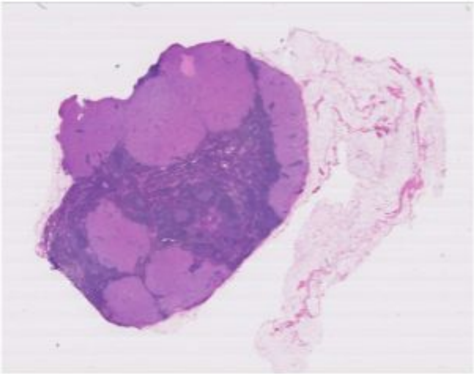
Detecting Cancer Metastases on Gigapixel Pathology Images

Ben Kepecs

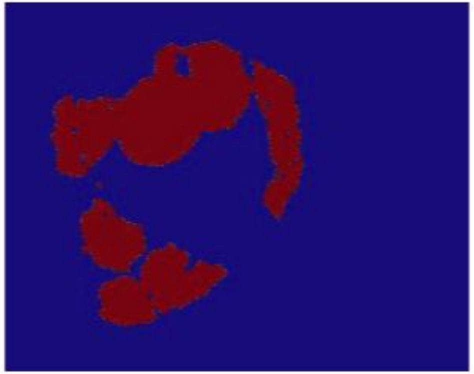
Reference: https://arxiv.org/pdf/1703.02442.pdf

Defining the Problem

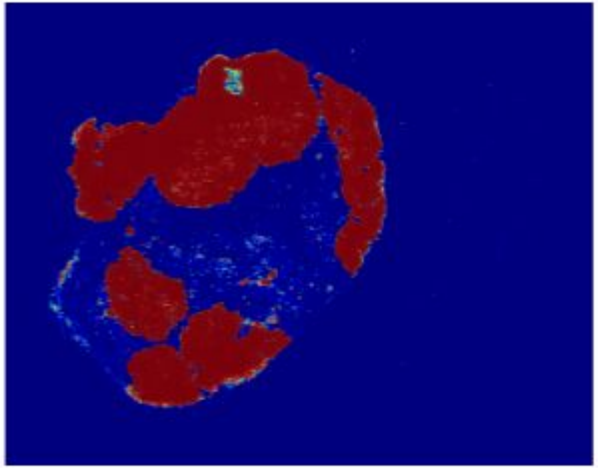
Given labeled pathology slides, can we segment a new slide between cancerous and benign regions?



Biopsy image



Ground truth
(from pathologist)

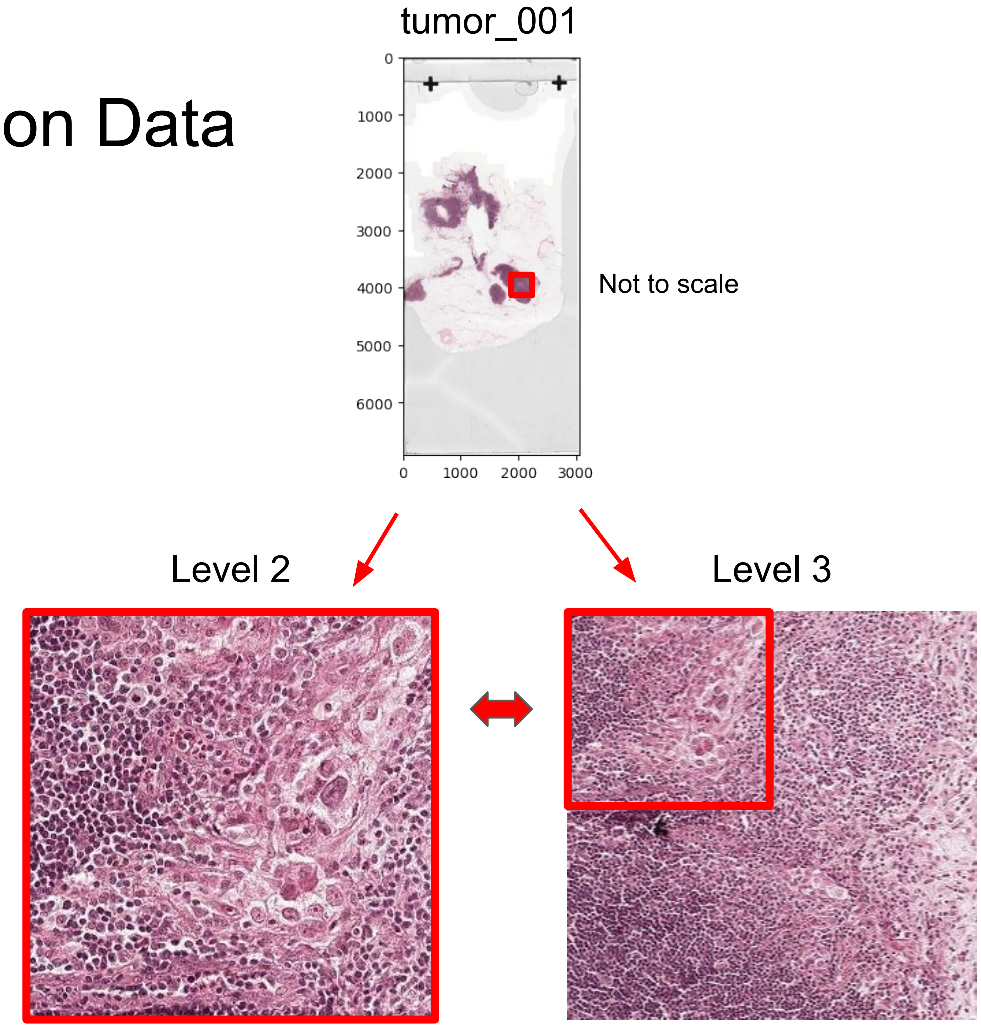


Model predictions

Collecting Training/Validation Data

Protocol:

- Downloaded 11 slides with ground-truth masks
- Allocated 10 slides for training, 1 slide for validation
- Extracted 299x299 patches with percent tissue > 50% from each slide at **level 2** and **level 3**
- Annotated each patch as **cancer** if any pixel in the 299x299 region of the corresponding ground-truth mask was labeled as cancer; otherwise, labeled it as **benign**



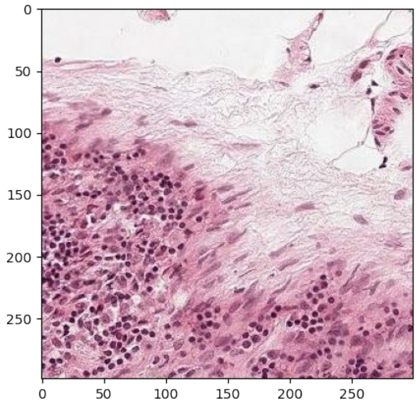
Balancing the Data

- The number of **benign** patches collected outnumbered **cancer** patches by ~10:1
- To balance the dataset, a random subset of the benign patches was removed (levels 2 and 3 were removed as a pair)
- After balancing, the final number of patches used for training was **394** benign and **394** cancer patches at level 2, and the corresponding **394** benign and **394** cancer patches at level 3
- The final number of patches used for validation was **69** benign and **69** cancer at level 2, and the corresponding patches at level 3
- In total, **~1600** patches were used for training and **~280** patches were used for validation (counting both magnifications)

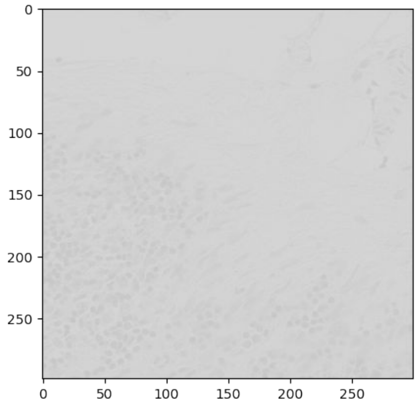
Augmenting the Training Data

- To reduce overfitting, several forms of data augmentation were applied to the training images:
    - Saturation was adjusted randomly to 0-0.25x of original
    - Hue was adjusted randomly with a maximum delta of 0.04
    - Contrast was adjusted randomly to 0-0.25x of original
    - Images were randomly flipped horizontally and rotated by a factor of 0.2

Original Image



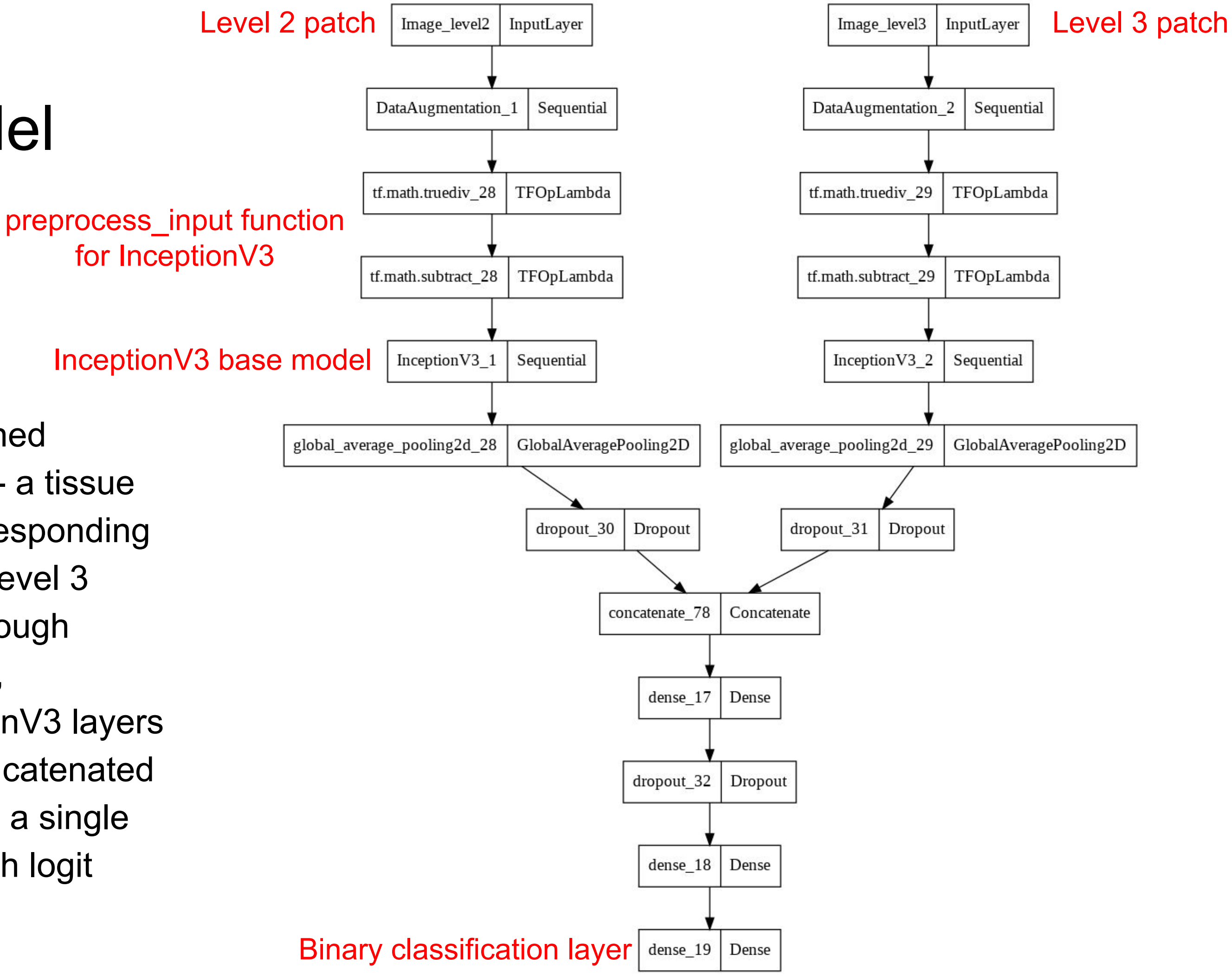
Augmented Image



Classifier will be less color-sensitive

Defining the Model

- A multiscale model was defined
- The model takes two inputs - a tissue patch at level 2 and the corresponding zoomed-out tissue patch at level 3
- The two images proceed through separate data augmentation, pre-processing, and InceptionV3 layers before being pooled and concatenated
- Several dense layers lead to a single binary classification layer with logit output

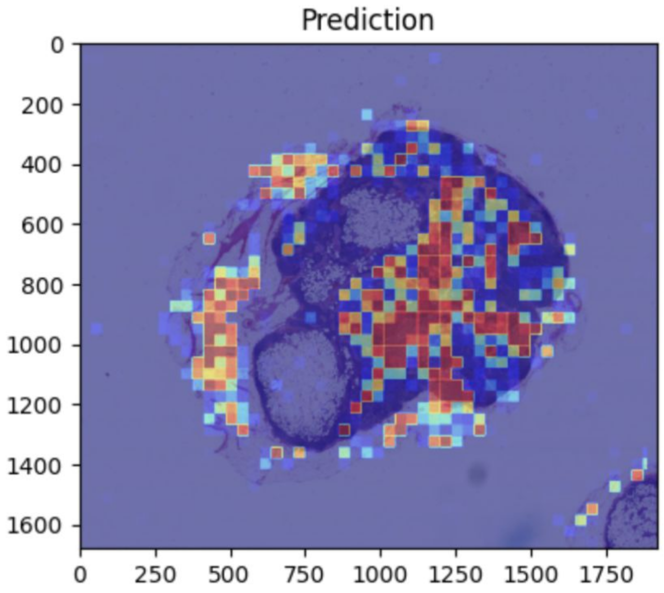


Training the Model

- The model was trained using transfer learning
- The InceptionV3 base models were downloaded with pre-trained ImageNet weights and without their final classification layers (include_top=False)
- The base model layers were frozen and the model was trained for 10 epochs
- All layers from layer 150 and on in the base models were unfrozen and the model was trained for an additional 10 epochs at a tenth of the original learning rate
- The final validation accuracy was evaluated to be **75%** (i.e, the percent of tissue patches that were correctly labeled)
- This seems low…but in production this model was much more accurate!

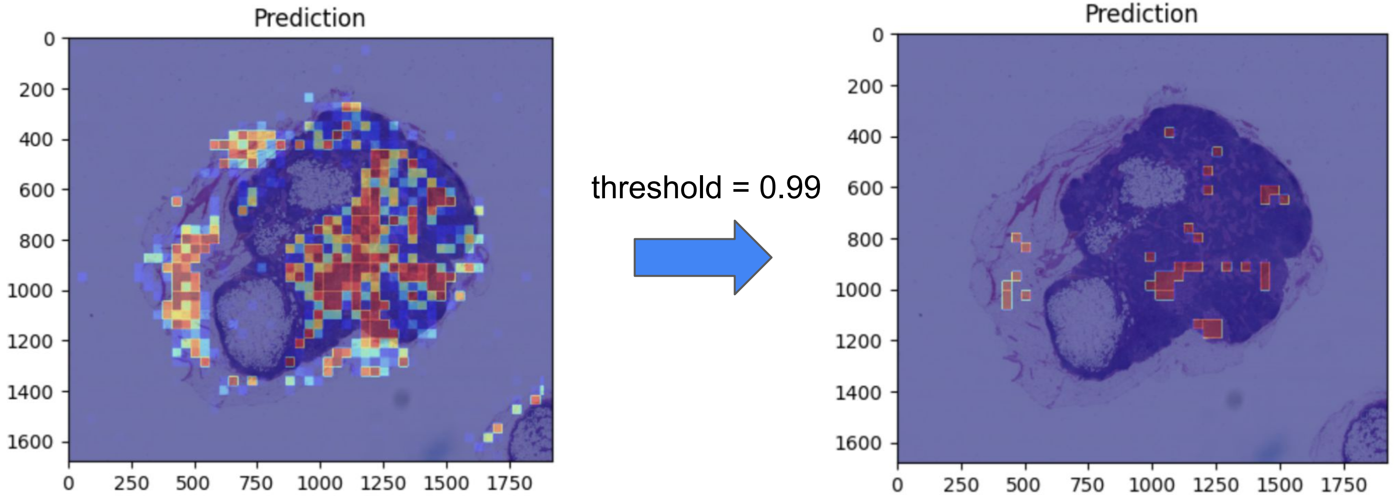
Testing the Model

- 3 unseen slides were downloaded for testing
- For each slide, a heatmap is defined with dimensions equal to the dimensions of the slide at level 5
- An inference script loops through each 299x299 tissue patch in the slides at levels 2 and 3
- model.predict() is called on each (level 2, level 3) pair of images, and the single output is converted to a [0,1] scale using the sigmoid function
- The downsampled region of the level 5 heatmap corresponding to the 299x299 patch of the level 2 image is set equal to the scaled result
- The final heatmap is returned as the prediction mask for the slide



Reducing Noise in the Results

- To eliminate noise in the prediction, a threshold was tuned for each testing image
- All pixels in the heatmap with value < threshold were zeroed out, and all pixels in the heatmap with value >= threshold were set to 1



Evaluation Metrics

- Three metrics for each testing slide were computed to evaluate the performance of the model
- **False positive rate** - the percent of tissue pixels which were predicted as cancer but for which ground truth was benign
- **False negative rate** - the percent of tissue pixels which were predicted as benign but for which ground truth was cancer
- **Overall accuracy** - the percent of tissue pixels where predicted labels matched ground truth labels

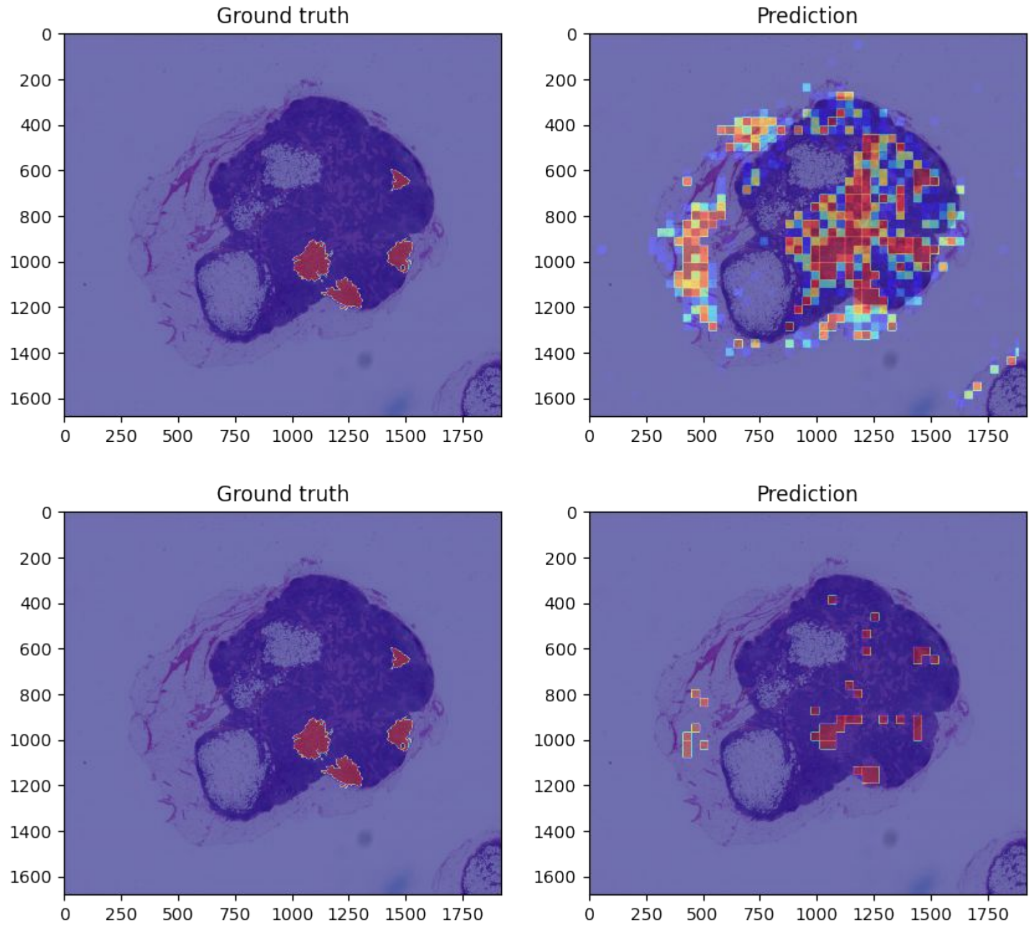
Results on tumor_091

False positive rate = 3.06%

False negative rate = 2.98%

Overall accuracy = 93.96%

Threshold 0.99



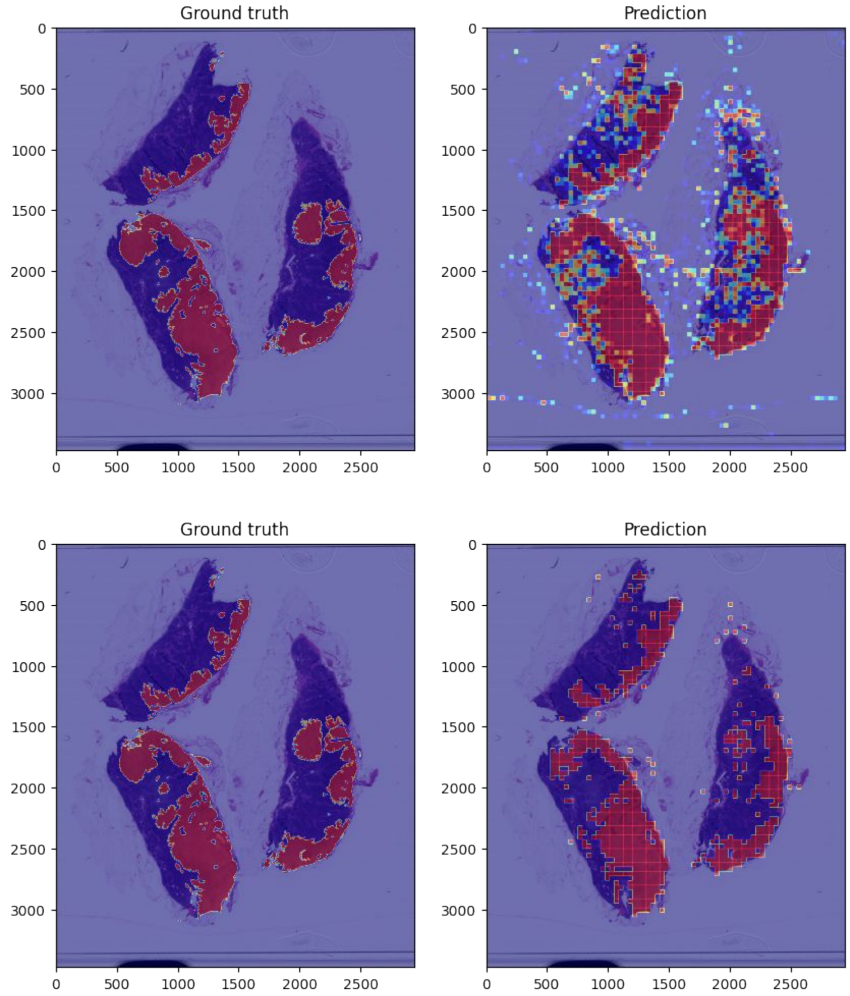
Results on tumor_078

False positive rate = 7.34%

False negative rate = 6.21%

Overall accuracy = 86.46%

Threshold 0.93



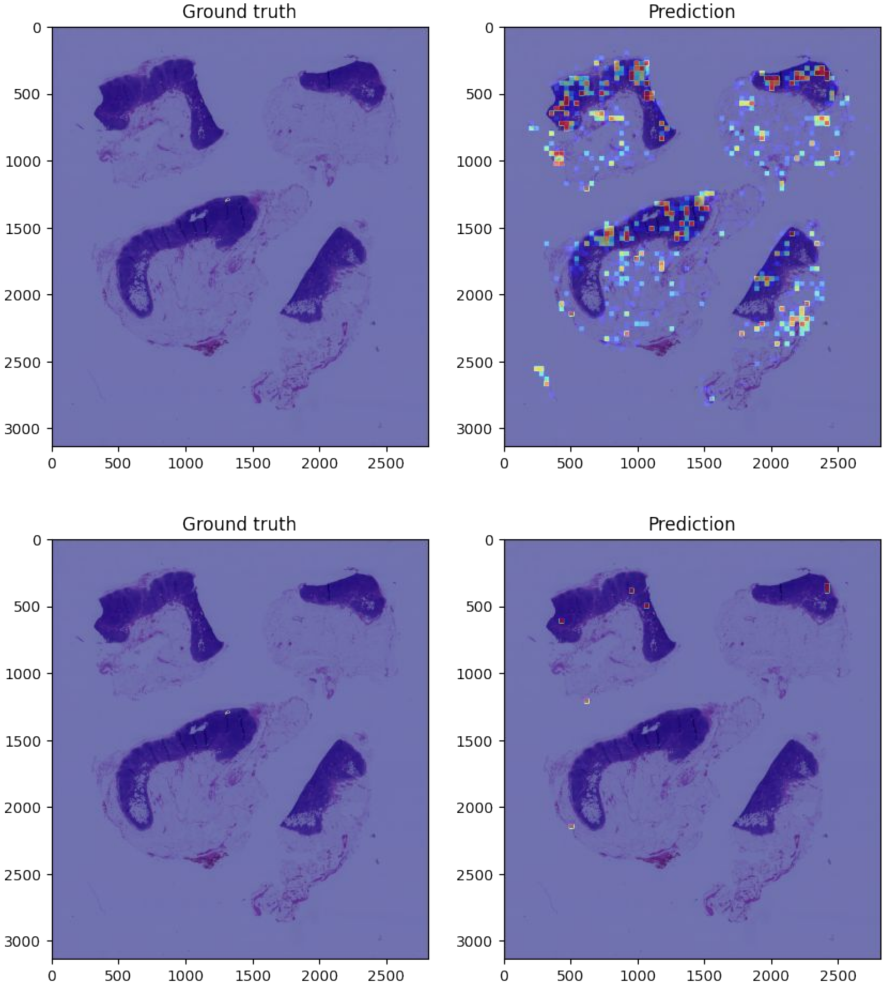
Results on tumor_081

False positive rate = 0.54%

False negative rate = 0.03%

Overall accuracy = 99.42%

Threshold 0.99



Conclusion and Future Directions

- The model achieves satisfactory results on slide images
- Although the model cannot currently replace pathologists, it can assist them
- In future experiments, I will:
  - Experiment with different types of data augmentation. The data augmentation from the paper reduced performance relative to the data augmentation I used, and this will be investigated.
  - Experiment with different model architectures and training methods
  - Use smaller patch sizes (e.g 128x128 patches within the 299x299 patches) to make the final heatmaps more accurate
  - Use higher magnifications and additional levels