Summary: A Dataset for Breast Cancer Histopathological Image Classification

IFA WiSe 22/23 - Assignment 3 (Data Science)

Group 7

2022-11-06

Data & Background

Dababase: BreakHis, contains 7909 microscopic images collected from 82 patients with maginificatio level: 40x, 100x, 200x, 400x.

Reference set: 2480 benign, and 5429 malignant.

Idea: Classification performance of a baseline pattern recognition system ← features & extractors.

1. *Preparation steps*: fixation, dehydration, cleaning, infiltration, embedding, and trimming.
2. Mounting on slides, sections (~ $3\mu m$) are cut using a microtome, and then, covered with a glass coverslip.

Staining → hematoxyline, and eosin (HE) ← standard paraffin process

Images ← RGB TrueColor (24-bit, 8 bits/color channel)

Pixel size ← physical pixel size using relay lens magnification (ROI : region of interest)

Benign has A, F, PT, and TA while Malignant has DC, LC, MC, PC.

Feature extraction

Texture representation was used instead of explicit segmentation (shape, glandular shape etc.).

1. *Local Binary Patterns (LBP)*: computing the distribution of binary patterns in the circular neighborhood of each pixel.
2. *Completed LBP*: variant of local binary pattern based on three components:

- Center pixel*
- Sign
- Magnitude

3. **Local Phase Quantization**: quantized phase information of discrete fourier transform (DFT).
4. **GLCM**: used to characterize texture images.
5. **PFTAS**: parameter-free threshold adjacency statistics.
6. **ORB**: keypoint detection ~ computational cost + matching performance.

*coded by binary code ≥ average query level of the whole image.

Classifiers

1-NN: discriminant power of features

QDA: normally distributed linear discriminant analysis

SVM: classification and regression

RF: decsion trees

Data partition: Training (70 %) | Testing (30 %)

Decision is patient wise (at the patient level)

Let N_p be cancer images of patient P, if N_{rec} images are correctly classified, the patient score is

$$S = \frac{N_{rec}}{N_p}$$

Global recognition rate is

$$GR = \sum \frac{S}{N}$$

If the the proportion of positive to negative instances changed in a test set, ROC curves will not change.

Results

PFTAS outperformed other classifiers with *85 % accuracy* for 200x magnification samples.

Root of confusion

When a benign tumor is classified as malignant (high FPR) $\Leftarrow$ one of the benign tumor present in the dataset shares similar properties with a malignant tumour

Reference

F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte, “A Dataset for Breast Cancer Histopathological Image Classification,” in IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455-1462, July 2016, doi: 10.1109/TBME.2015.2496264.