Neural Network Classification of Colon Cancer Histopathological Images

Presentation Outline

1. Problem
2. Data
3. Modeling
4. Recommendation
5. Next Steps

Problem

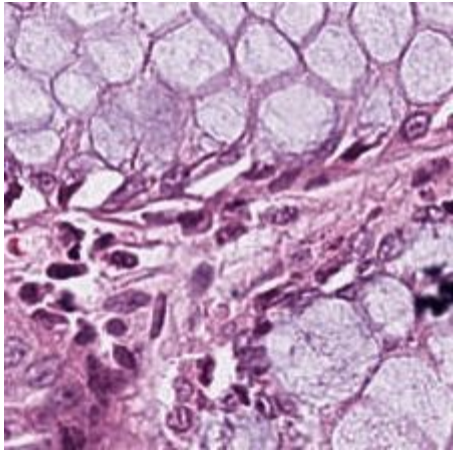
- Colon cancer is the #3 most common cancer in US[1]
- 19 million colonoscopies are performed each year in the US[2]
- It can take 2-3 days to get biopsy results[3]
- Although tests aren't 100% accurate all the time, false positives and false negative results are very low. Generally is thought to occur in 1 to 2% of surgical pathology cases[4]

https://www.cancer.org/treatment/understanding-your-diagnosis/tests/testing-biopsy-and-cytology-specimens-for-cancer/how-long-does-testing-take.html

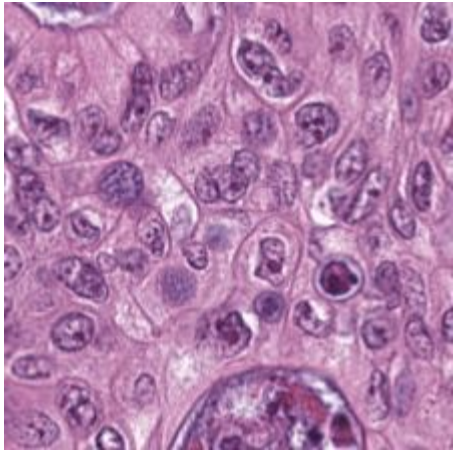
Data

- Zenodo dataset
- Whole dataset: 100,000 images
    - 9 classes separated by a subdirectory
- TUM (cancer cells): 14,317 images
- NORM (normal cells): 8,763 images
- Image size: Standard 224x224
- Cells were preprocessed (staining)

Normal



Cancer

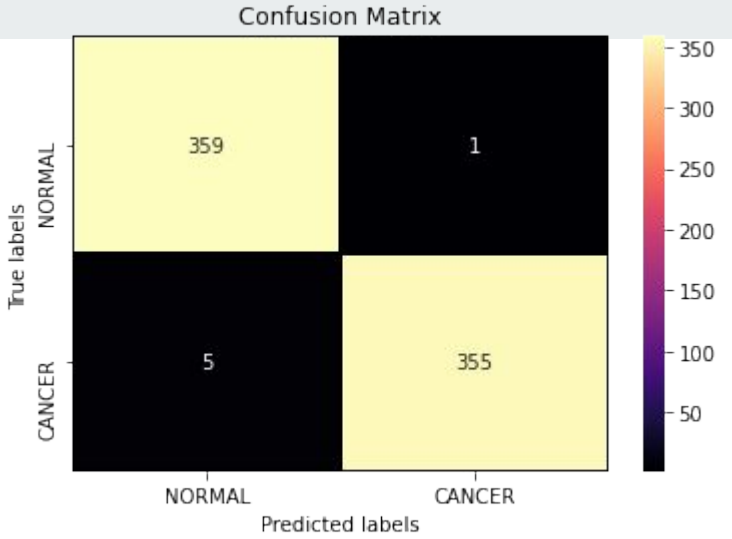


Pre-processing and Modeling

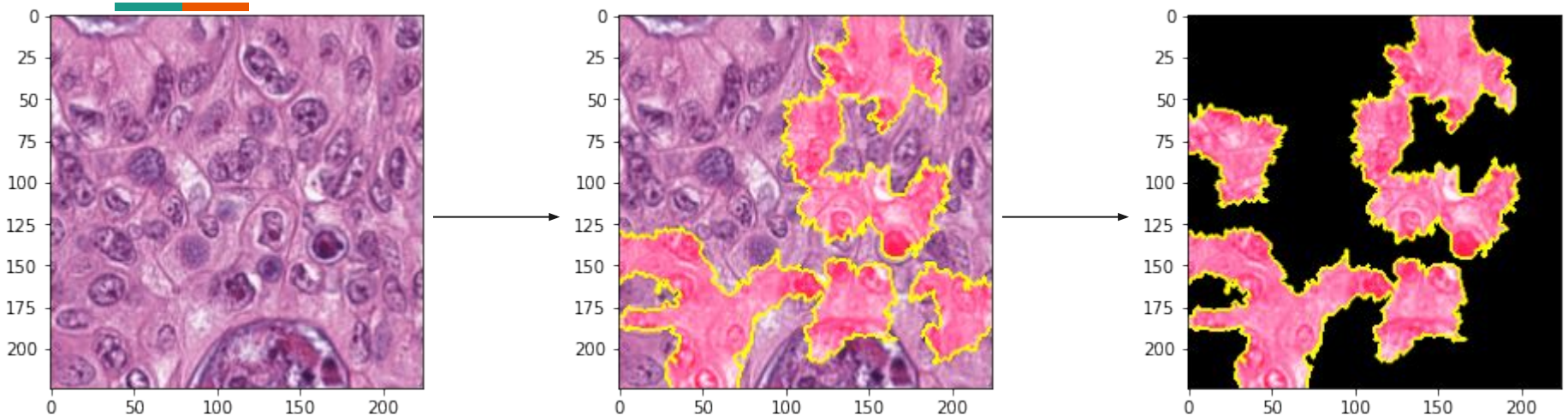
- To prevent our images from losing data, the image size remained the same at 224x224
- Before running the model, images were separated into a training, validation, and testing subdirectories.
- The dataset was undersampled to prevent a class imbalance.
    - 8000 images for each class for training
    - 400 images for each class for validating
    - 360 images for each class for testing
- Images were converted to JPEG from TIF
- Training: images were augmented by adding random rotations, random brightness, random zoom range, horizontal flip, and rescaling
- Validation: images were only rescaled
- Testing: images were only rescaled

Modeling

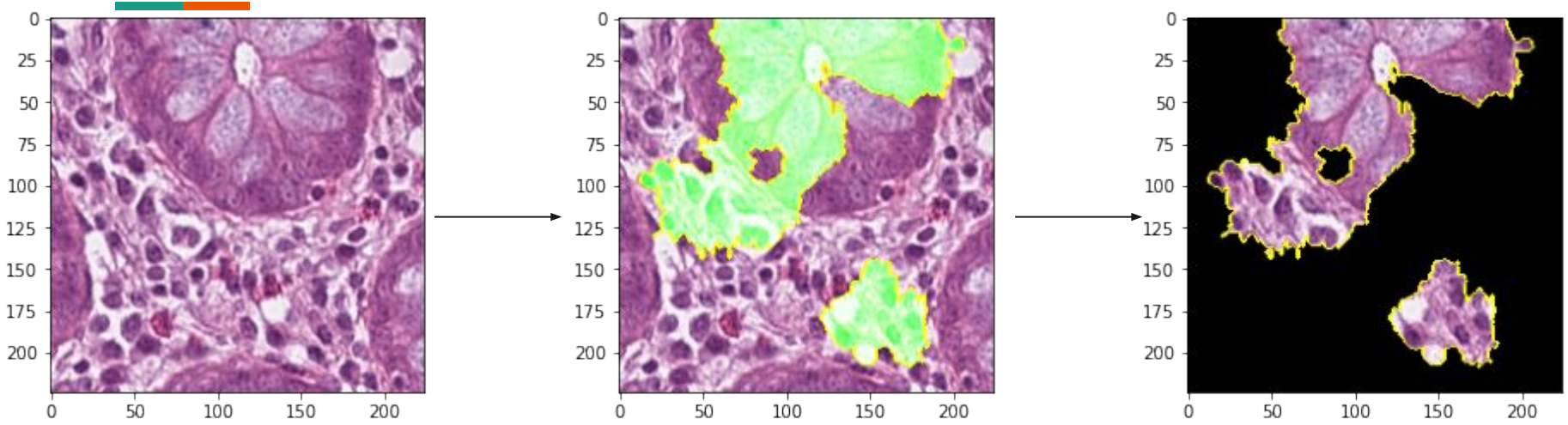
- Three models were iteratively created
    - All of them severely under fitted to the images
- Transfer learning was used as the final, fourth, model
    - ResNet50: was able to predict with 99.3% precision.
    - From 720 images not used in the training process, 5 were incorrectly classified.



Visualizing the Model (CANCER)



Visualizing the Model (NORMAL)



Recommendation

What does the model tell us

- The predictions generated by the trained model took only a little over a couple of seconds to classify 720 images.
- Due to the high efficiency and high results. This model can become an extra tool in the medical field
    - Prioritization of patients with CANCER diagnosis
    - A monitored deployment with a validation team

Next Steps

- Research the market to see where this model would be of the most help

- Use other datasets to view how the model performs

- Consider using an “unstained” version of images to see if the model holds up to the same performance

Questions?

Github: @edgarbarr1

LinkedIn: www.linkedin.com/in/edgar-barr1/

Email: edgarraul98@icloud.com