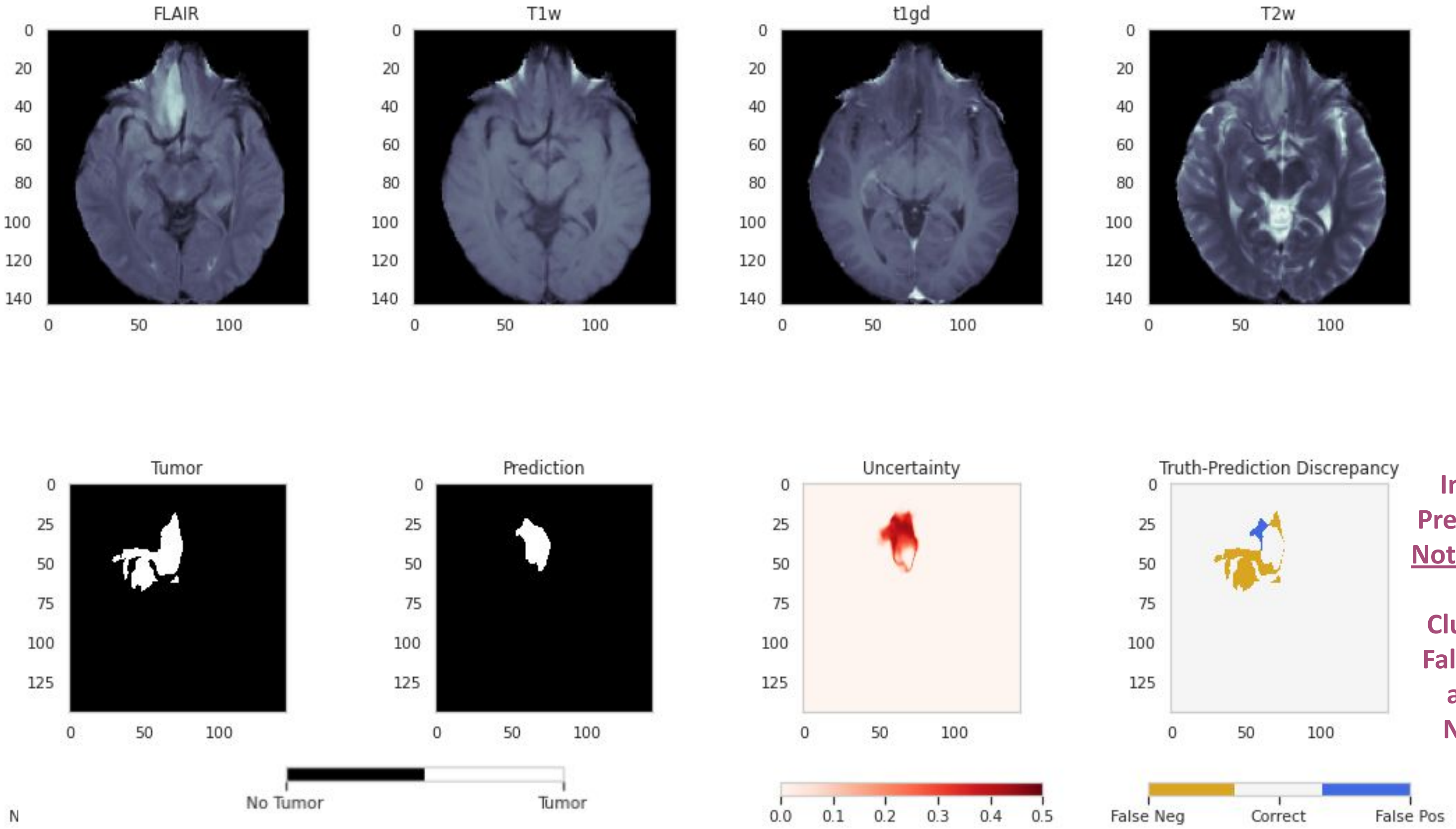
Tissue Source Site: Case Western - St. Joes
Study: Brain Lower Grade Glioma
Histology: oligodendroglioma (G3)

INPUTS

Female, age 41
37.13 month survival time

Tissue Source Site: Case Western - St. Joes
Study: Brain Lower Grade Glioma
Histology: oligodendroglioma (G3)

OUTPUTS

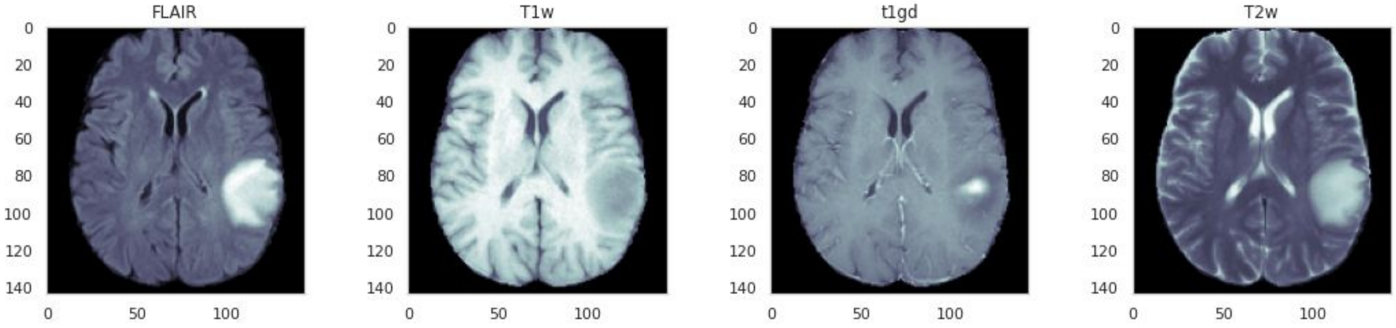


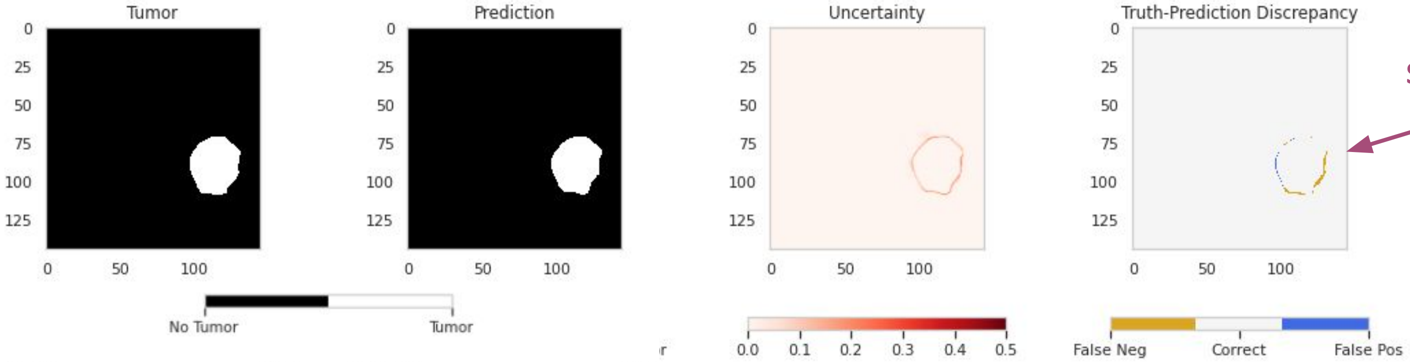
Inaccurate Prediction but Not Uncertain?

Clustering of False Positive and False Negative?

Female, age 66, 15.97 month survival time

Tissue Source Site: Duke
Study: Glioblastoma multiforme
Histology: glioblastoma (G3)





High Sensitivity

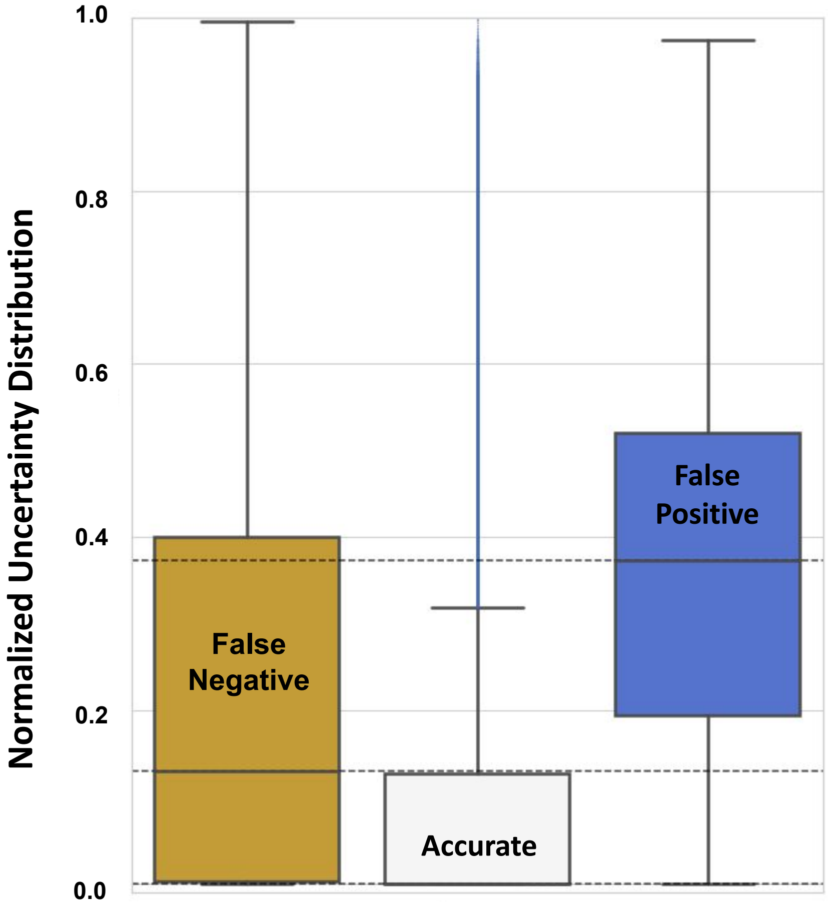
Higher Uncertainty in Predicted Boundary Regions

Comparing Uncertainty Across Truth Prediction Discrepancy Values

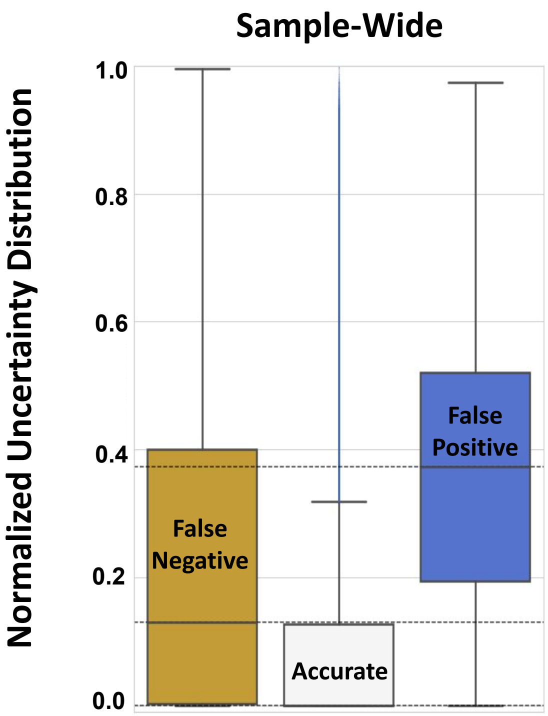
More certain for **accurate** classification.

More certain for **false negatives** than false positives.

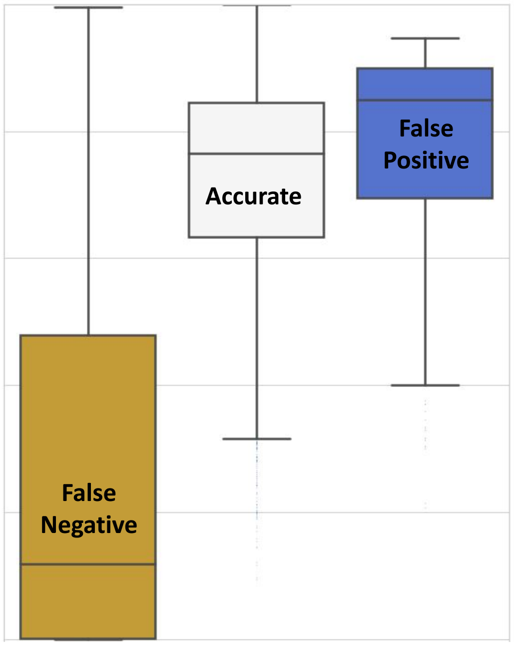
- Less certain when classifying a pixel as "tumor".
- More likely to be falsely confident that a pixel is "non-tumor" than "tumor".



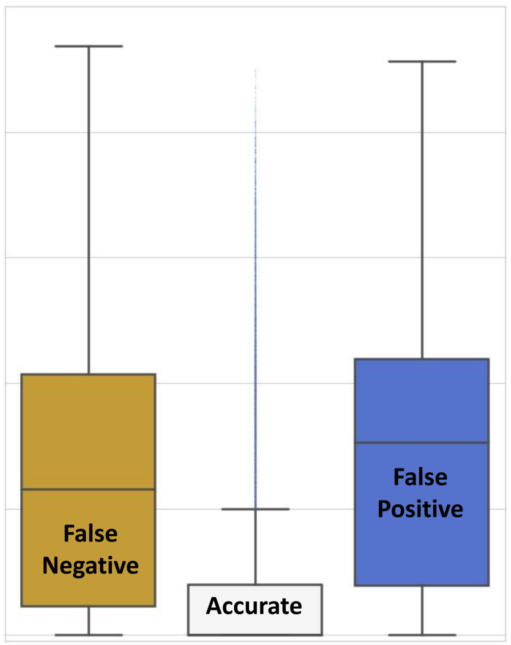
Sample-wide Certainty ≠ Individual Level Certainty



Male, age 67, 7.69 month survival time
Tissue Source Site: Thomas Jefferson University
Study: Lower Brain Grade Glioma
Histology: Astrocytoma (G3)



Female, age 70, 5.32 month survival time
Tissue Source Site: Case Western St. Joes
Study: Lower Brain Grade Glioma
Histology: Astrocytoma (G3)



These patients' clinical info are highly similar

	Case	Tissue source site	Study	Histology	Grade	Age (years at diagnosis)	Gender	Survival (months)	Vital status (1=dead)
1	TCGA-CS-4941	Thomas Jefferson University	Brain Lower Grade Glioma	astrocytoma	G3	67.0	male	7.688047	1.0
318	TCGA-HT-A5RC	Case Western - St Joes	Brain Lower Grade Glioma	astrocytoma	G3	70.0	female	5.322494	1.0

…But the Normalized Uncertainty Distributions Vary

Especially in False Positive and Accurate Discrepancies

Future Work

Collaborating with clinicians to better understand **why model fails in specific brain regions**, and why false positive and false negative results tend to cluster.

Comparing model performance and uncertainty levels **across various subsets** (e.g. different tumor histologies, tissue source sites, patient sex, vital status, etc.).

Investigating the **implications** of the different kinds of **model failure on clinical outcomes**. Investigating what kind of model failure is considered more dangerous by clinicians.

References

Bhatt, Umang, Javier Antorn, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, et al. 2021. "Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–13. AIES '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3461702.3462571.

Prabhudesai, Snehal, Nicholas Wang, Vinayak Ahluwalia, Xun Huan, Jayapalli Bapuraj, Nikola Banovic, and Arvind Rao. 2021. "Stratification by Tumor Grade Groups in a Holistic Evaluation of Machine Learning for Brain Tumor Segmentation." *Frontiers in Neuroscience* 15 (October). https://doi.org/10.3389/fnins.2021.740353.

Snehal Prabhudesai, Dingkun Guo, Jeremiah Hauth. 2022. "Partially Bayesian Neural Networks: Low-Cost Bayesian Uncertainty Quantification for Deep Learning in Medical Image Segmentation."

Thank you!

Claire Chu	cychu@email.unc.edu
Sara Colando	skca2020@mymail.pomona.edu
Dhruba Nandi	nandidhruba2019@gmail.com
Xavier Serrano	serranox17@gmail.com