Ovarian Cancer Subtype Classification and Outlier Detection

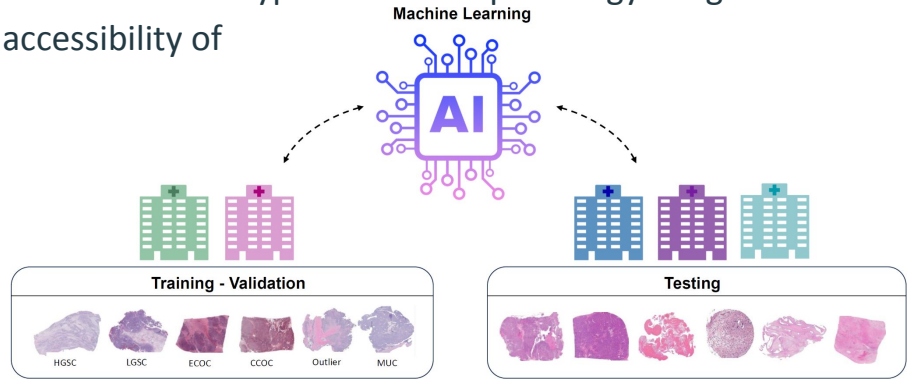
Jonathan DEGOUVE, Victor CHEVREAU, Eva URANKAR, Kathleen GUILLET

TEAM Kakus

- Jonathan DEGOUVE, project configuration
- Victor CHEVREAU, models testing and tuning
- Eva URANKAR, tiling
- Kathleen GUILLET, project configuration

ABOUT THE PROJECT

- Kaggle competition, Sponsored by: University of British Columbia (UBC), BC Cancer, and Partners
- Ovarian Carcinoma: most lethal female reproductive cancer
  - 5+ subtypes
  - Challenging diagnosis needed for subtype specific treatment
- Data:
  - Most extensive and diverse ovarian cancer dataset
  - 550 GB of high resolution images
- Goal: improve the accuracy of identifying ovarian cancer subtypes from histopathology images
- Why : help to enhance the applicability and accessibility of accurate ovarian cancer diagnoses.

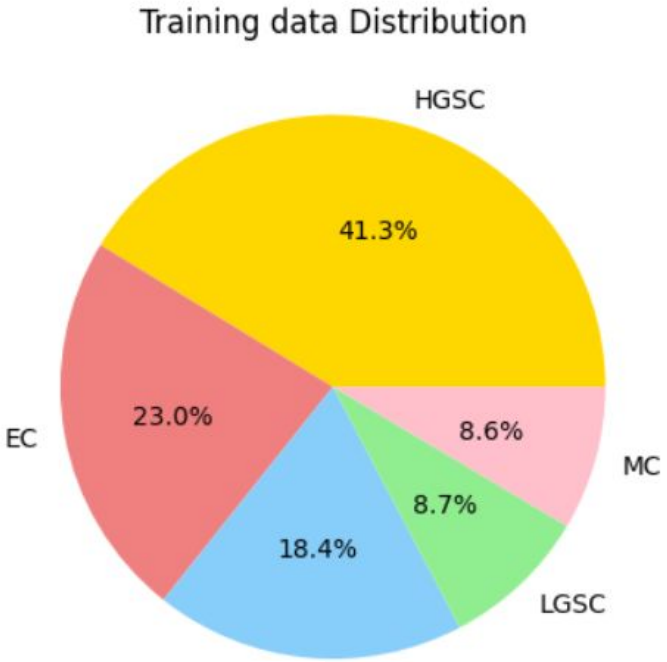


APPROACH

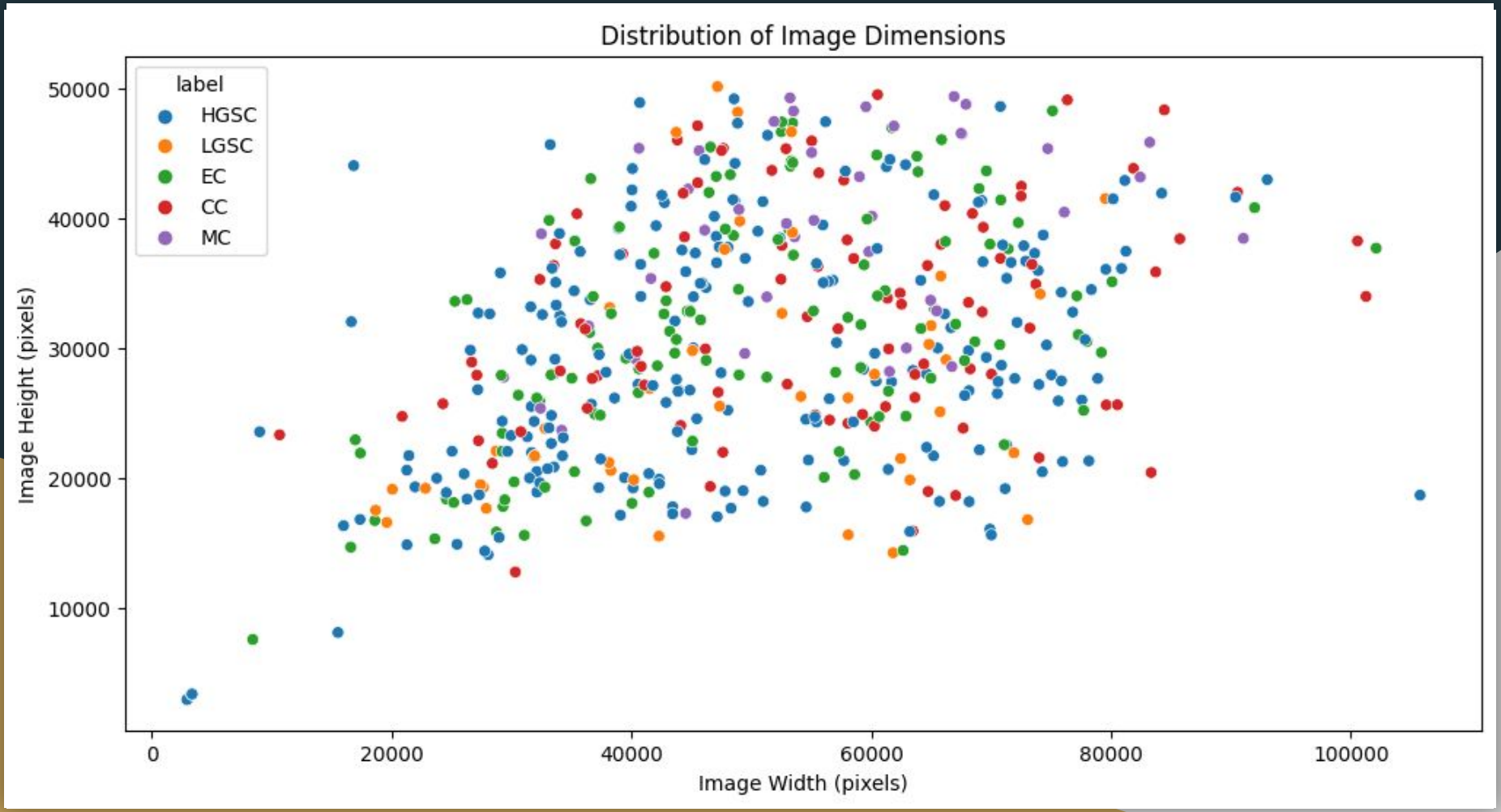
1- Data exploration :

a. Difference between the dataset of thumbnail images and the one with bigger images.
b. Dataset unbalanced.
c. Deciding on the approach : using the thumbnails dataset, applying tiling and ImageDataGenerator for data augmentation.

2- Loading the dataset and opening the images.

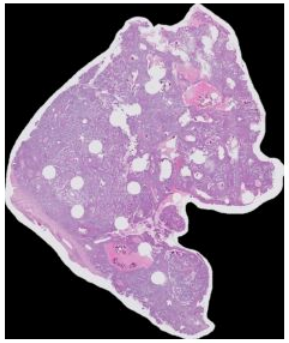



```
Missing Values:
image_id        0
label           0
image_width     0
image_height    0
is_tma          0
dtype: int64
```


Distribution of Image Dimensions
label
HGSC
LGSC
EC
CC
MC
Image Height (pixels)
50000
40000
30000
20000
10000
0
20000
40000
60000
80000
100000
Image Width (pixels)

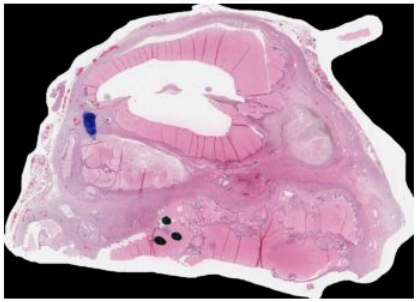
Tiling

Class: EC

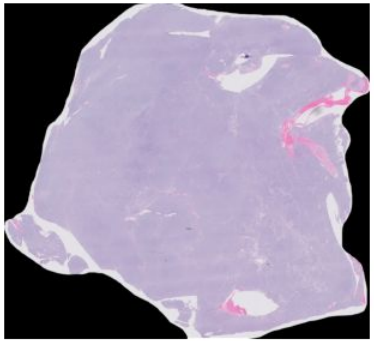


Class: HGSC

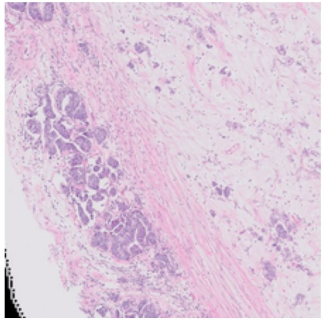
Class: EC



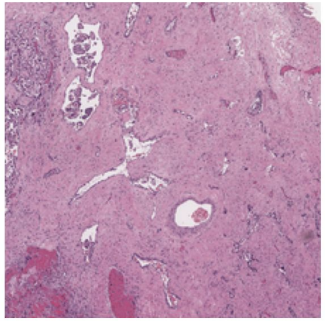
Class: EC



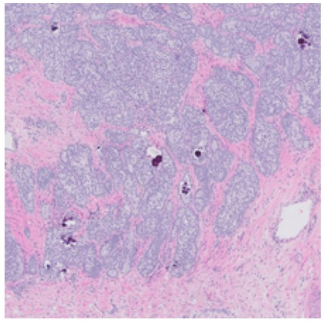
Tile 1



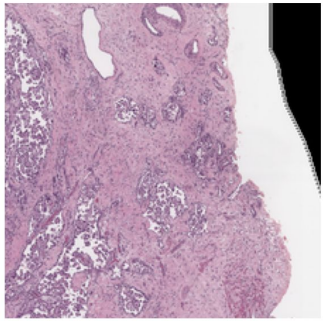
Tile 2



Tile 3



Tile 4



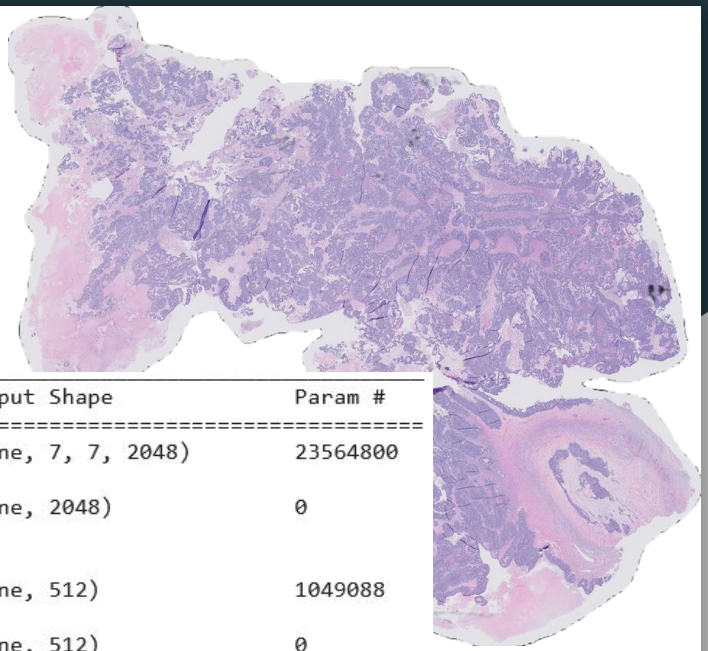
...Approach

3- Training models, testing various pre-trained models (ResNet, EfficientNet) and different hyperparameters.

4- Selecting the best model.

Model

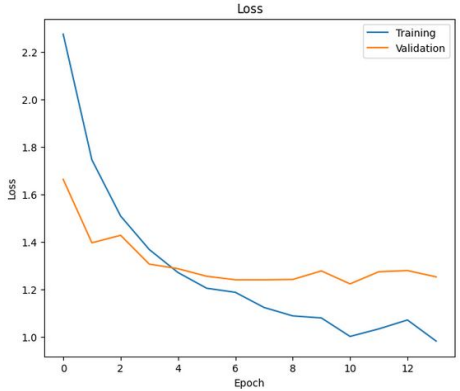
- Normalization
- 80% Train, 20% Test
- Image Augmentation
  - horizontal flip
  - rotation angle (15 degrees)
  - zoom range
  - shift range
- ResNet50v2

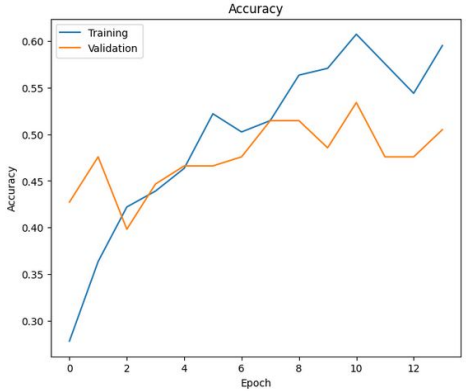


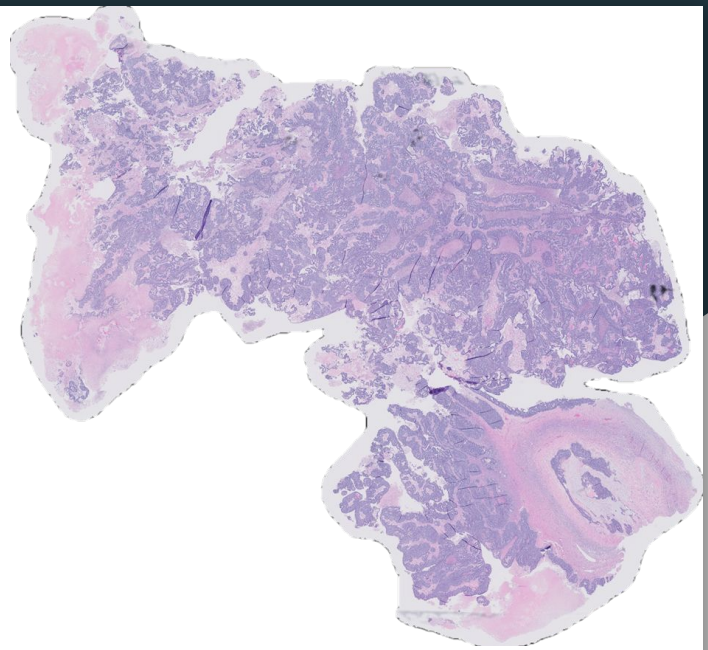
Layer (type)	Output Shape	Param #
resnet50v2 (Functional)	(None, 7, 7, 2048)	23564800
global_average_pooling2d_4 (GlobalAveragePooling2D)	(None, 2048)	0
dense_13 (Dense)	(None, 512)	1049088
dropout_7 (Dropout)	(None, 512)	0
dense_14 (Dense)	(None, 5)	2565

https://www.kaggle.com/code/evaurankar/ml-project-randomforest
https://www.kaggle.com/code/evaurankar/ml-project-resnet

RESULTS







	precision	recall	f1-score	support
0	0.36	0.31	0.33	13
1	0.43	0.33	0.38	27
2	0.56	0.76	0.64	46
3	1.00	0.25	0.40	8
4	0.83	0.56	0.67	9

accuracy			0.53	103
macro avg	0.64	0.44	0.48	103
weighted avg	0.56	0.53	0.52	103

LESSONS

- Rightfully choose the project according to yours skills.
- The difficulty of treating images, especially when they are substantials and their characteristics are sparse.
- The computational resources needed to treat images.
- The power of working in team, where the range of skills can be wide.

IMPROVEMENT

- Training on the dataset of the big images.
- Detection of outliers
- Use annotated masks
- Improve the tiling

Thank you for listening