STAT 8002: Project Presentation

Applying Deep Learning Algorithms in Predicting Prostate Cancer

Jason Chan Jin An   UID: 2012610156

Supervisor: Dr. Dora Zhang Yan

Agenda

Dataset Introduction & Objectives

Masks Explanation and Challenges

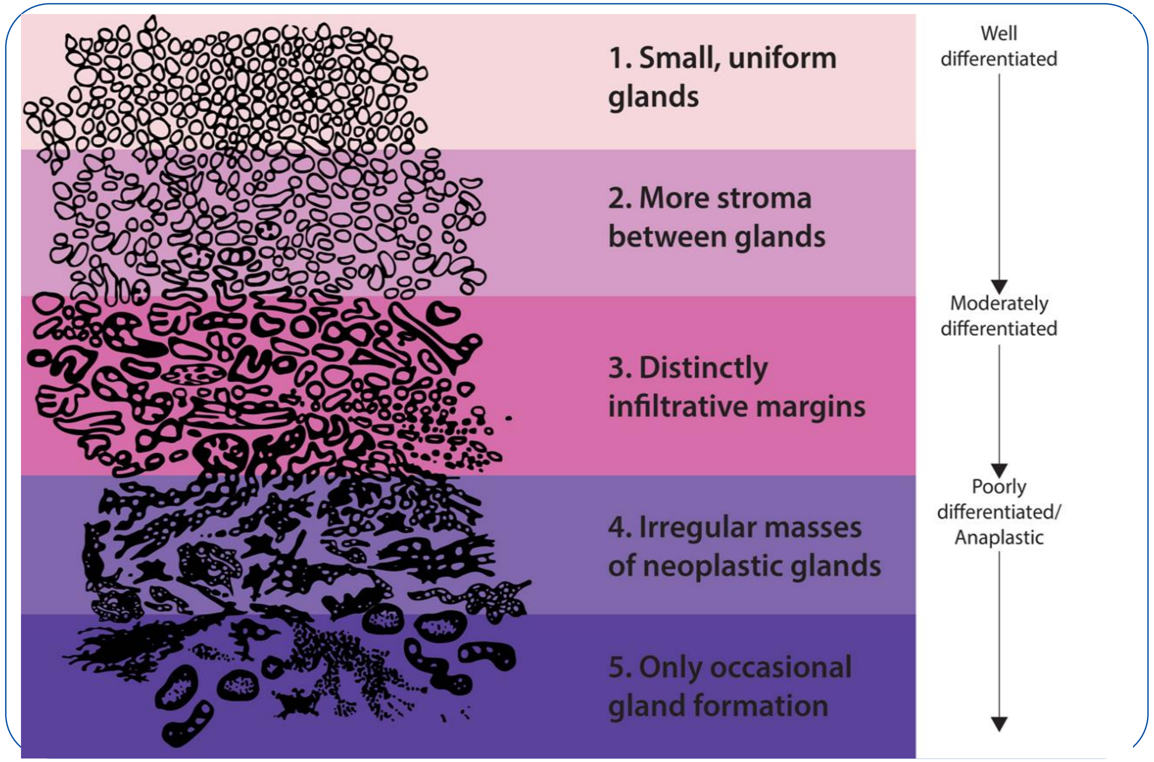
Tiling Images

One Stage Model

Two Stage Model

Results & Conclusion

Further Improvements & QnA



Deep Learning Algorithms in Predicting Prostate Cancer

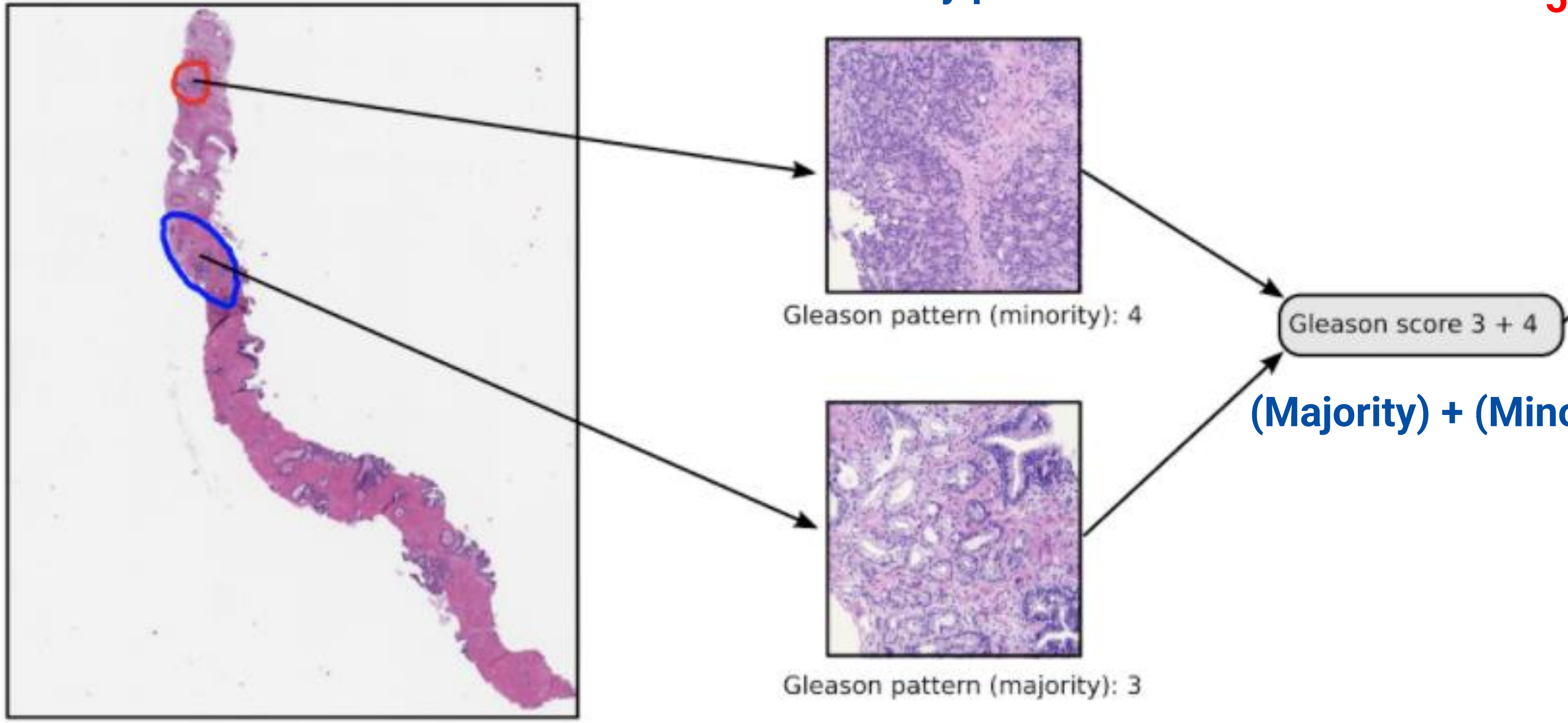
Dataset Objective

Dataset Objective – Image Classification

What is the input and what are we trying to predict?

Input

Pathologists labels the majority and minority patterns



Whole slide image of a prostate biopsy

(Majority) + (Minority)

Output

ISUP* grade is what we want to predict. 0 means not cancerous. ISUP **from 0 to 5**.

Gleason score	ISUP grade
3+3	1
3+4	2
4+3	3
4+4	4
3+5	4
5+3	4
4+5	5
5+4	5
5+5	5

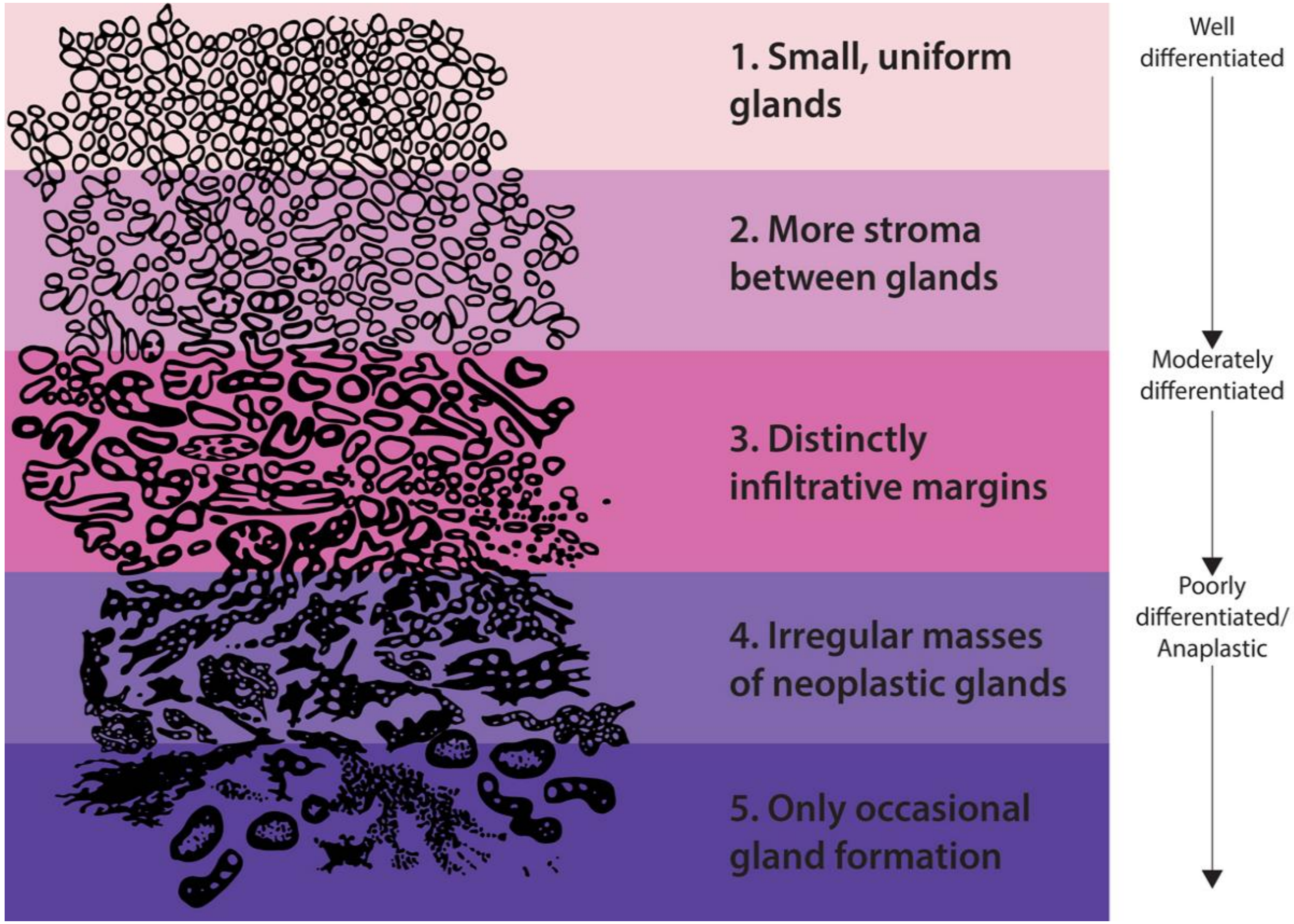
Less aggressive

Aggressive

**ISUP = International Society of Urological Pathology*

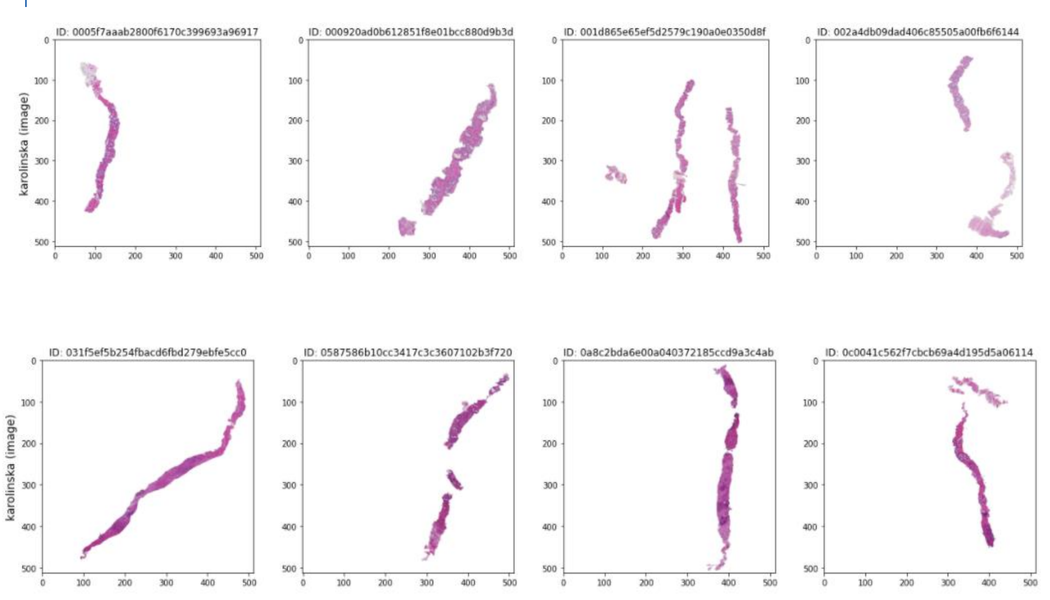
Dataset Objective

Severity of Prostate Cancer



Dataset Introduction – What's the source?

Dataset



~10,000 H&E stained prostate biopsy images

- **Size:** 383 GB
- **Format: TIFF (multi level image)**
- **Labels:** Yes (labelled by pathologists)
- **Masks:** Yes (labelled areas of interest by pathologists)

Source from Organizations ~(50:50)

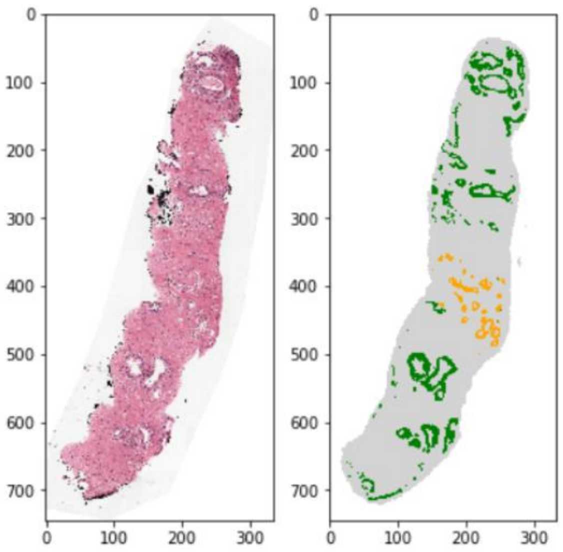


Image Mask*

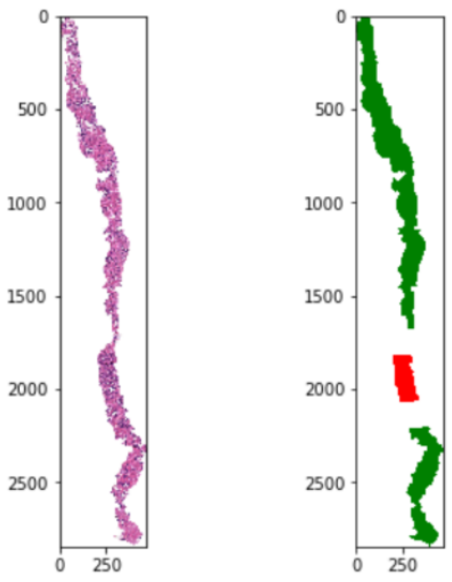


Image Mask*

Radboudumc university medical center

Karolinska Institutet

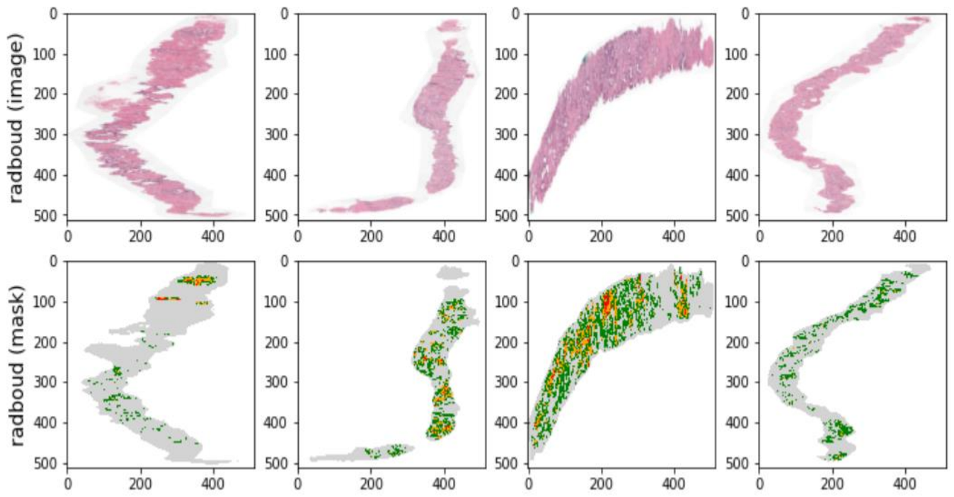
Note that these 2 sources have different masking techniques

Masking Explanation – What do the masks mean?

Radbound

Image

Mask



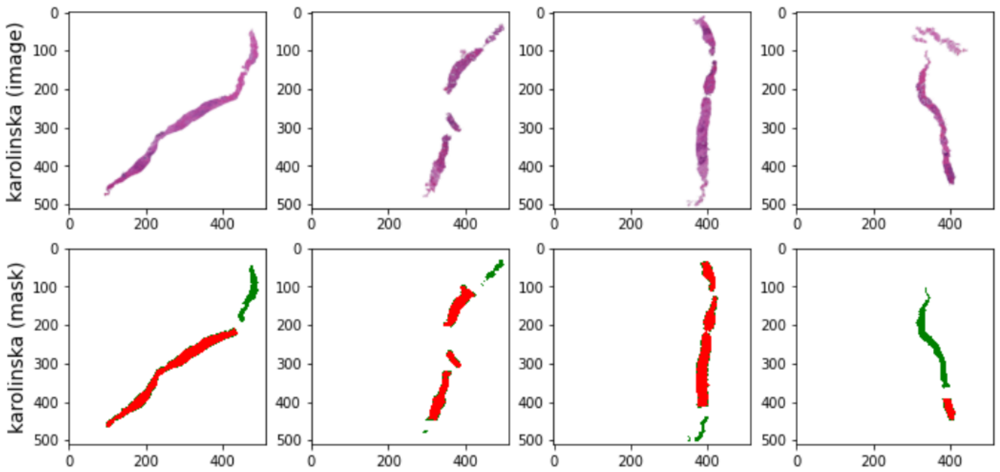
Mask colors and description

- 0: background (non tissue) or unknown
- 1: stroma (connective tissue, non-epithelium tissue)
- 2: healthy (benign) epithelium
- 3: cancerous epithelium (Gleason 3)
- 4: cancerous epithelium (Gleason 4)
- 5: cancerous epithelium (Gleason 5)

Karolinska

Image

Mask



Mask colors and description

- [0]: background (non tissue) or unknown
- [1]: benign tissue (stroma and epithelium combined)
- [2]: cancerous tissue (stroma and epithelium combined)

Challenges with Masks

Masks are not the same

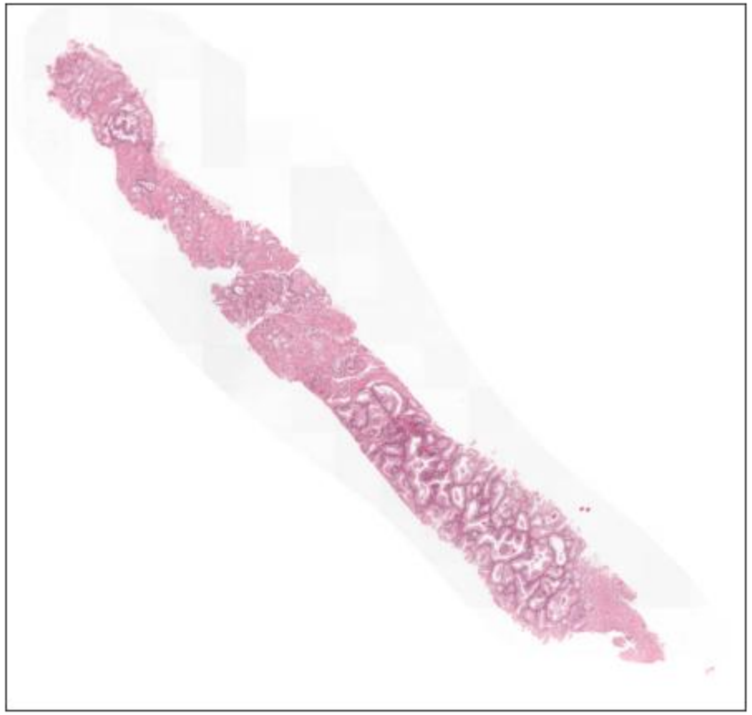
- From previous slide, obvious that both sources have different masking techniques
- If want to use, just use the more granular source, which is Radboud

No masks in production

- In production, just the image will be available and no masks will be available
- A segmentation model must be trained to produce/predict masks of the images

Dataset Introduction – Image & Mask Resolution

Highest



Avg size: 40MB

Medium

Avg size: 20MB

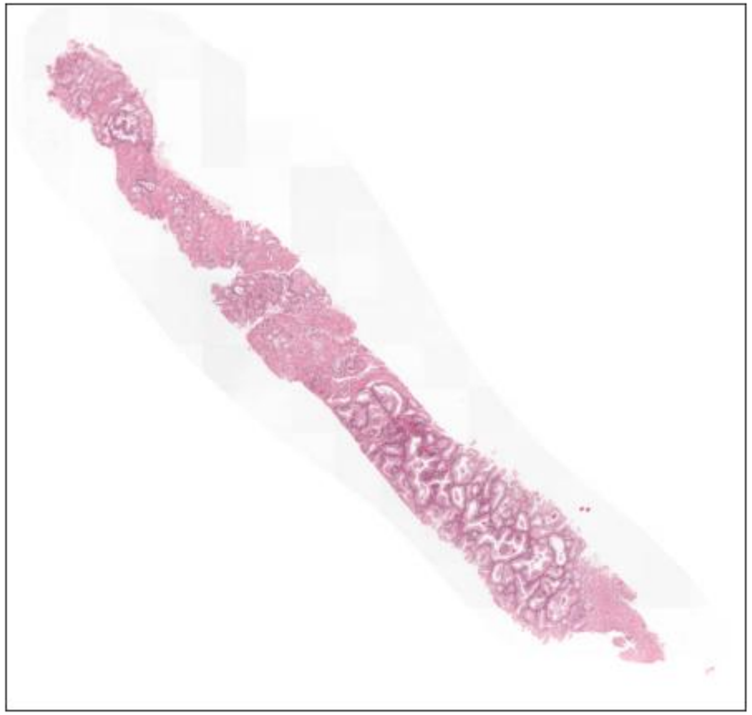
Lowest

Avg size: 10MB

- Images come in TIFF multilevel format, has 3 resolutions available

Dataset Introduction – Image Resolution

Highest



Avg size: 40MB

Medium

Avg size: 20MB

Lowest

Avg size: 10MB

- Images come in TIFF multilevel format, has 3 resolutions available
- **Lowest resolution is chosen for this project due to limited hardware resources (GPU and RAM)**
- Lowest resolution is actually already quite large in size

Evaluation – Quadratic Weighted Kappa (QWK)

QWK with N classes

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$$

i = actual

j = predicted

Observed Confusion Matrix

N x N confusion matrix of prediction classification

Expected Confusion Matrix

N x N confusion matrix under the assumption of no correlation among classes. Outer product of actual and predicted labels.

E = np.outer(actual, predicted)

Penalty matrix, N x N

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

N x N confusion matrix of penalties. If predicted is equal to actual, zero penalty.

Why use QWK?

- Allocate a higher penalty score if our prediction is further away from the actual value.
- Hierarchy matters in this case. Predicting ISUP grade 1 for sample supposedly grade 5 has huge consequences.

isup_grade
0
1
2
4
3
5

Actual = 2 , Predicted = 1

Penalty:

$$\frac{(2-1)^2}{(5-1)^2} = 0.0625$$

Actual = 2 , Predicted = 4

Penalty:

$$\frac{(2-4)^2}{(5-1)^2} = 0.25$$

Evaluation – Quadratic Weighted Kappa (QWK)

QWK Values Interpretation

Range of Quadratic Weighted Kappa	Concordance
Negative	poor
0.01–0.20	slight
0.21–0.40	fair
0.41–0.60	moderate
0.61–0.80	substantial
0.81–1	almost perfect

Project target

Usage in Python

From sklearn library

```
cohen_kappa_score(actual, pred, labels=None, weights= 'quadrati
c', sample_weight=None)
```

Data Cleansing – Omitting Suspicious Images

Images with no masks

```
Image name: e4215cfc8c41ec04a55431cc413688a9
ISUP Grade: 2
Data Provider: karolinska
Original Resolution of Image: (1792, 1248)
Original resolution of Mask: (1792, 1248)
```

4 images

No Cancerous Regions but labelled cancerous

```
Image name: fe79209ab178c89a9be62bc05b63f083
ISUP Grade: 4
Data Provider: radboud
Original Resolution of Image: (624, 256)
Original resolution of Mask: (624, 256)
```

82 images

- Removing these images removes noise while training model

Modelling

Modelling Approach

One Stage Model

Stained Image

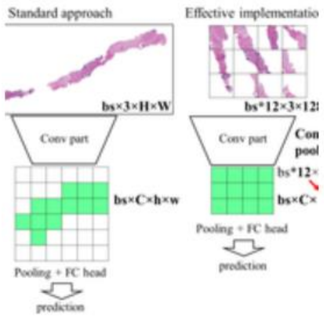
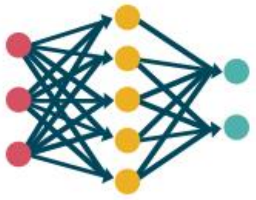


Image Tiling



Deep Learning Model

isup_grade
0
1
2
4
3
5

Classification

Two Stage Model

Stained Image

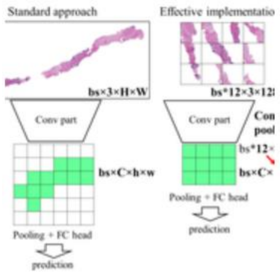
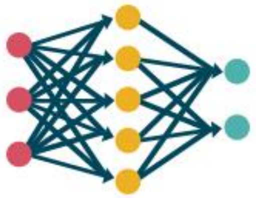
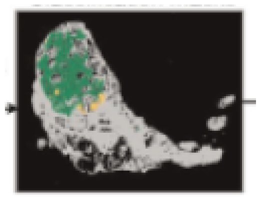


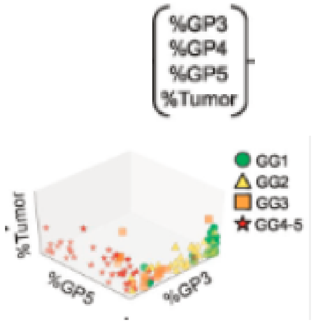
Image Tiling



Segmentation Model



Segmentation of Masks



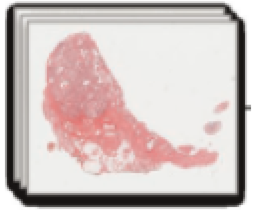
Feature Extraction & 2nd Model

isup_grade
0
1
2
4
3
5

Classification

Modelling Approach Recap

One Stage Model



Stained Image

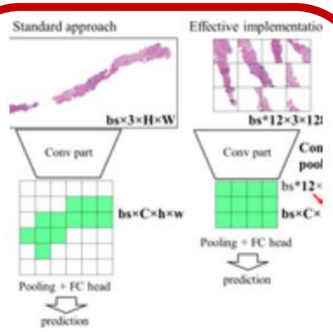
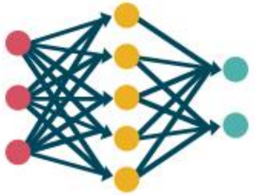


Image Tiling

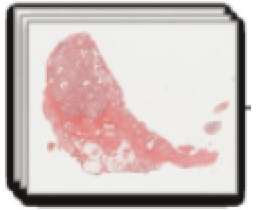


Deep Learning Model

isup_grade
0
1
2
4
3
5

Classification

Two Stage Model



Stained Image

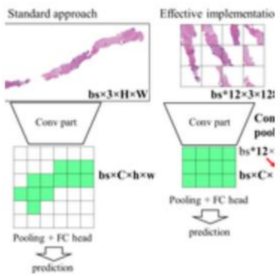
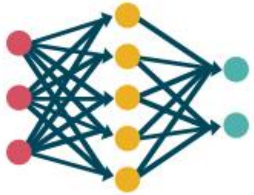
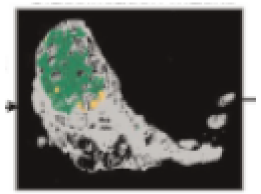


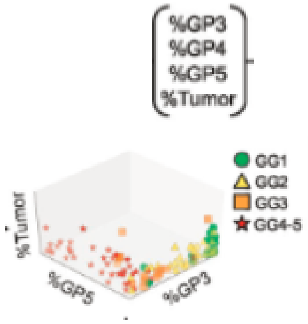
Image Tiling



Segmentation Model



Segmentation of Masks



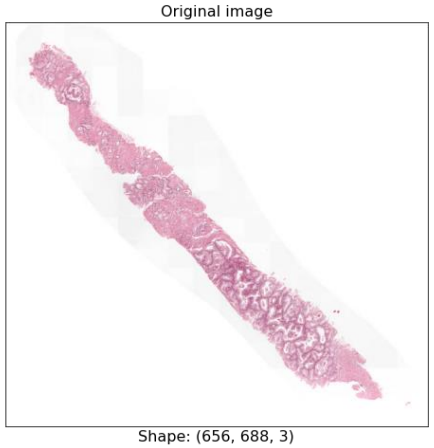
Feature Extraction & 2nd Model

isup_grade
0
1
2
4
3
5

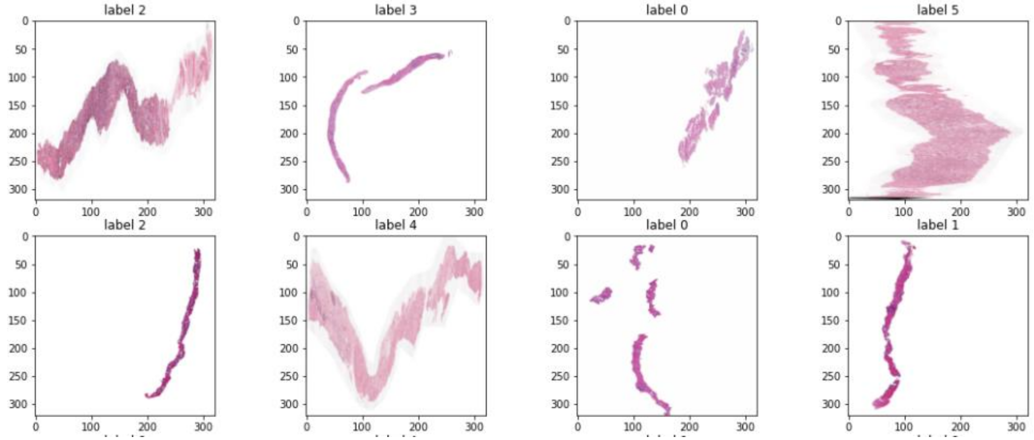
Classification

Why Must Images be Tiled?

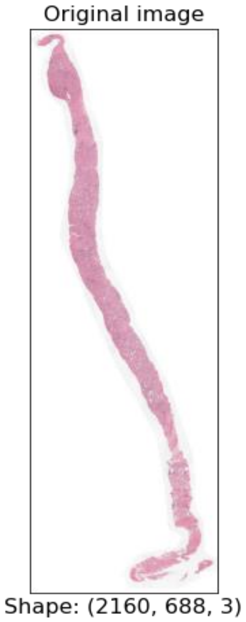
Deep learning model needs to train on **same sized images**



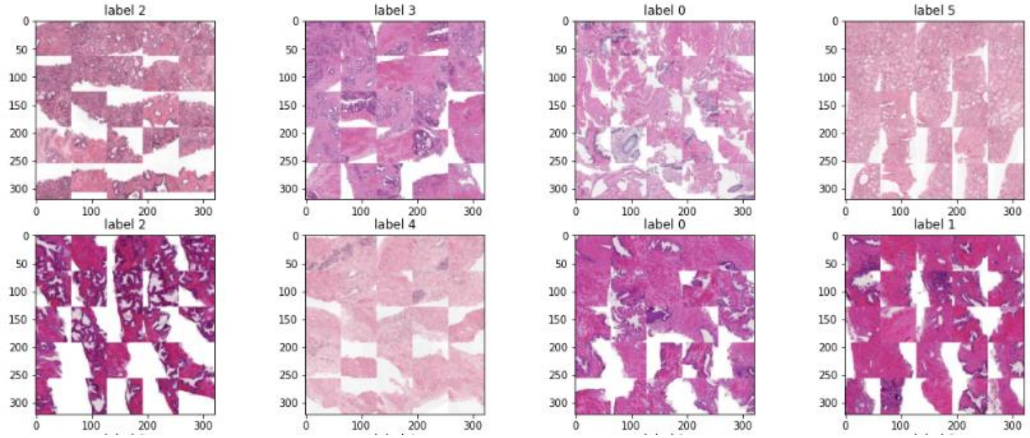
Option 1: Resize images to fixed size



- A lot of white space and hard to interpret
- **Serves as a good benchmark**



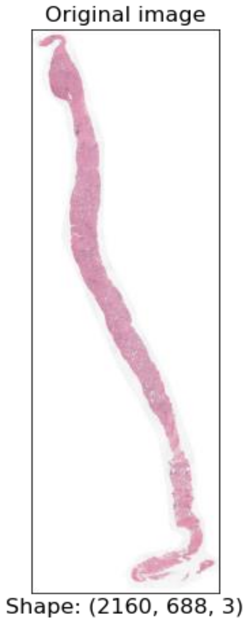
Option 2: Tile Images To fixed size



- A lot less white space
- Easier for model to interpret

Static Tiling Algorithm

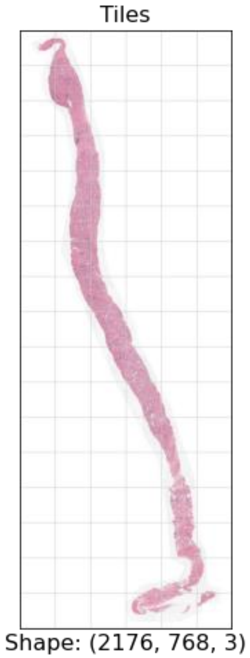
16 Tiles, each tile 128 x 128 (16 x 128 x 128)



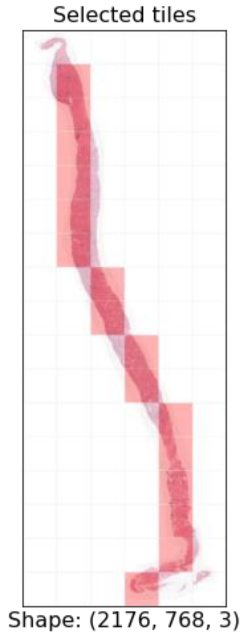
Original Image

Pad Image to nearest multiple of 128, here is 1024 & 512

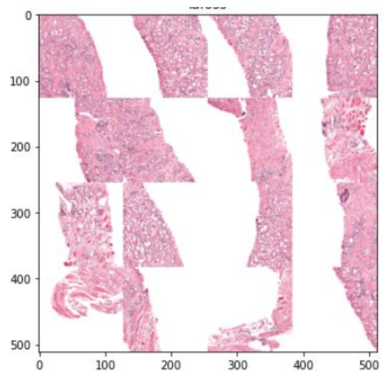
Grey pad to illustrate, in reality, I use white.



Fill Image with 128 by 128 tiles



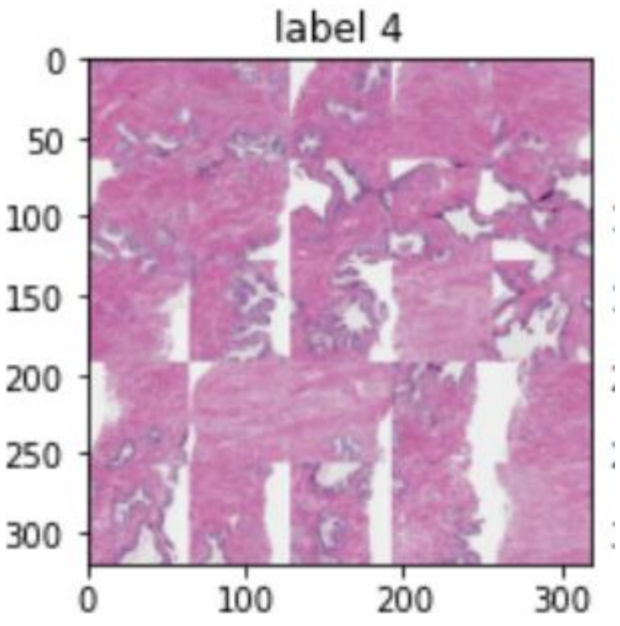
Select top 16 highest concentration tiles



What if not enough tiles (a lot of white space)?

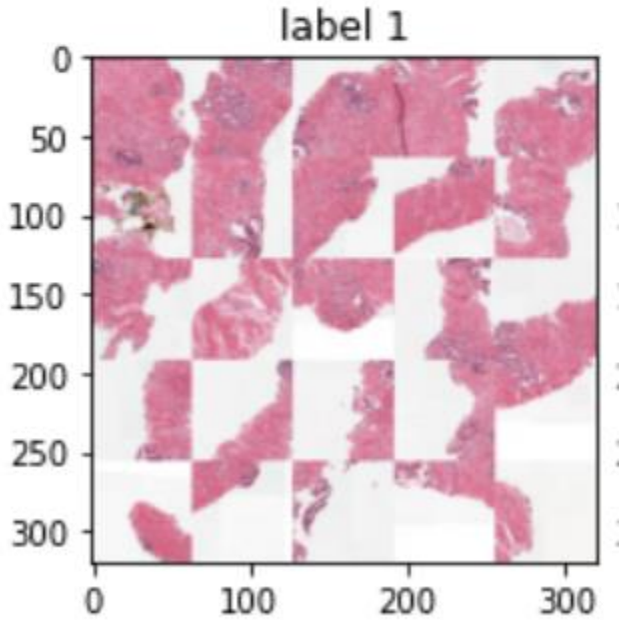
Pad the remaining tiles needed to complete 16 tiles
Can consider excluding the images that has not enough to tile, doesn't make sense to tile if too little non white.

Good vs Bad Tiles (312 x 312)



Good Tile

- Not too much white space
- Capture a lot of content

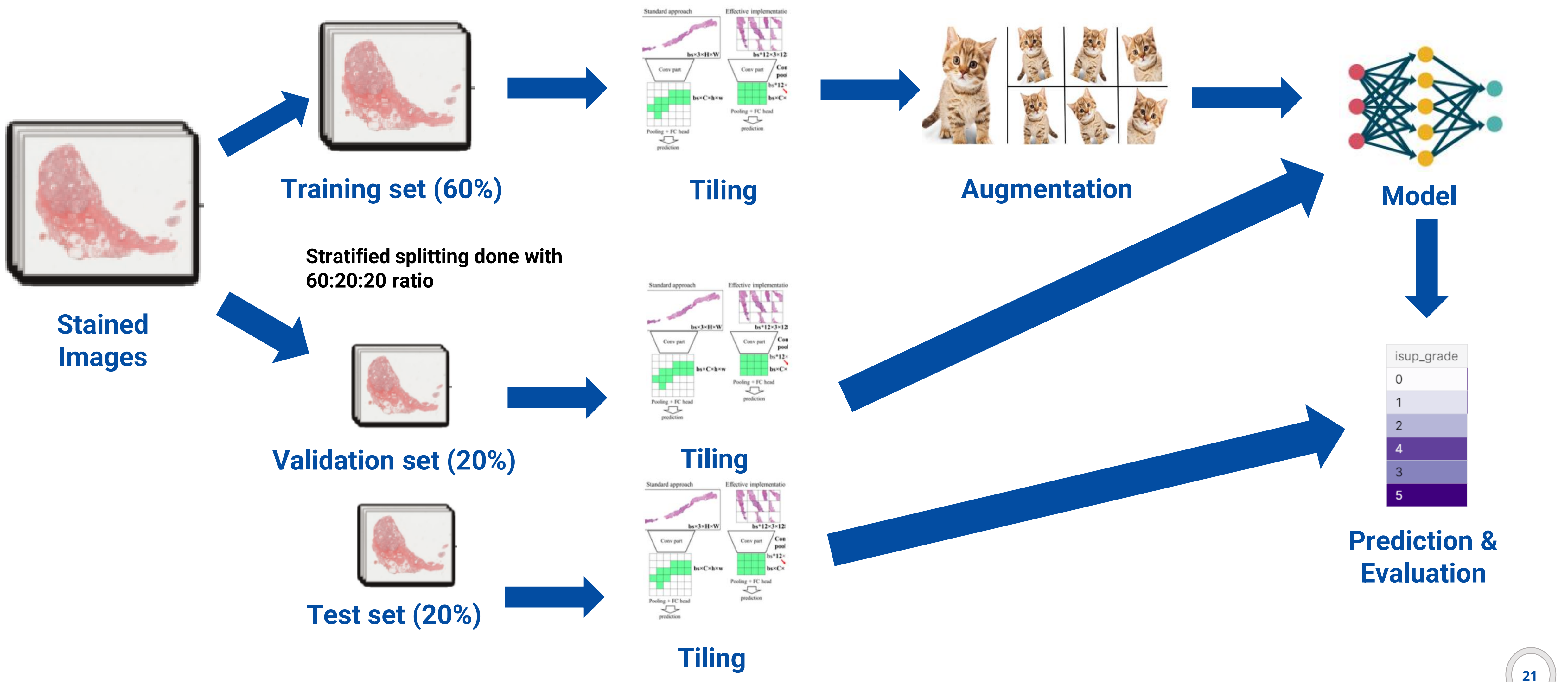


Bad Tile

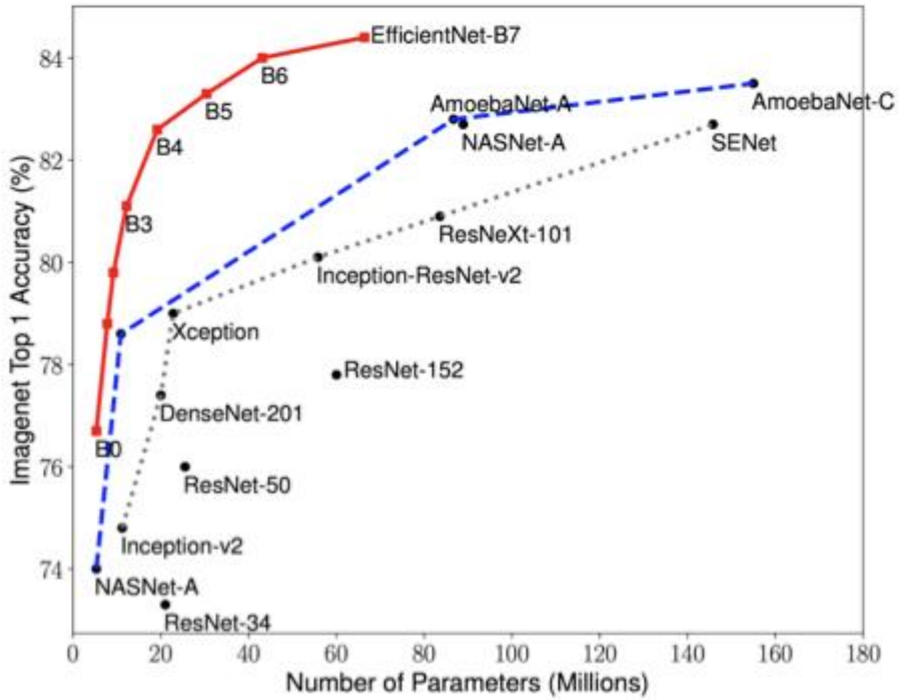
- Too much white space
- Can't capture a lot of content and too much white space not useful for model

One Stage Model

One Stage Detailed Modelling Workflow



Models Tested for One Stage Modelling



The following models were tested:

- Xception
- ResNet-50
- Inception-ResNet-v2
- Inception-v3
- EfficientNet-B5

Wide range of complexity (in terms of number of parameters)

One Stage Model Results – Tiling improves model

Tiling

* Use pre-trained weights because those weights have been trained over many images and if train from scratch, accuracy increases very slow and will overfit

Tiles	Weight*	Training Time	Epoch	Backbone Model	Train QWK	Validation QWK	Test QWK
25 x 64 x 64	imagenet	8 hrs	25	Xception	**0.78**	**0.76**	**0.75**
25 x 64 x 64	imagenet	8 hrs	25	ResNet50	0.73	0.71	0.73
25 x 64 x 64	imagenet	8 hrs	25	InceptionResNet	0.74	0.73	0.69
25 x 64 x 64	imagenet	8 hrs	25	Inception	0.73	0.69	0.69
25 x 64 x 64	imagenet	9 hrs	25	EfficientNet-B5	0.65	0.68	0.68

No Tiling

Tiles	Weight	Training Time	Epoch	Backbone Model	Train QWK	Validation QWK	Test QWK
None	imagenet	12 hrs	25	Xception	**0.62**	**0.57**	**0.57**

One Stage Model Results – Tiling improves model

Tiling

* Use pre-trained weights because those weights have been trained over many images and if train from scratch, accuracy increases very slow and will overfit

Tiles	Weight*	Training Time	Epoch	Backbone Model	Train QWK	Validation QWK	Test QWK
25 x 64 x 64	imagenet	8 hrs	25	Xception	**0.78**	**0.76**	**0.75**
25 x 64 x 64	imagenet	8 hrs	25	ResNet50	0.73	0.71	0.73
25 x 64 x 64	imagenet	8 hrs	25	InceptionResNet	0.74	0.73	0.69
25 x 64 x 64	imagenet	8 hrs	25	Inception	0.73	0.69	0.69
25 x 64 x 64	imagenet	9 hrs	25	EfficientNet-B5	0.65	0.68	0.68

No Tiling

Tiles	Weight	Training Time	Epoch	Backbone Model	Train QWK	Validation QWK	Test QWK
None	imagenet	12 hrs	25	Xception	**0.62**	**0.57**	**0.57**

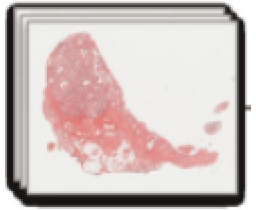
One Stage Model Conclusion

- Tiling the images were useful, compared to not tiling them.
- Significant increase in QWK score when tiling. Average of 0.59 vs 0.76
- Xception model performs the best and QWK scores across train, validation, and test shows no signs of overfitting

Two Stage Model

Modelling Approach Recap

One Stage Model



Stained Image

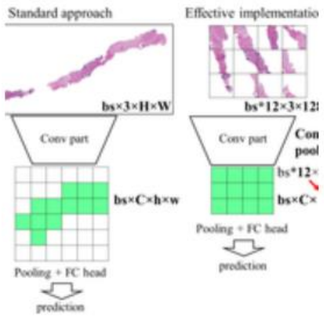
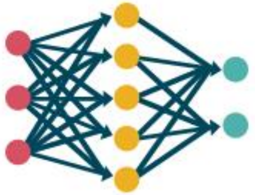


Image Tiling

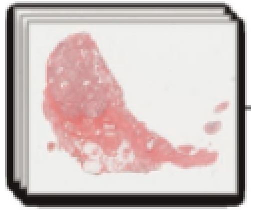


Deep Learning Model

isup_grade
0
1
2
4
3
5

Classification

Two Stage Model



Stained Image

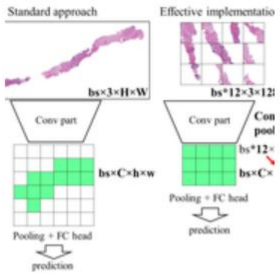
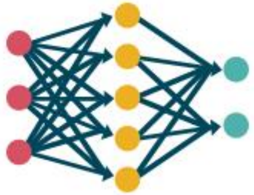
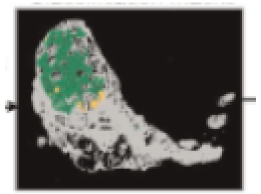


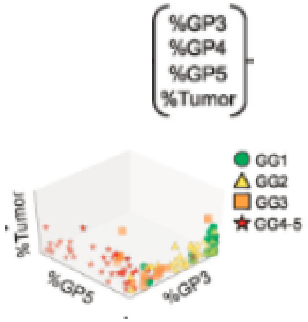
Image Tiling



Segmentation Model



Segmentation of Masks



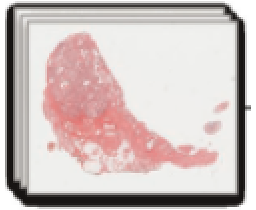
Feature Extraction & 2nd Model

isup_grade
0
1
2
4
3
5

Classification

Modelling Approach Recap

One Stage Model



Stained Image

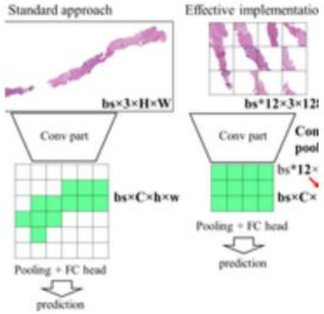
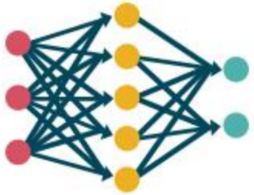


Image Tiling

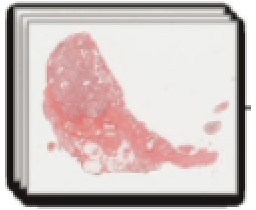


Deep Learning Model

isup_grade
0
1
2
4
3
5

Classification

Two Stage Model



Stained Image

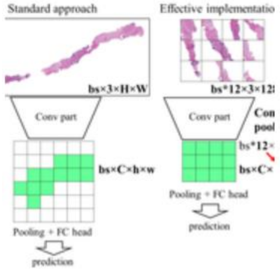
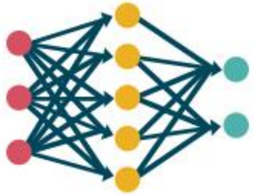
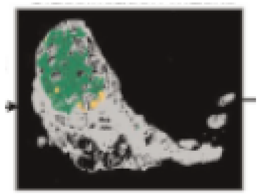


Image Tiling

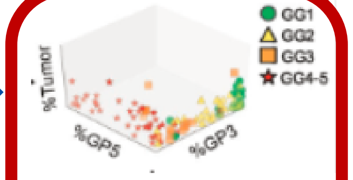


Segmentation Model



Segmentation of Masks

%GP3
%GP4
%GP5
%Tumor



Feature Extraction & 2nd Model

isup_grade
0
1
2
4
3
5

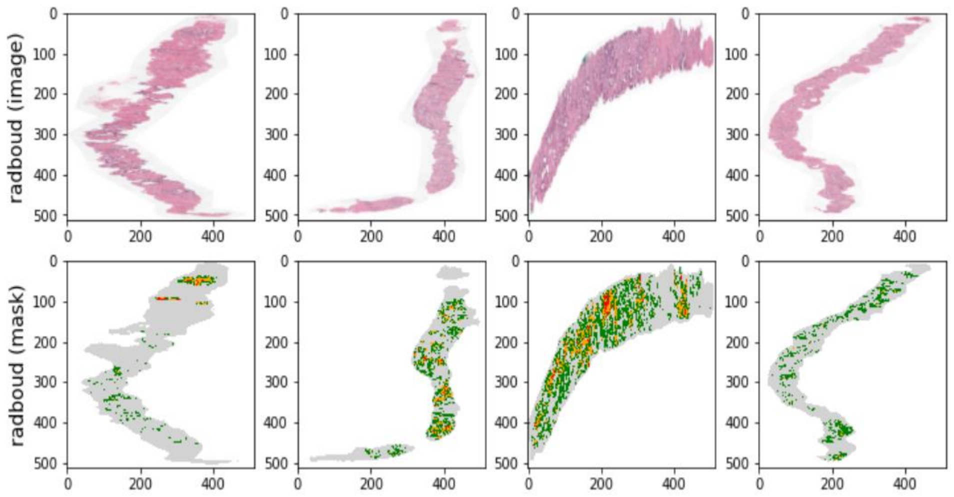
Classification

Different Masks

Radbound

Image

Mask



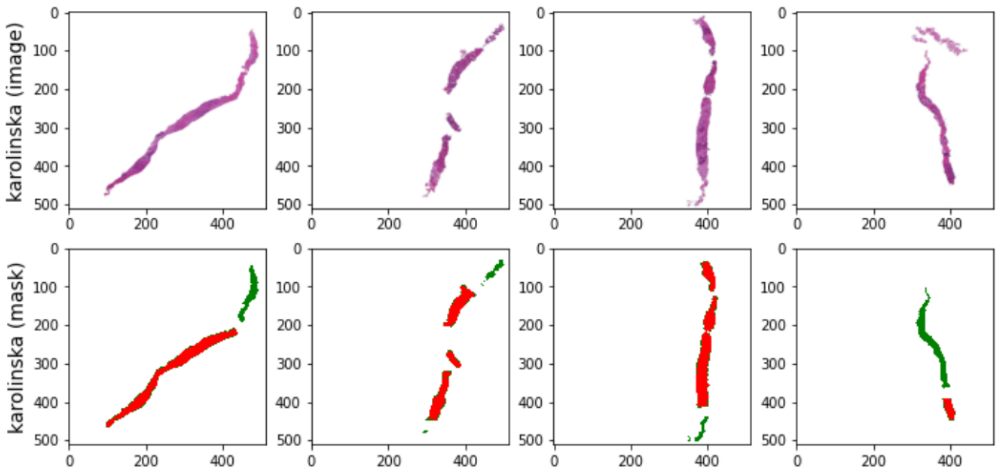
Mask colors and description

- 0: background (non tissue) or unknown
- 1: stroma (connective tissue, non-epithelium tissue)
- 2: healthy (benign) epithelium
- 3: cancerous epithelium (Gleason 3)
- 4: cancerous epithelium (Gleason 4)
- 5: cancerous epithelium (Gleason 5)

Karolinska

Image

Mask



Mask colors and description

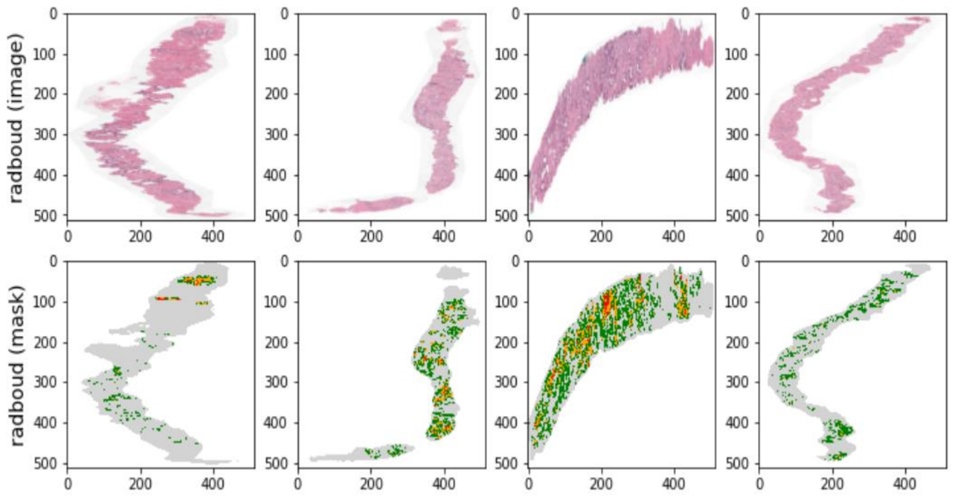
- [0]: background (non tissue) or unknown
- [1]: benign tissue (stroma and epithelium combined)
- [2]: cancerous tissue (stroma and epithelium combined)

Different Masks

Radboud

Image

Mask



Mask colors and description

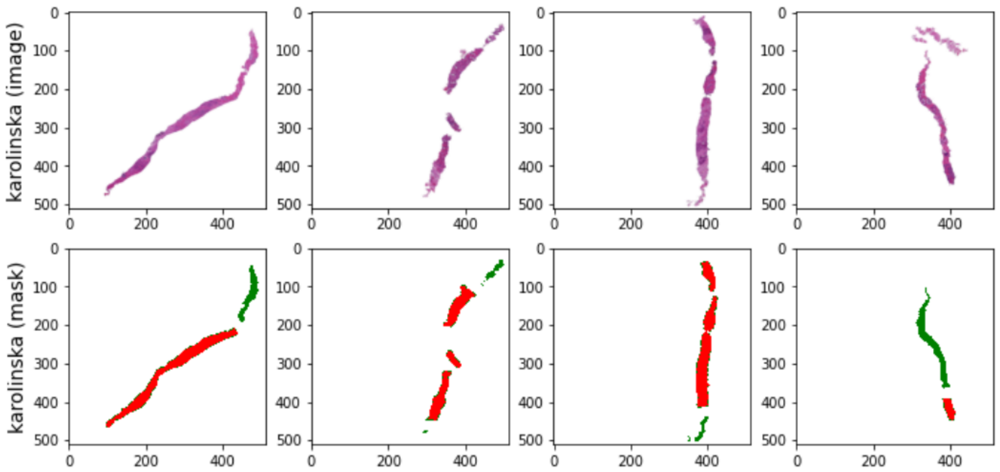
- 0: background (non tissue) or unknown
- 1: stroma (connective tissue, non-epithelium tissue)
- 2: healthy (benign) epithelium
- 3: cancerous epithelium (Gleason 3)
- 4: cancerous epithelium (Gleason 4)
- 5: cancerous epithelium (Gleason 5)

Use Radboud to train model as it is more granular

Karolinska

Image

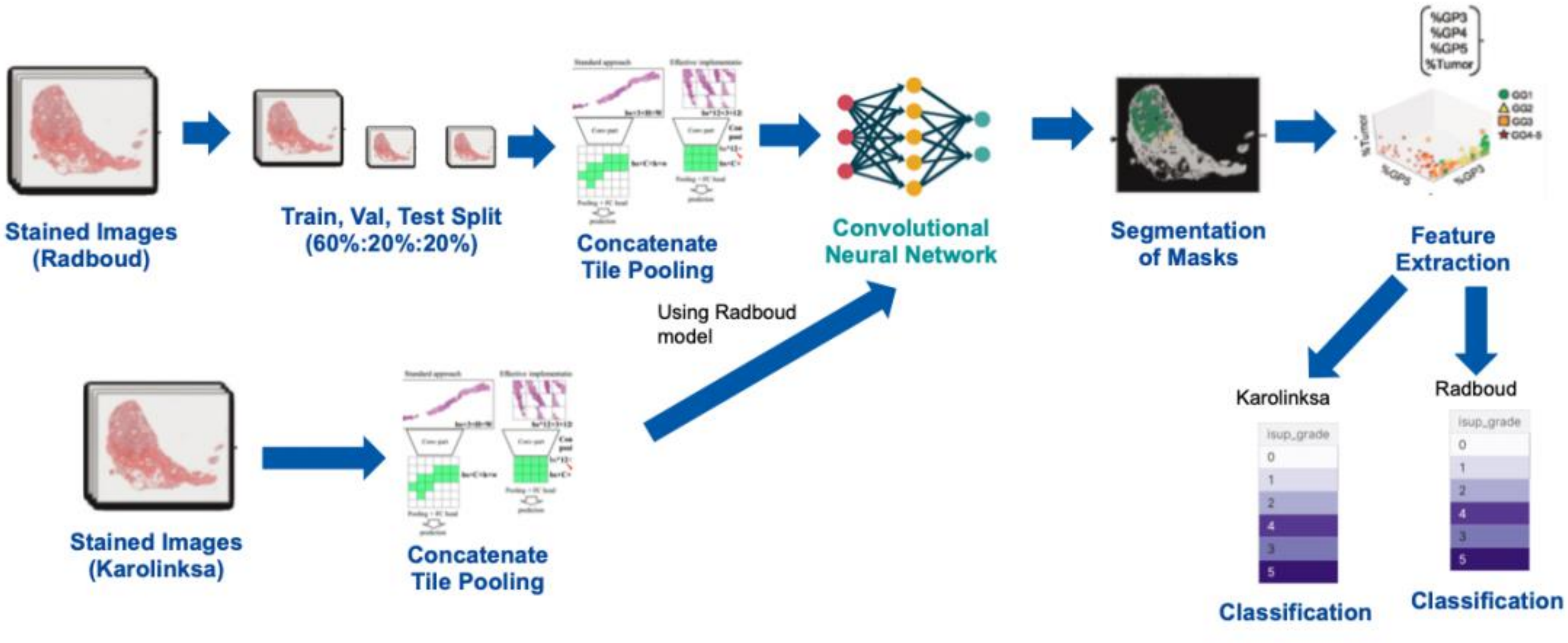
Mask



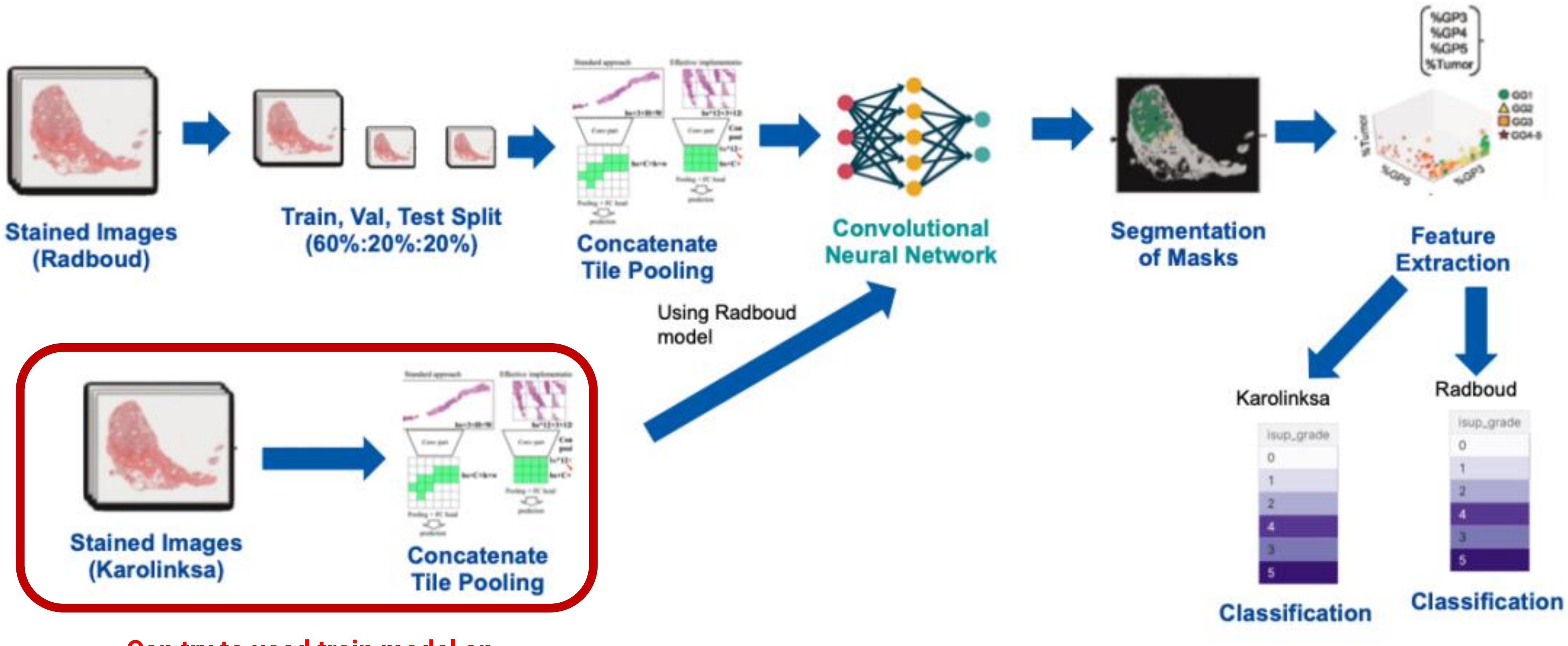
Mask colors and description

- [0]: background (non tissue) or unknown
- [1]: benign tissue (stroma and epithelium combined)
- [2]: cancerous tissue (stroma and epithelium combined)

Two Stage Model Detailed Workflow

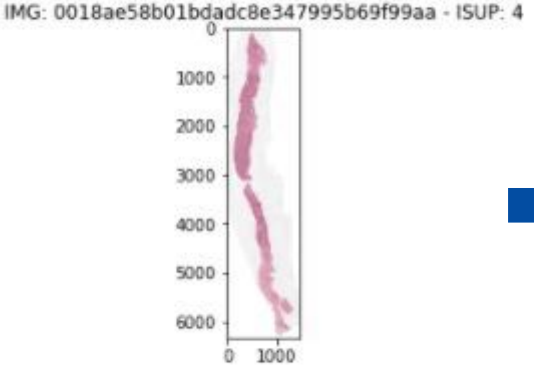


Two Stage Model Detailed Workflow



Can try to used train model on Radboud images to predict on Karolinksa images to see if we get good results

Image Segmentation Workflow



Biopsy Scan Images

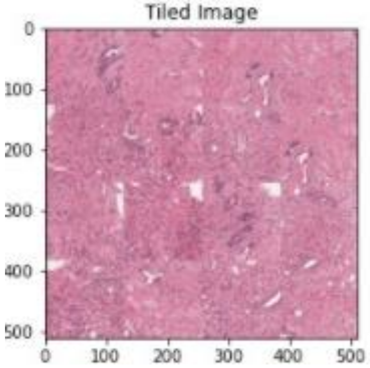
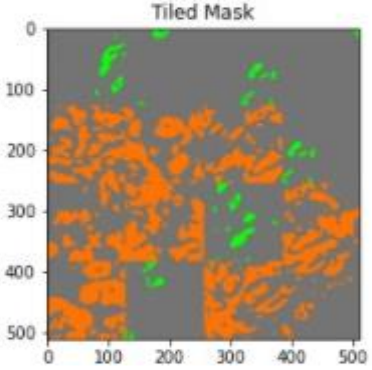
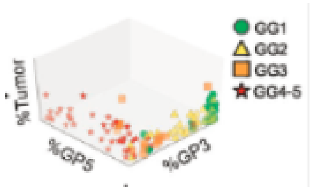


Image Tiling



Predicted Segmentation

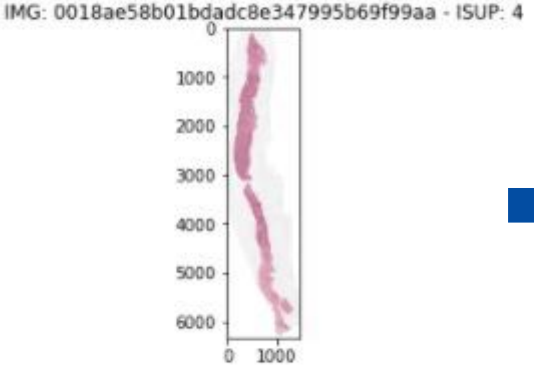


Feature Extraction & Classifier Model

isup_grade
0
1
2
4
3
5

Prediction

Features Extracted



Biopsy Scan Images

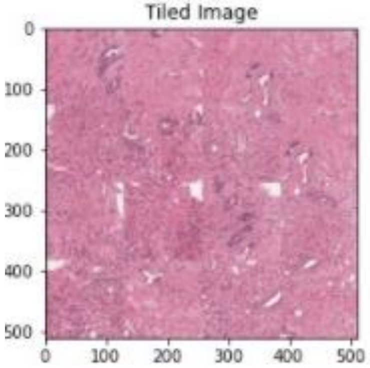
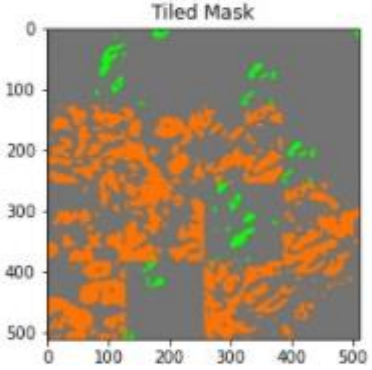
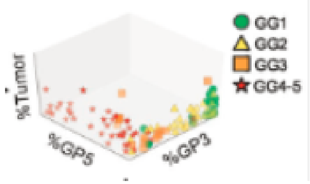


Image Tiling



Predicted Segmentation



Feature Extraction & Classifier Model

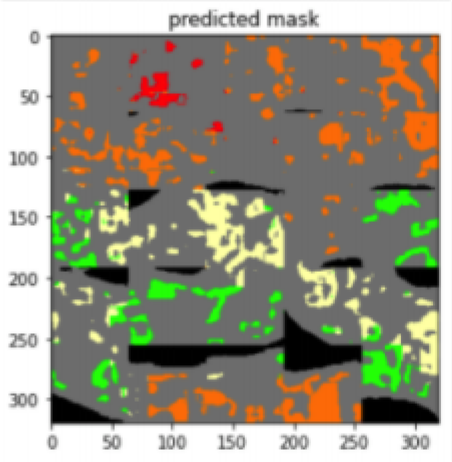
isup_grade
0
1
2
4
3
5

Prediction

Features Extracted

Mask colors and description

- 0: background (non tissue) or unknown
- 1: stroma (connective tissue, non-epithelium tissue)
- 2: healthy (benign) epithelium
- 3: cancerous epithelium (Gleason 3)
- 4: cancerous epithelium (Gleason 4)
- 5: cancerous epithelium (Gleason 5)

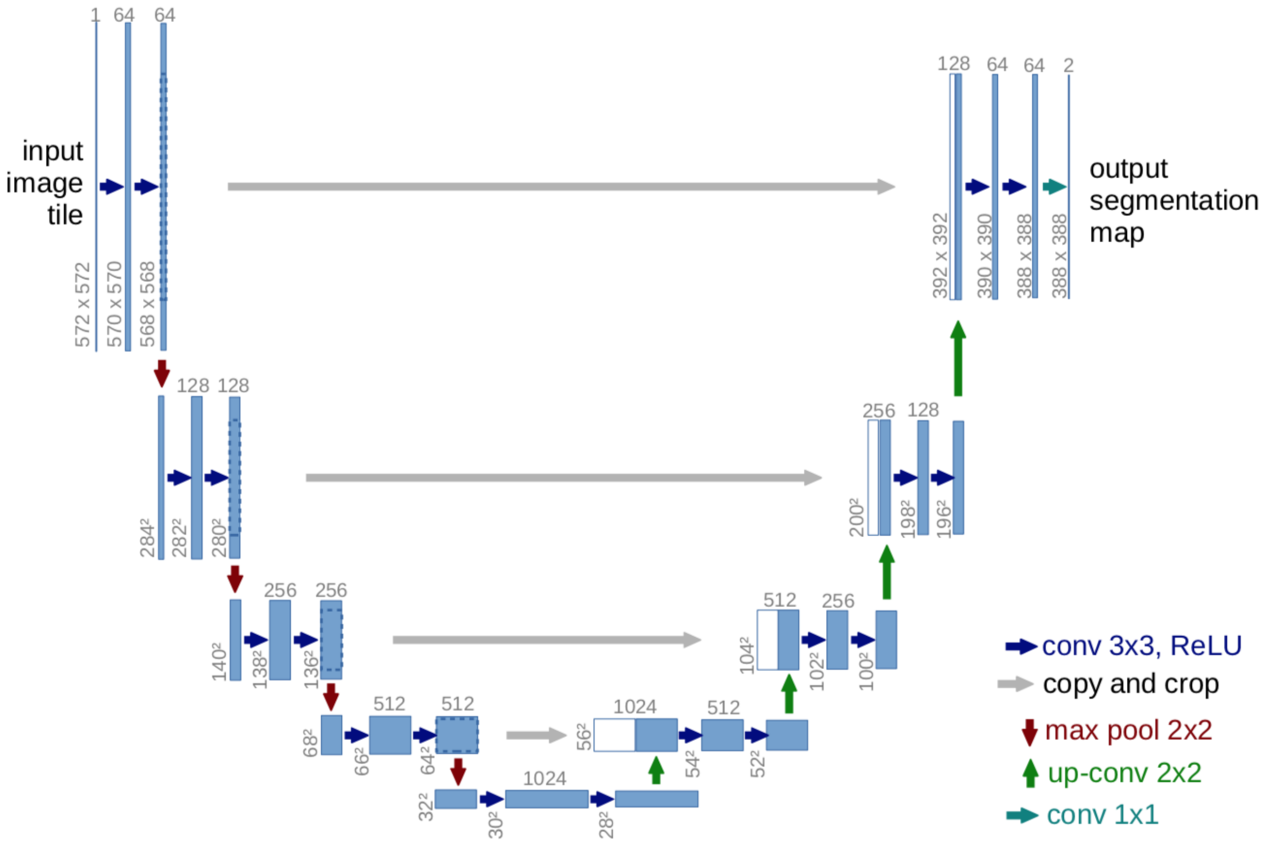


Feature extraction

	features
image_id	0550b23f29085f41b10d165a46ad4371
data_provider	radboud
isup_grade	5
percent_1	0.79431
percent_2	0.0477429
percent_3	0.0590833
percent_4	0.0910798
percent_5	0.00778406
count_1	75716
count_2	4551
count_3	5632
count_4	8682
count_5	742

- **Count and % of mask colors**

Segmentation Model – U-Net Model



- Common segmentation model used in medical imaging
- Convolutional layers with encoders and decoders
- U – shaped architecture

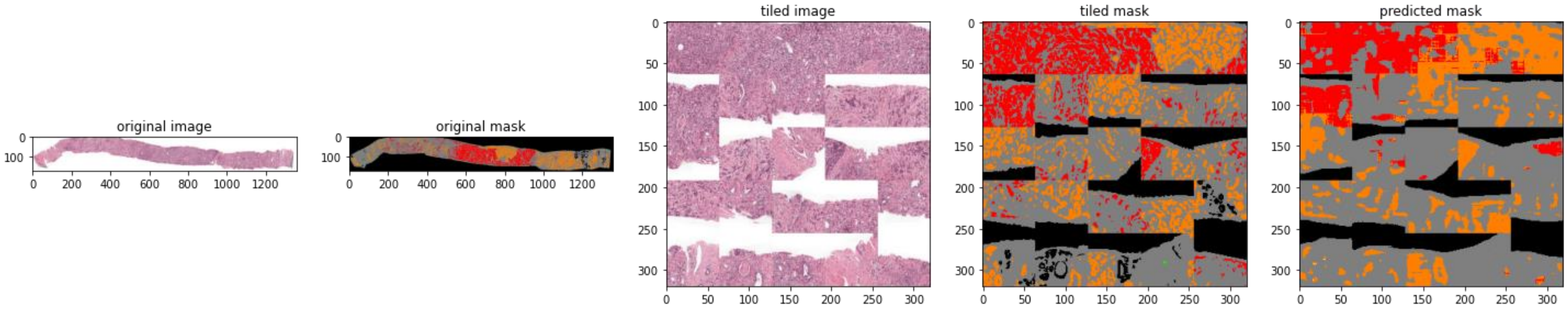
2 Stage Model Results & Findings

Tiles	Pre-trained Weight	Segmentation Model	Classifier Model	Train QWK	Validation QWK	Test QWK	Karolinska QWK
25 x 64 x 64	imagenet	U-Net	XGBOOST	**0.70**	**0.68**	**0.63**	-0.002
25 x 64 x 64	imagenet	U-Net	Logistic Regression	0.64	0.63	0.62	-0.006
25 x 64 x 64	imagenet	U-Net	Random Forest	0.90	0.62	0.59	-0.012
25 x 64 x 64	imagenet	U-Net	KNN Classifier	0.62	0.61	0.61	-0.017
25 x 64 x 64	imagenet	U-Net	Support Vector Classifier	0.54	0.50	0.50	-0.021

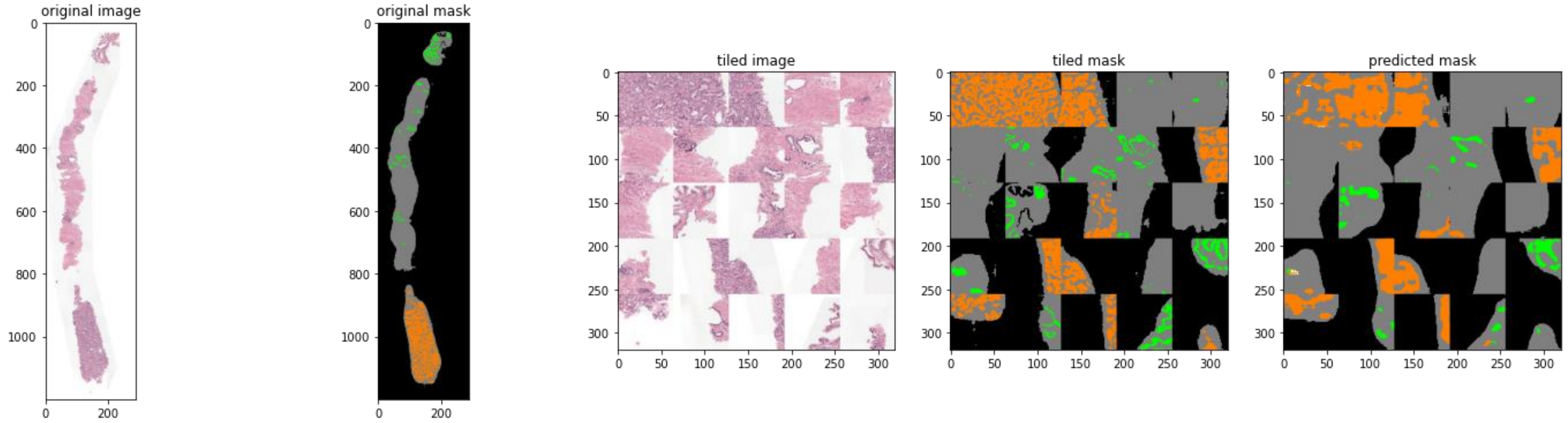
- U-Net with XGBOOST model performs the best
- Scores not as good as 1 stage model. But bear in mind, we have less data to train on since only using images from Radboud
- Using model trained on Radboud images are not useful to predict Karolinska

Examples of Accurate Predictions

img_id: ecad7ce730fc63ac02df4a9da719fa1c ISUP Grade: 5, Predicted ISUP Grade: 5

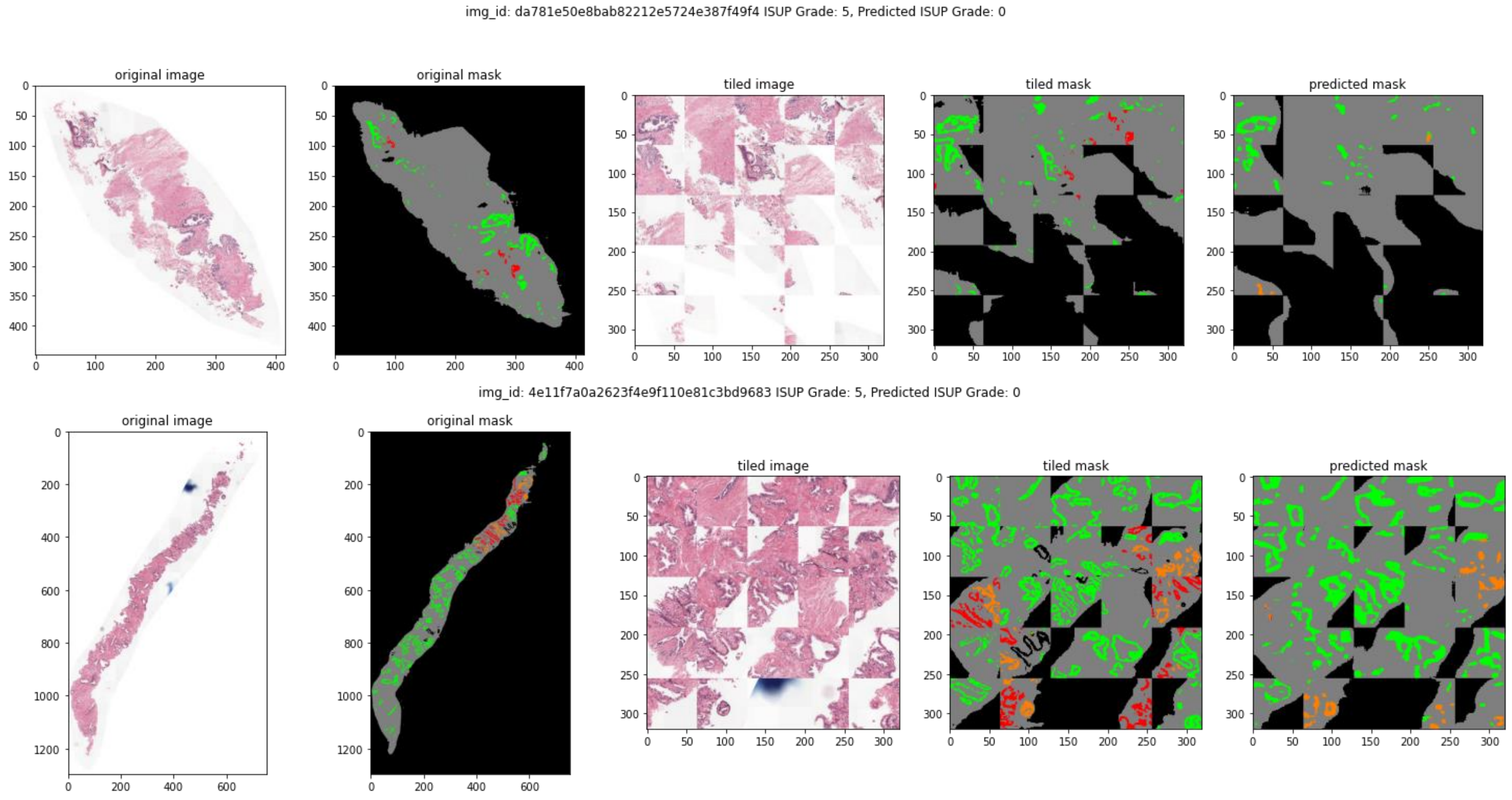


img_id: e17b55972ff36346f98e17f963a057f0 ISUP Grade: 4, Predicted ISUP Grade: 4



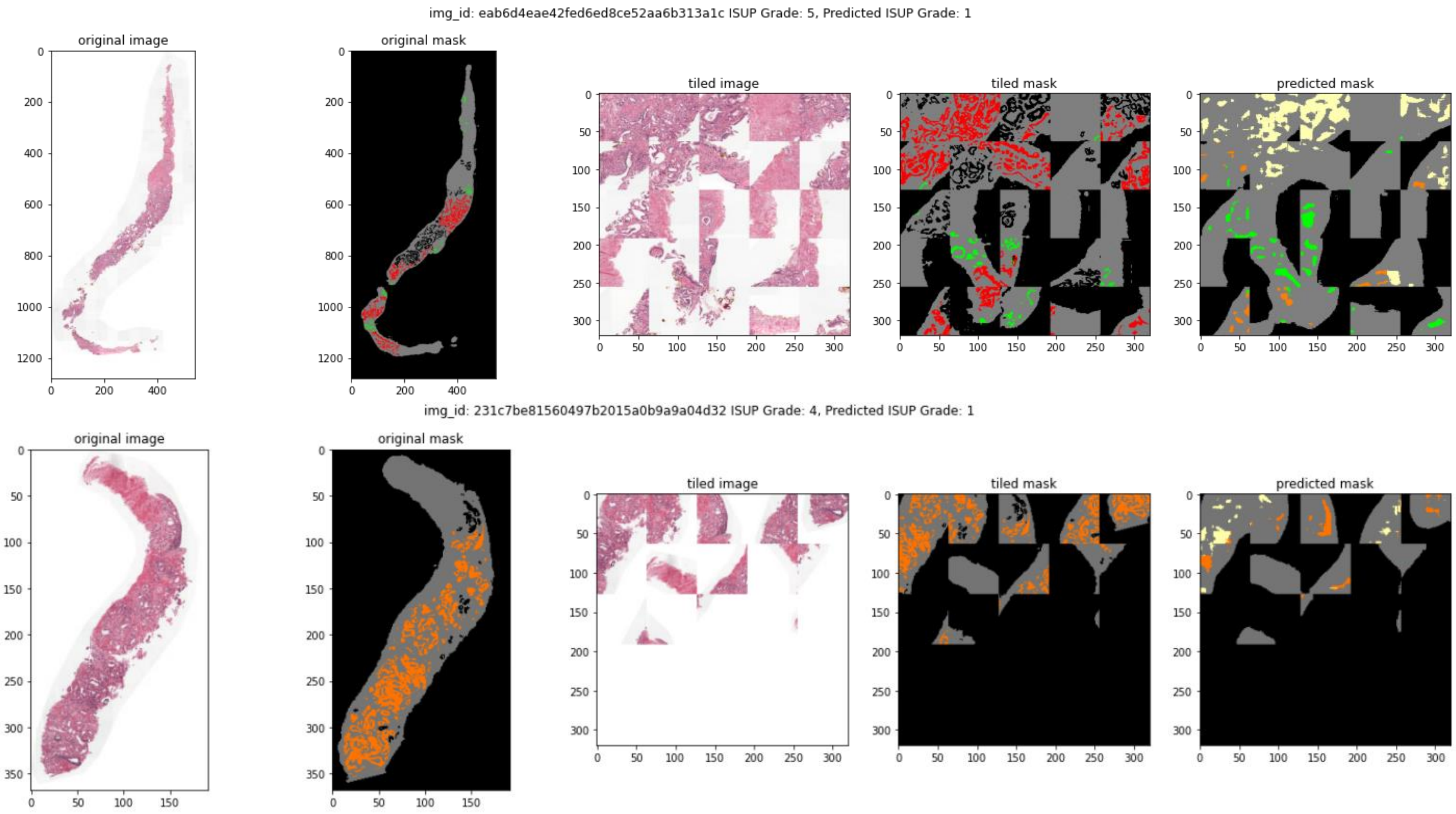
- Model is very accurate when cancerous regions are very obvious from images

Examples of Inaccurate Predictions



- Model is not accurate when very little cancer regions but labelled as highly cancerous
- Perhaps a mistake or mislabel? But it is hard to identify such cases

Examples of Inaccurate Predictions

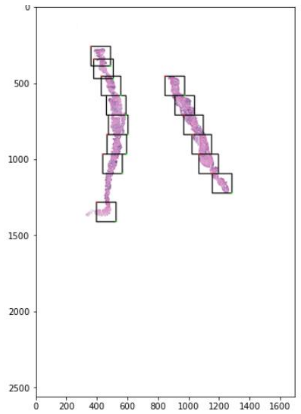


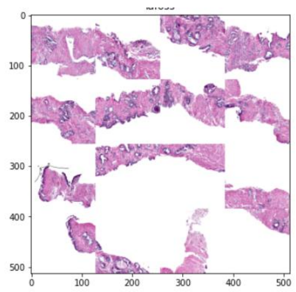
- Model completely fails to predict cancerous masks
- Image resolution is too low, could be fixed with high resolution images and more hardware resources

Further Improvements

Use Higher Resolution Images

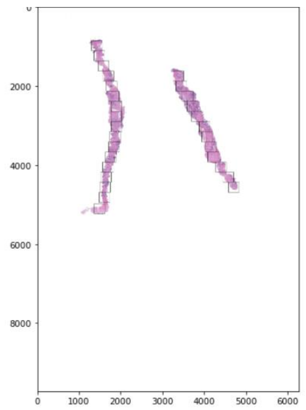
- Higher resolution images will enable larger tiles and more granular information to be captured
- Will make it easier for model to pick up on cancerous areas not detectable by lower resolution

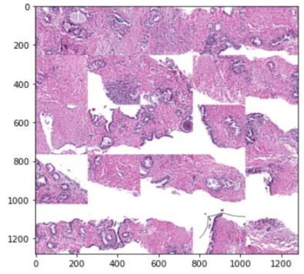




Low Res Image and Smaller Tiles

- 16 tiles, each 128 x 128 on lowest resolution of image
- Tiling will result in more padded white tiles





High Res Image and Larger Tiles

- 25 tiles, each 256 x 256 on highest resolution of image
- Can afford to have more tiles
- Tiling results are most utilized

Further Improvements

Standardize Masks

- Find a way to standardize masks
- Allows usage of all images in dataset for 2 stage model. Has potential to improve 2 stage model
- Apple to apple and fair comparison

More GPU Resources

- With more VRAM in GPU, can train deeper models, reduce training time
- Load and train over high resolution images. Improve both 1 stage and 2 stage models
- Reduced training time and tune hyperparameters more frequently

Conclusion

Tiling Adds Value to Predictions

- Tiling the images improves the predictions compared to not tiling
- Tiling reduces the white spaces in final image and zoom into area of interest
- **Actually part of pathologist job that is being replicated by tiling**

One Stage Model Produced Best Results

- Xception model in one stage modelling produced the best results (QWK 0.7 on test set)
- Uses all images in the dataset

Two Stage Model Has Potential

- Successfully built two stage model, consists of image segmentation and classifier model
- Best model has QWK of 0.63 on test set
- Although not as good as one stage model, but bear in mind only used Radboud images (50% of dataset

Further Improvements

- Use high resolution images – more tiles and larger tiles, more granular
- Standardize masks – use all images in training of 2 stage model
- More GPU resources – reduce training time, deeper models and can load and train over high res images

QnA Session

THANK YOU