Camelyon16: Detecting Cancerous Cells in Gigapixel Images

Applied Deep Learning Fall 2020

Alberto Munguia Cisneros - am5334

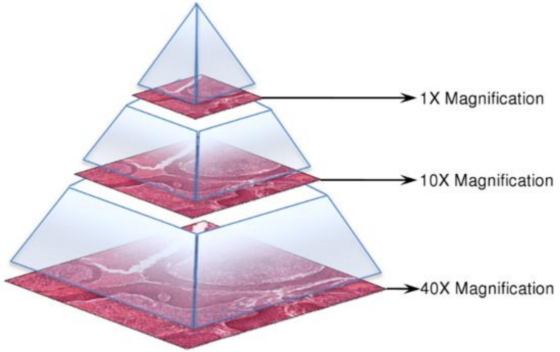
Goal

- The goal of this project is to create an automated process for breast cancer detection to assist pathologists.
- The cancer diagnosis is a high clinical relevance task but requires large amounts of reading time from pathologists.
- An automated process to assist pathologists in cancer detection could help to reduce their workload and reduce the subjectivity in diagnosis.
- This project is based on the paper “Detecting Cancer Metastases on Gigapixel Pathology Images” from Liu et al.

Data

- The source data for the project is the CAMELYON16 challenge dataset.
  - 400 WSI (whole slide images) collected independently from two medical centers in the Netherlands.
  - A subset of 22 images was provided by Prof. Josh Gordon.

- Image files contain multiple downsampled versions of the original image.
- And contain side level annotations.



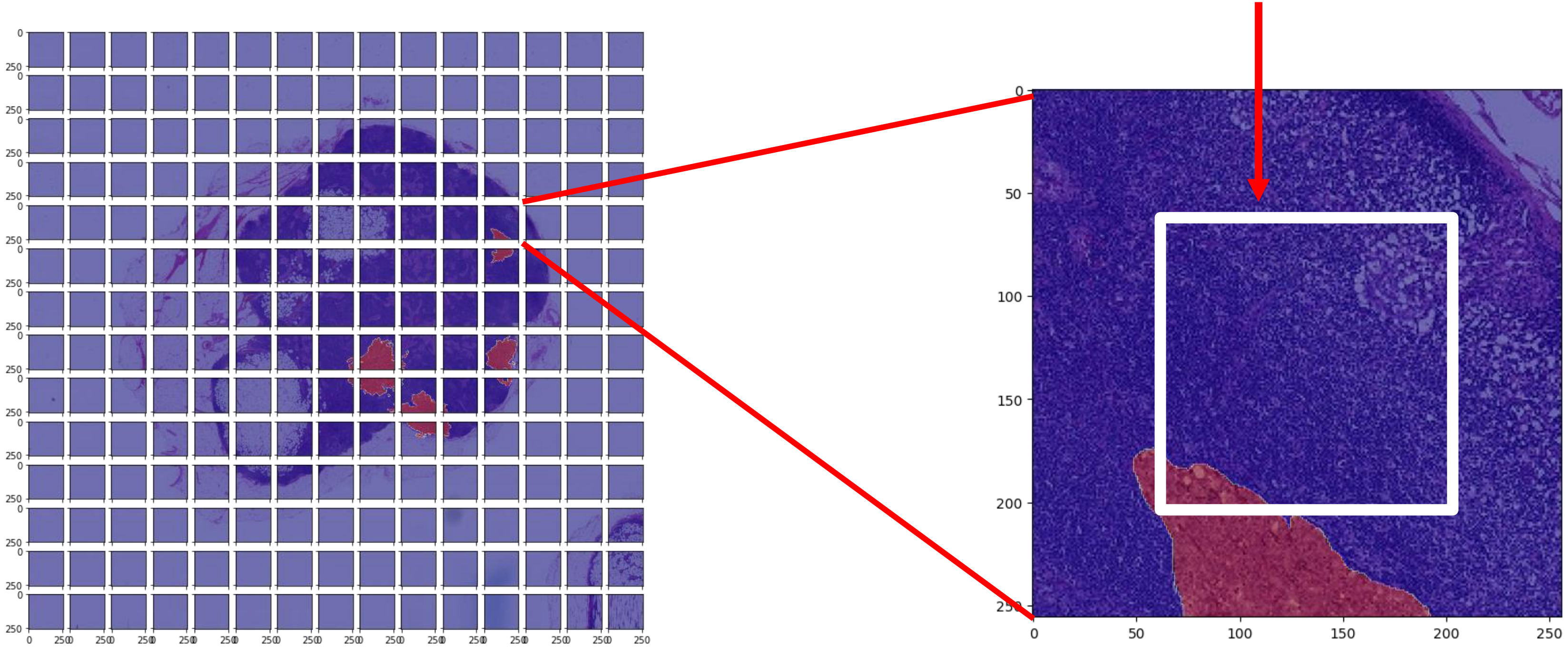
Data Extraction

- Data sets for zoom levels 3, 4, and 5.
- We generated the tiles/patches that constituted our training and test sample with the DeepZoomGenerator function ‘get_tile.'
    - Advantage: Fast and generates all the zoom levels at the same time
    - Disadvantage: Compared with a sliding window method, the number of tiles/patches is limited by its size and the level of depth.
- For labeling each tile, we followed the same approach as the reference paper; we verified if the center of the image (128, 128) contained cancerous cells.

Data Extraction

Get_tile create a grid of tile, for every zoom level

Verify the center of each tile to create the label of tumor or no-tumor



Small and Imbalanced Data

- To deal with the limited number of samples from the previous stage, we used some data augmentation techniques to increase the size and diversity of training sample and to avoid overfitting in our models. Some of the image transformations that we included are:
  - Horizontal and Vertical Flip
  - Random and fixed rotations
  - Shear and zoom range
  - Width and Height shift range

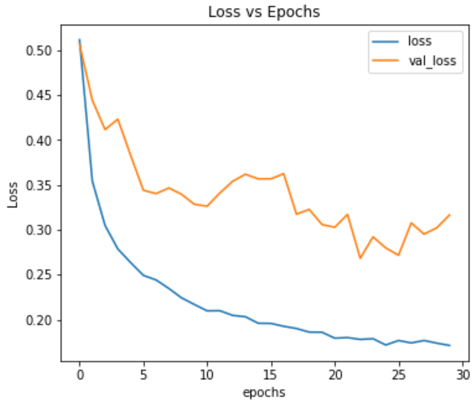
Small and Imbalanced Data

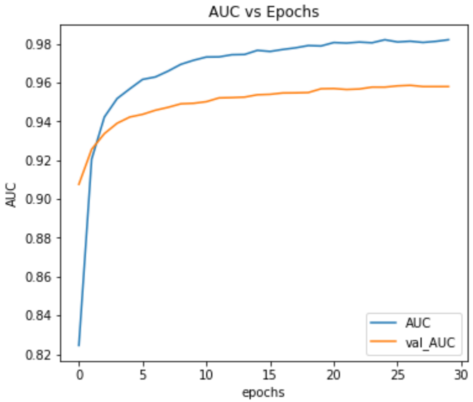
- Also, the data was highly imbalance. The number of tile samples with tumor presence was outnumbered by the tiles with healthy tissue. To overcome this issue, we implemented 2 different approaches to create a balance training set:
    - Oversampling: Match the number of healthy images by over sampling the non-healthy images
    - Undersampling: Match the number of non-healthy images by under sampling the healthy images

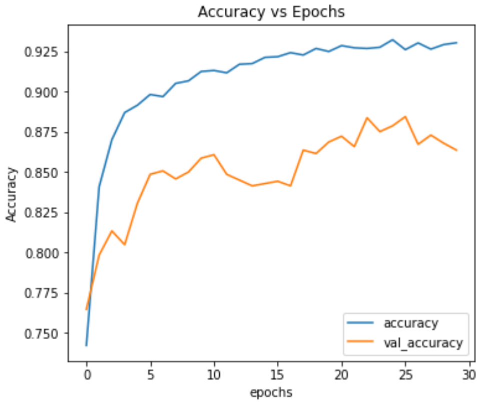
Models and Results

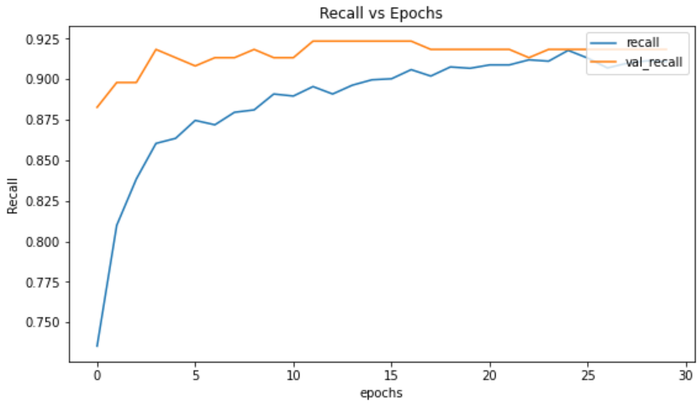
Model	Level	Imbalance Treatment	Metrics on Test Set
Base-CNN (3Conv + 2MaxPool + Dense)	3	Oversampling	Test loss: 0.202, AUC: 0.822, Recall: 0.387, Precision: 0.387, Accuracy: 0.928
Base-CNN (3Conv + 2MaxPool + Dense)	4		Test loss: 0.718, AUC: 0.731, Recall: 0.636, Precision: 0.636, Accuracy: 0.764
Base-CNN (3Conv + 2MaxPool + Dense)	5		Test loss: 0.746, AUC: 0.716, Recall: 0.700, Precision: 0.700, Accuracy: 0.556
VGG16 - Pretrained	4		Test loss: 0.562, AUC: 0.477, Recall: 0.455, Precision: 0.455, Accuracy: 0.719
MobileNet - Pretrained	4		Test loss: 0.171, AUC: 0.771, Recall: 0.091, Precision: 0.091, Accuracy: 0.953
MobileNet - Pretrained	3&4		Test loss: 0.148, AUC: 0.826, Recall: 0.333, Precision: 0.333, Accuracy: 0.957
MobileNet - Pretrained	3		Test loss: 0.164, AUC: 0.861, Recall: 0.452, Precision: 0.452, Accuracy: 0.949

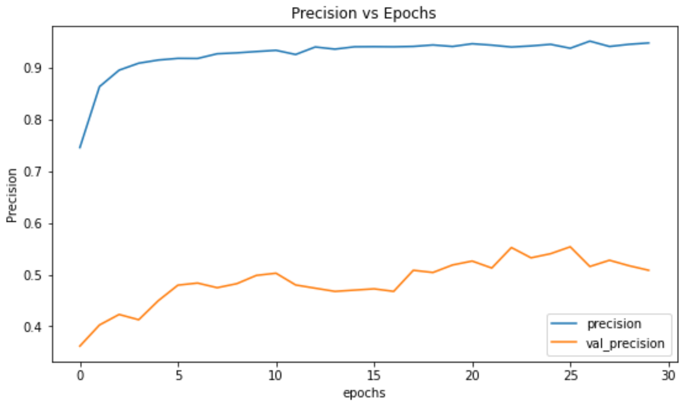
Results MobileNet pretrained zoom level 3







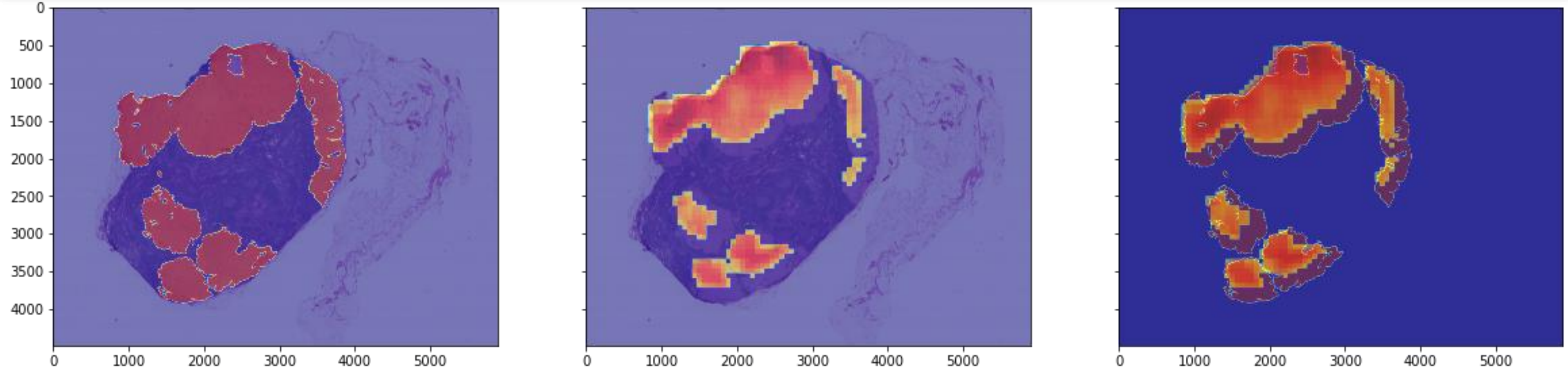




Heat Map and Metrics

- For the creation of the heatmaps we followed a sliding window methodology, where for each window, we predicted the probability of having a tumor.
- Furthermore, we evaluated how good were our models, we implemented the metrics :
  - Confusion Matrix (TP, FP, TN, False Negative)
  - Recall - *TP / (TP +FN)*
  - Precision - *TP / (TP +FP)*
  - F1-score - *2* (Recall * Precision) / (Recall + Precision)*
  - Accuracy – *(TP + TN) / (TP + TN + FP +FN)*

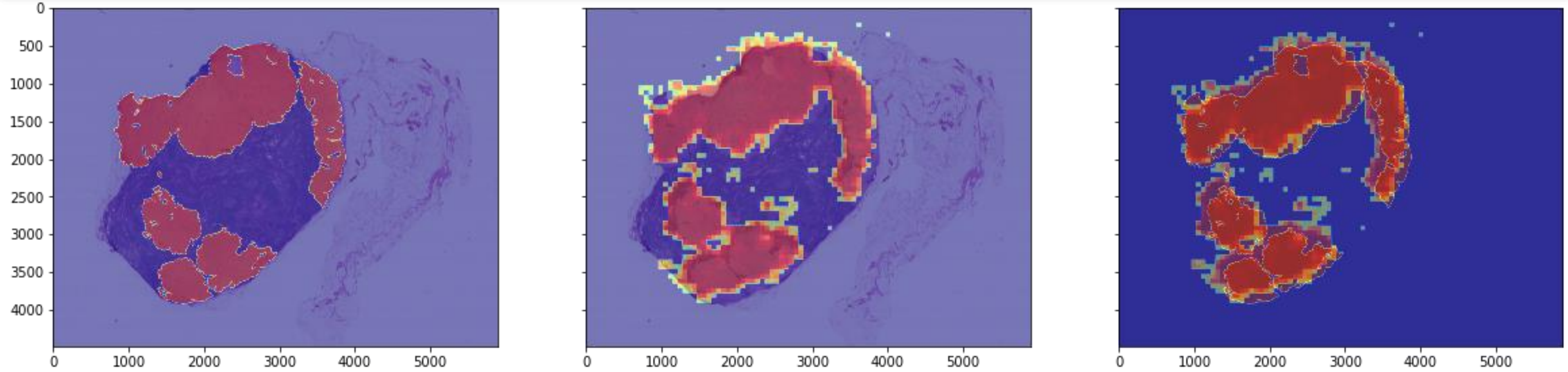
Heatmap Base - CNN




```
ROC AUC score:  0.7984005018587917
True Negative: 22000347
False Positive: 265638
False Negative: 1608997
True Positive: 2503258

              precision    recall  f1-score   support

    No tumor       0.93      0.99      0.96  22265985
       Tumor       0.90      0.61      0.73   4112255

    accuracy                           0.93  26378240
   macro avg       0.92      0.80      0.84  26378240
weighted avg       0.93      0.93      0.92  26378240
```

- If we observe the difference precision and recall, we can infer that the model tend to have a high number of False Negatives (Type Error II)
- The image shows that our base model in fails to predict the cancerous cells in some regions

Heatmap MobileNet pretrained zoom level 3

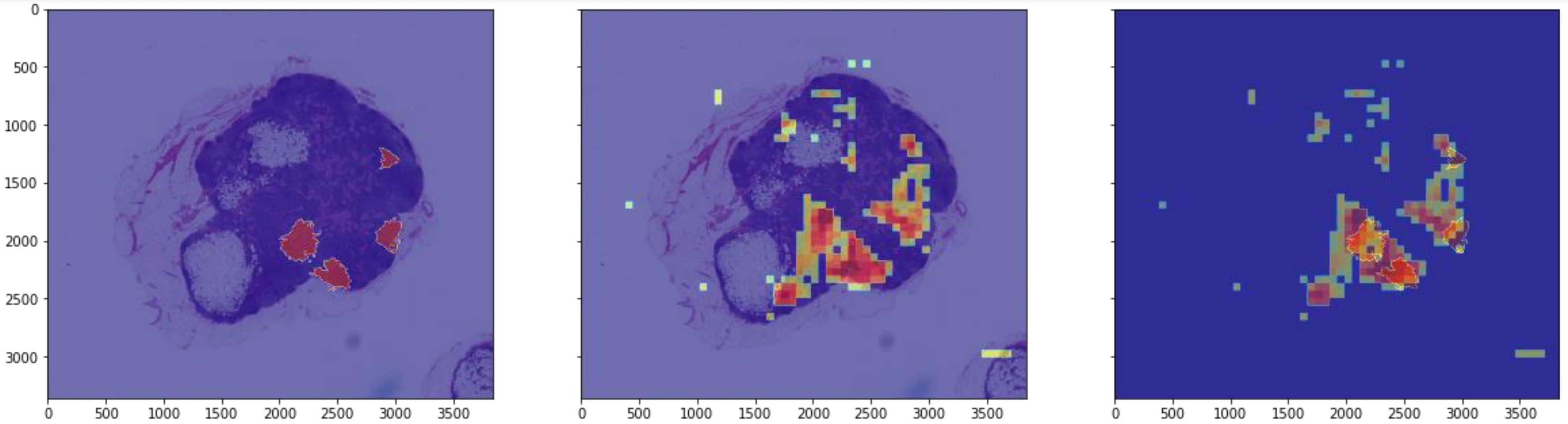



```
ROC AUC score:  0.9087370286271221
True Negative: 20619264
False Positive: 1646721
False Negative: 446464
True Positive: 3665791
```

	precision	recall	f1-score	support
No tumor	0.98	0.93	0.95	22265985
Tumor	0.69	0.89	0.78	4112255
accuracy			0.92	26378240
macro avg	0.83	0.91	0.86	26378240
weighted avg	0.93	0.92	0.92	26378240

- This model predicts better the cancerous cells
- The recall and precision metrics are more balance, with a slightly tendency of our model to predict False Positives (Type Error I)

Heatmap MobileNet pretrained level 3




```
ROC AUC score:  0.8602587417899545
True Negative: 12044882
False Positive: 670469
False Negative: 42414
True Positive: 144635
```

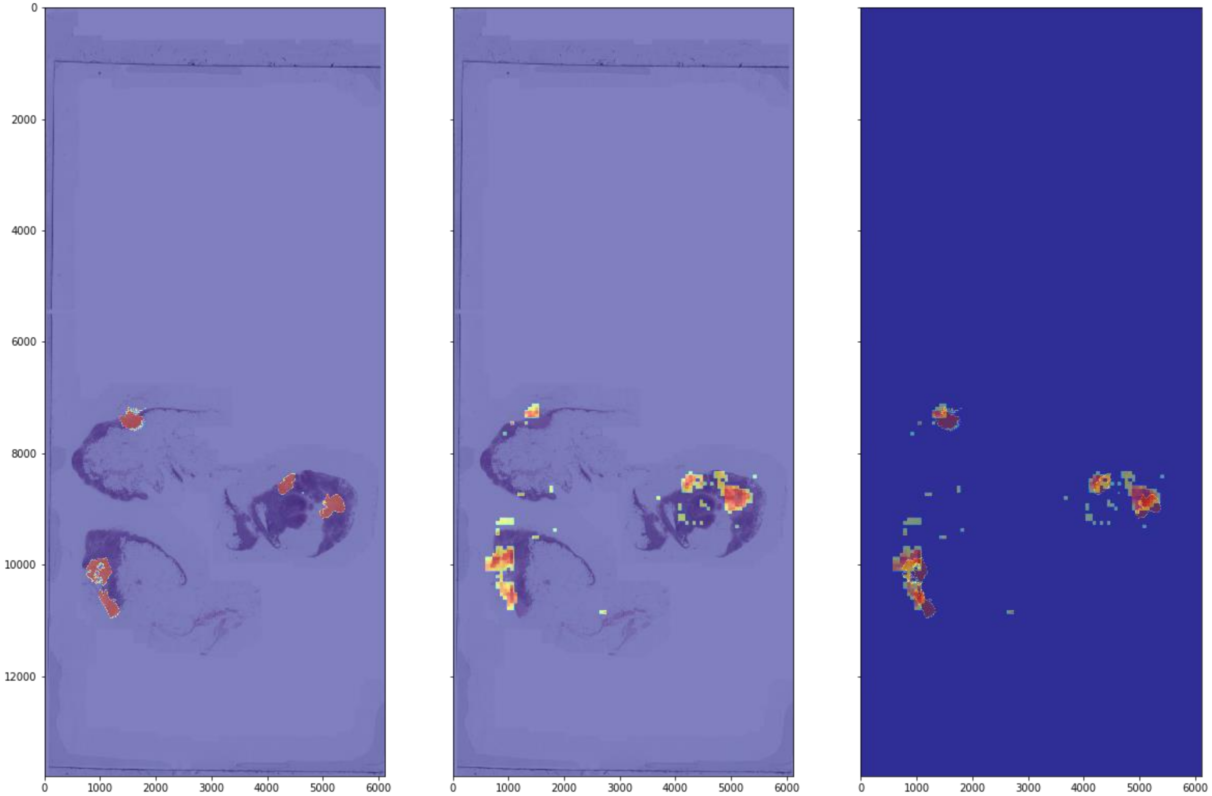
	precision	recall	f1-score	support
No tumor	1.00	0.95	0.97	12715351
Tumor	0.18	0.77	0.29	187049
accuracy			0.94	12902400
macro avg	0.59	0.86	0.63	12902400
weighted avg	0.98	0.94	0.96	12902400

- We confirm that the best model tend to predict False Positives (Type Error I)

Heatmap MobileNet pretrained level 3

```
ROC AUC score:  0.7324960474044238
True Negative: 83119983
False Positive: 653959
False Negative: 275601
True Positive: 247161
```

	precision	recall	f1-score	support
No tumor	1.00	0.99	0.99	83773942
Tumor	0.27	0.47	0.35	522762
accuracy			0.99	84296704
macro avg	0.64	0.73	0.67	84296704
weighted avg	0.99	0.99	0.99	84296704



Conclusions and Future Work

- Data
  - Augmenting the data and creating balance sample had a positive impact on the performance of the models.
  - The best models were those who were trained with the highest zoom level (level 3).
  - Hence, an area of opportunity is to generate data from levels 0, 1, and 2 to train our models, although more computational power will be needed.

Conclusions and Future Work

- Model Architecture
  - Although some of our models, like MobileNet trained, achieved an acceptable performance; they were prone to type I error.
  - An area of opportunity is the Implementation of finetuned models and multilevel models, for example twin tower architectures.

Code

The code was divided in 6 books:

Book	Description	Input	Output
Book 1 Data Exploration	This book is an initial exploratory analysis to get familiar with the data and with open_slide library	Images	None
Book 2 Data Generation	This book generates the training, validation and test sets and stores them in google drive.	Images + Masks	Training, Validation and Test Sets
Book 3 Model Training CNN	Convolutional model implementation	Training, Validation and Test Sets	Model
Book 4 Model Training VGG16	VGG16 implementation	Training, Validation and Test Sets	Model
Book 5 Model Training MobileNet	MobileNet implementation	Training, Validation and Test Sets	Model
Book 6 HeatMap	This load any model and generates the prediction heatmap and the performance report for a specific slide	Model + Images	Heatmap + Performance report

THANK YOU