Deep Learning Workshop: Breast Cancer Classification

HASHEMI Seyedali CHAHbAOUI Mohammed KARATAS Musab

Introduction

- build a machine learning model that can detect **Invasive Ductal Carcinoma (IDC)** in breast tissue images.

- IDC is a type of breast cancer most common type of breast cancer (about 80%).

- patches are taken from **162 whole-slide images** scanned at 40x magnification.

Problem to solve

- Classification: is the 50x50 patch benign or malignant ?
    - Bigger scope: detect patients with malignant tumor.

- Challenges:

- **Imbalanced data** (71% benign vs 29% malignant)
- **High accuracy needed** for medical use (no room for risk).
- Using the correct metrics to answer the problematic.

Data - Characteristics

- Dataset of 277 524 images

- Each image is **50×50 pixels** in size

- Images are extracted from **162 whole-slide images**

- Slides were scanned at **40x magnification**

- Each image belongs to one of two classes:

  - **Class 0**: Non present IDC (benign): 198,738 images.
  - **Class 1**: Present IDC (malignant): 78,786 images.

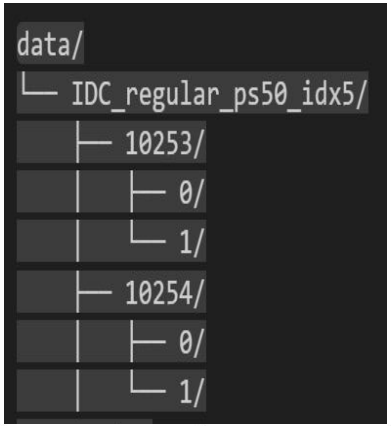
Data - Pre-processing

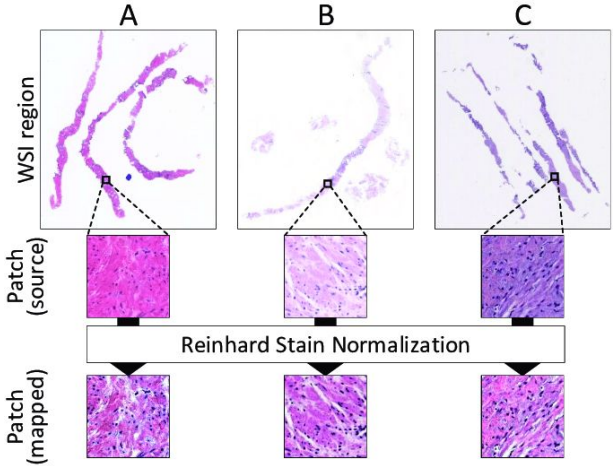
- Images stored in subfolders as shown :
  - Reading the and storing the image's paths in a DataFrame.
  - path - patient_ID - label (class)
    - To make Patient level splits.

- Reinhard normalisation :

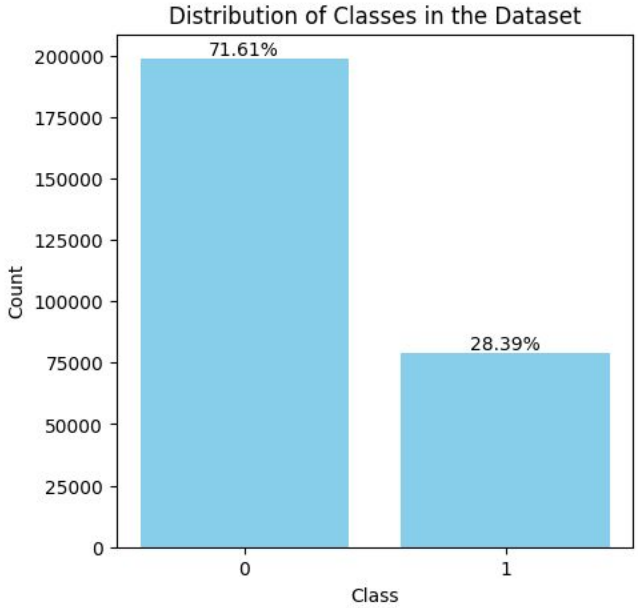
  - Good for this type of medical images for better distribution.

  - Random augmentation to help the smaller class.



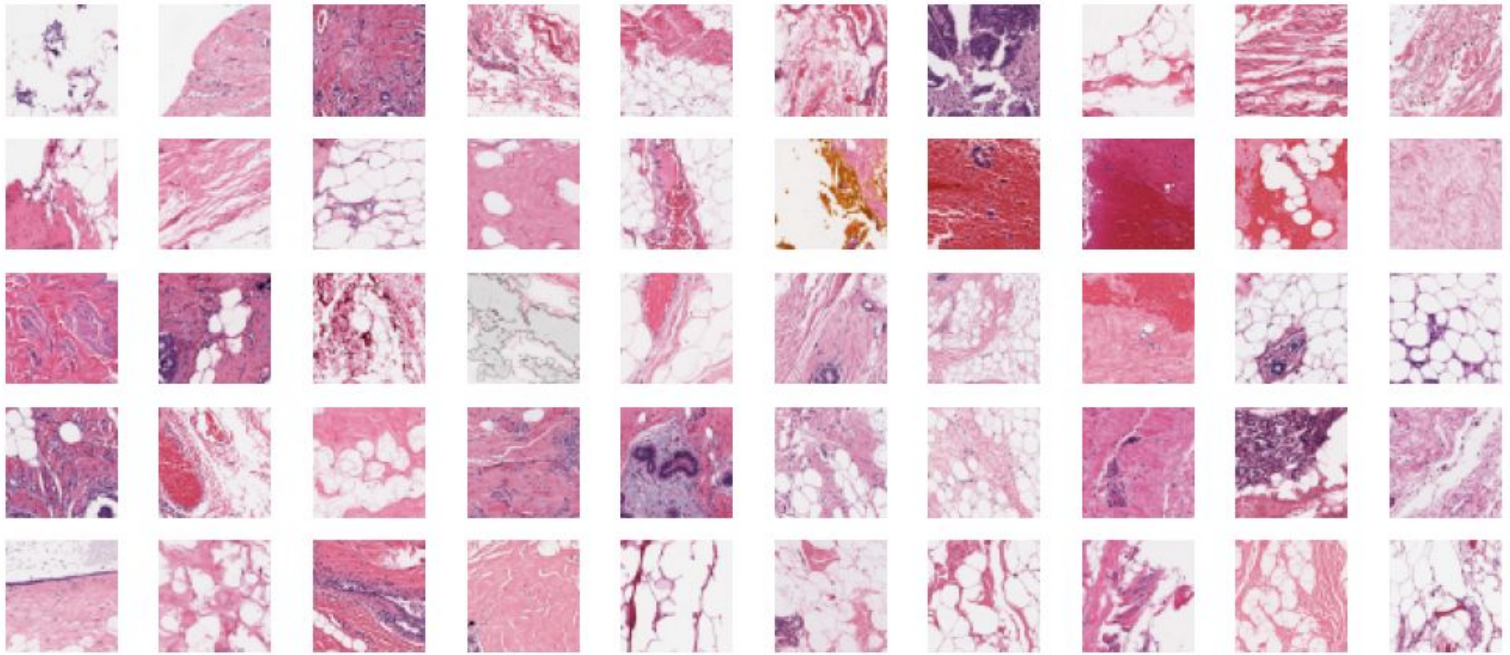


Data - Visualisation: class distribution



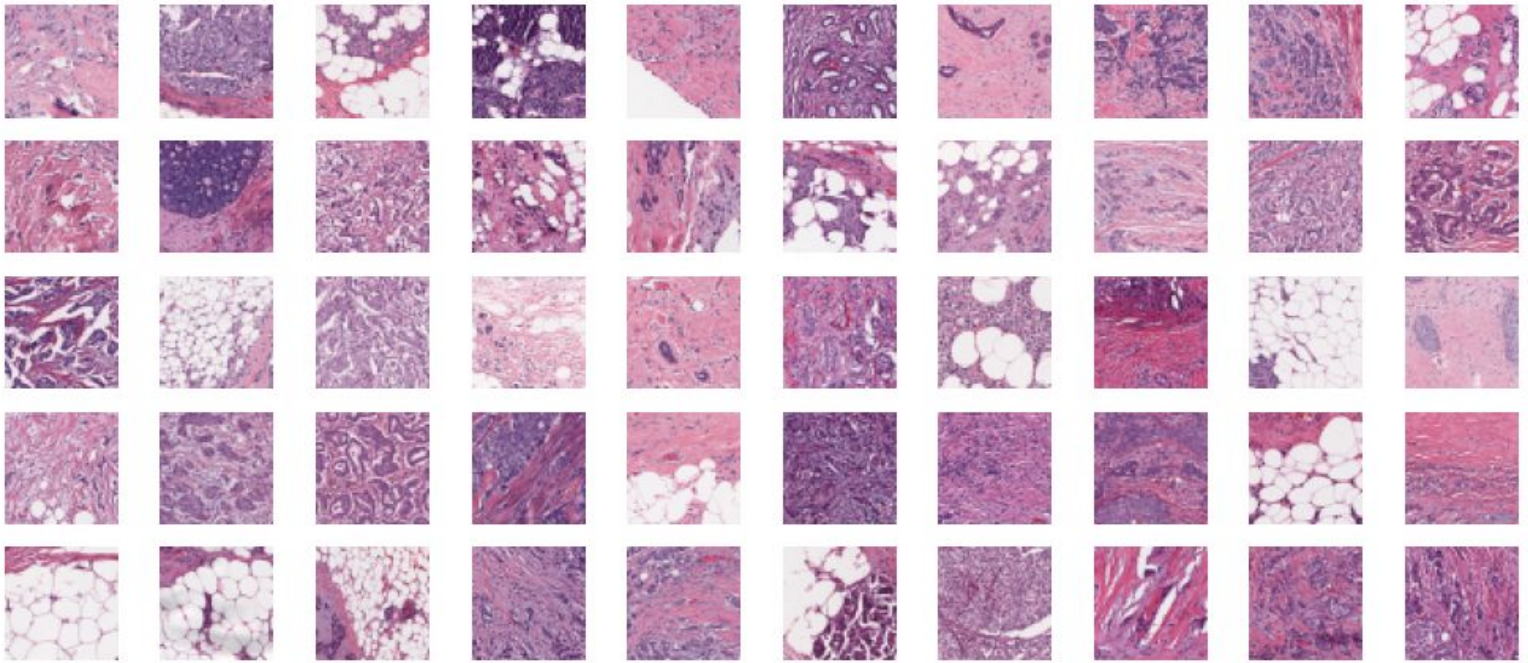
Data - Visualisation: Class0 images

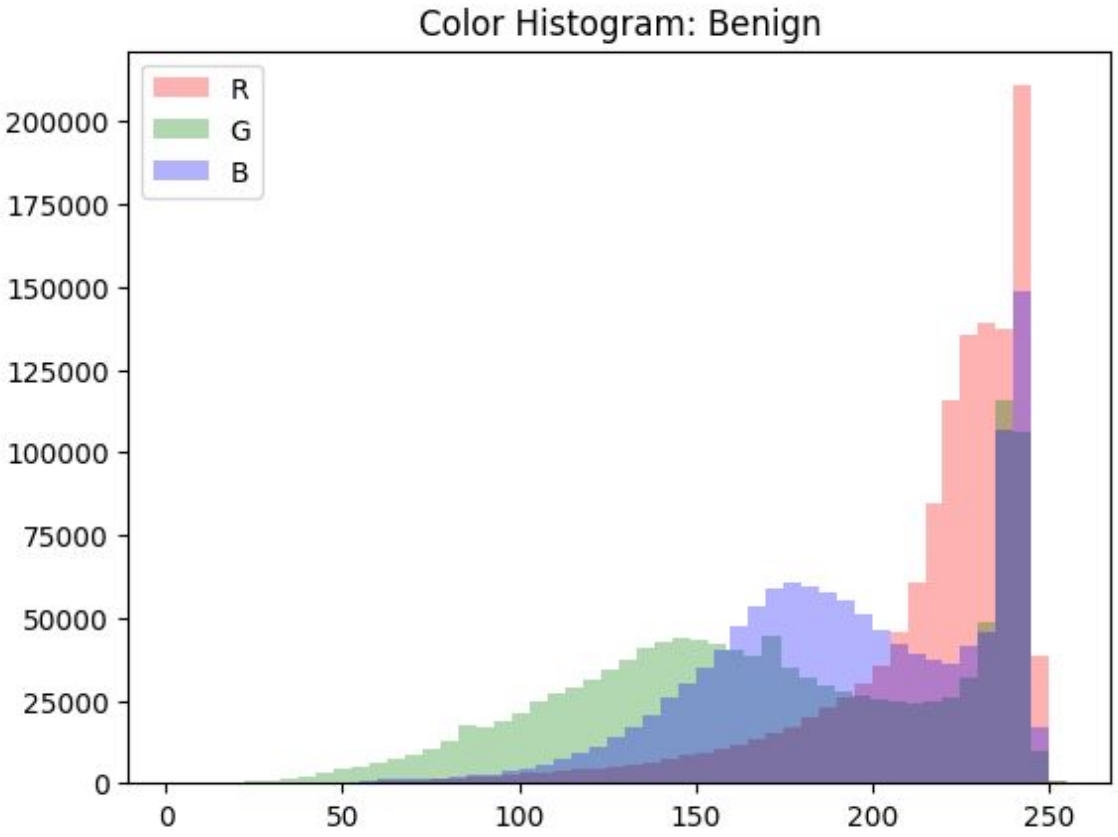
Benign Samples (5x10)

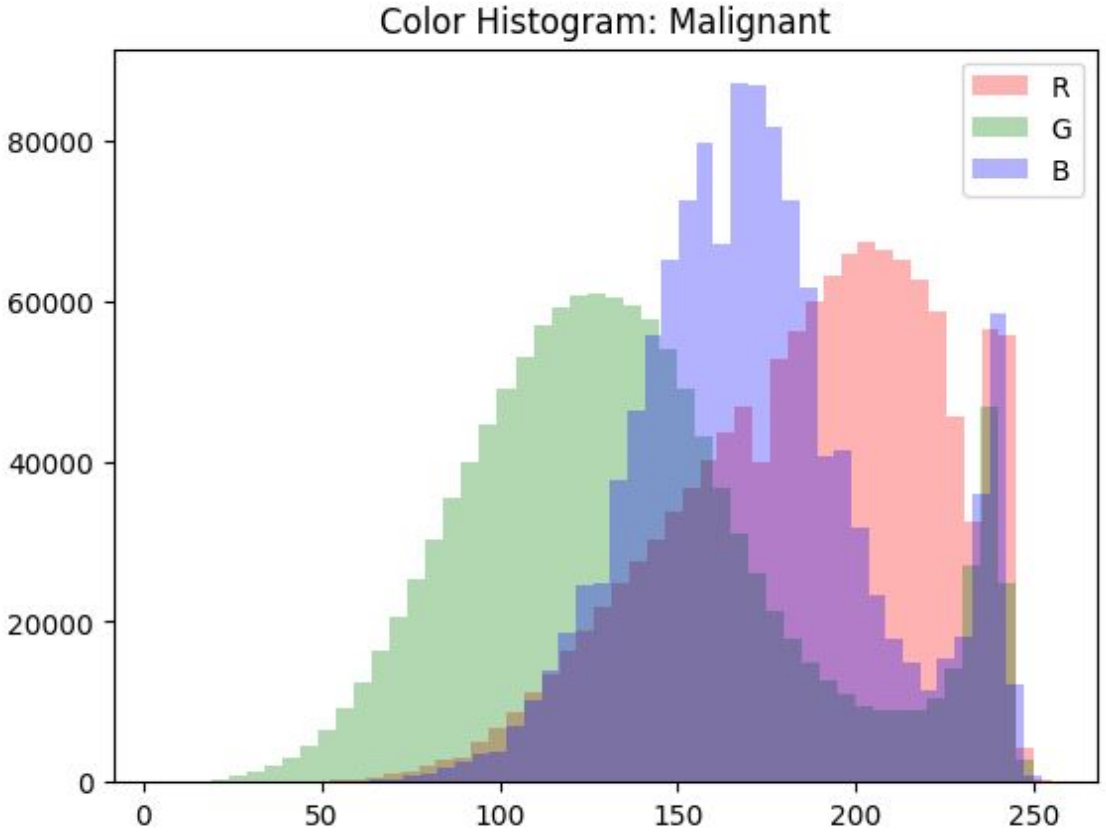


Data - Visualisation: Class1 images

Malignant Samples (5x10)

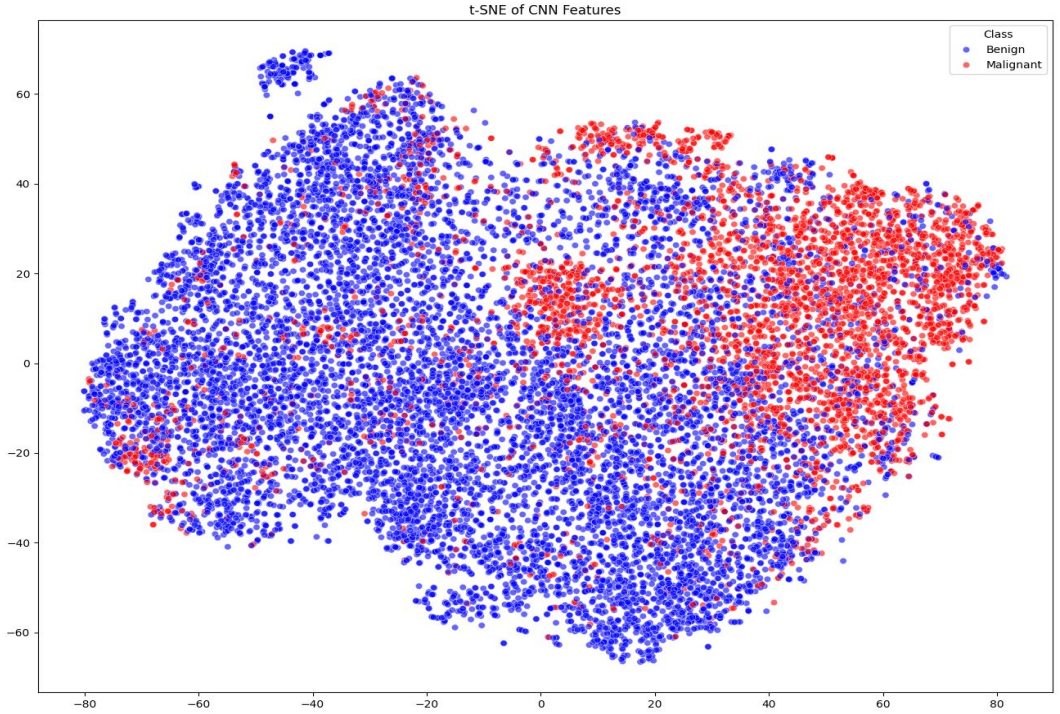



Color Histogram: Benign
R
G
B
200000
175000
150000
125000
100000
75000
50000
25000
0
0
50
100
150
200
250


Color Histogram: Malignant
R
G
B
80000
60000
40000
20000
0
0
50
100
150
200
250

Plotting the images embedding:

- feature extraction :
  - Base of the pretrained ResNet50.
- dimension reduction :
  - T-sne



Modeling: architecture

Dealing with images: CNN is a go to.

- 1st model : fin-tunning a pretrained model : Slow !
- 2nd model :
  - 3 Convolutional blocks : 2Conv2D layers + MaxPooling2D + Dropout with 20%
  - Loss function : binary crossEntropy.
  - Metrics :
    - Recall - Precision - Auc.

Modeling: training

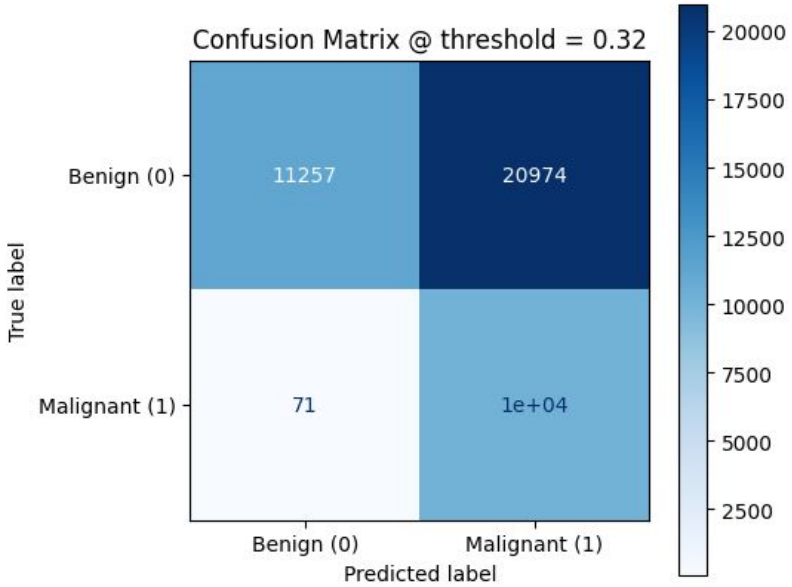
- Data splitting: Divided 162 patients into 70 % train, 15 % validation, 15 %.
- Balanced sampling: Computed class weights to up-weight the minority (malignant) class during training.
- Callbacks: Used ReduceLROnPlateau and EarlyStopping on validation recall to optimize sensitivity.

Modeling: metrics

Recall - Precision - Auc :

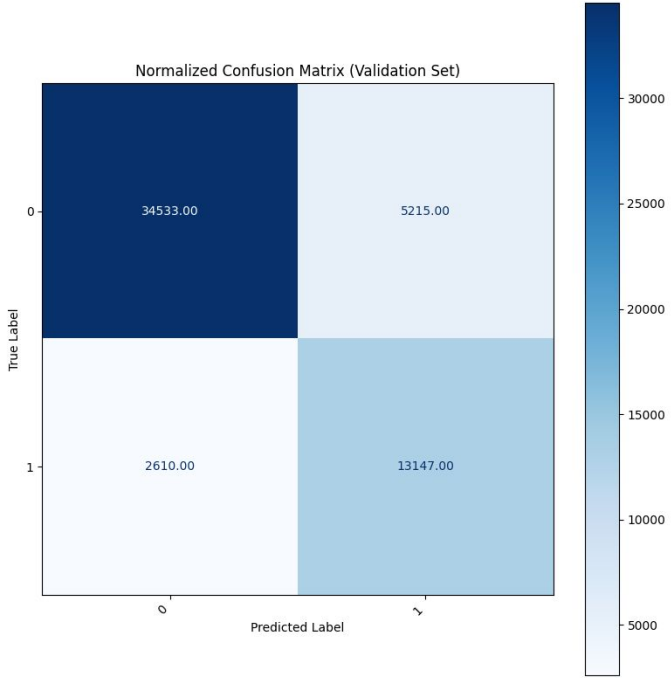
- Answer questions of the medical domain.
- Results :
  - precision: 0.3676 recall: 0.9691
  - precision: 0.3958 recall: 0.9409
  - precision: 0.3941 recall: 0.9453

Averaged precision : 0,3858 recall : 0.9517



Modeling: what it can do.

- In practice, this model can safely filter out the majority of benign images (low false negatives)
- still cutting down the number of tiles a doctor must review.
    - Lets pathologists and histotechnologists focus their attention only where it's needed.
- Precision is modest (≈ 1/3), that's the trade-off to hit the 99 %+ recall target.



Discussion

- optimisation with segmentation to localise tumoral zones.

    - the images describe only a segment of the whole image.

- Couldn’t à 90% of recall be sufficient ? And gain précision as result ?

Thank you for your attention