EE417 - Assigment #1
Biomedical image analysis
Breast cancer severity assessment

March 4th, 2025

1 Introduction

Breast cancer is one of the most common forms of cancer, and one of the leading causes of death among women.

In order to establish a treatment, medical professionals need to determine-measure how aggressive the case under study is. If it is an aggressive case, then a more intensive treatment (typically chemotherapy) is applied. Otherwise, a treatment with fewer unwanted side-effects might be preferred. But how can one measure the aggressiveness of cancer?

Cancer is by definition the case where body cells are divided without control. Consequently, one way of achieving the aforementioned goal, is to acquire a small amount of tissue from the patient (a.k.a. biopsy) and count under a microscope the number of cell divisions (a.k.a. mitosis) per millimetre squared (mm^2). The higher the resulting number, the more aggressive the cancer case.

As you can guess, this counting process is tedious, takes time, requires expertise, and expensive equipment. The alternative is to provide the acquired microscopy images to a software, that will do the counting automagically. This is the objective of this assignment.

2 Dataset

You are provided[1] with the images of 5 patients (A00, A01, A02, A03 and A04). You have a varying number of images per patient. The images are in color (artifically colored via chemical substances to improve visibility), and of size 2084×2084 pixels (about 4 Megapixels) in BMP format. Each image is also accompanied by a CSV file of the same name containing the coordinates of the mitosis instances present in them (one mitosis per row) (as visualized in Fig. 1). The data of patient A00 will be used for testing, the data of patient

[1] https://drive.google.com/file/d/1ZGzmx8UaqMaVJo0Rq1q6lBwXSGnLIFUp

A02 will be used for validation, and the data of the remaining 3 patients will be used for training/development.

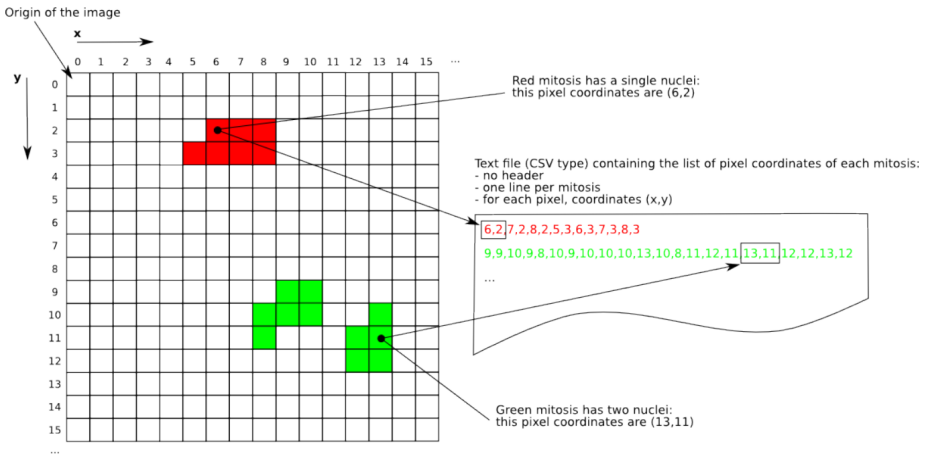


Figure 1: Organisation of a text file containing the list of pixel coordinates for the mitosis.

3 Objective and evaluation

The objective is to develop an image analysis method, that will enable the automatic detection (and hence counting) of the mitosis cases for a given breast tissue image. Performance will be measured in terms of two metrics. The first will be the overlap between the estimated mitosis detection ($E \subseteq \mathbb{Z}^2$ as a set of coordinates) and the actual mitosis instance ($A \subseteq \mathbb{Z}^2$ as a set of coordinates) denoted as *intersection over union* (IoU):

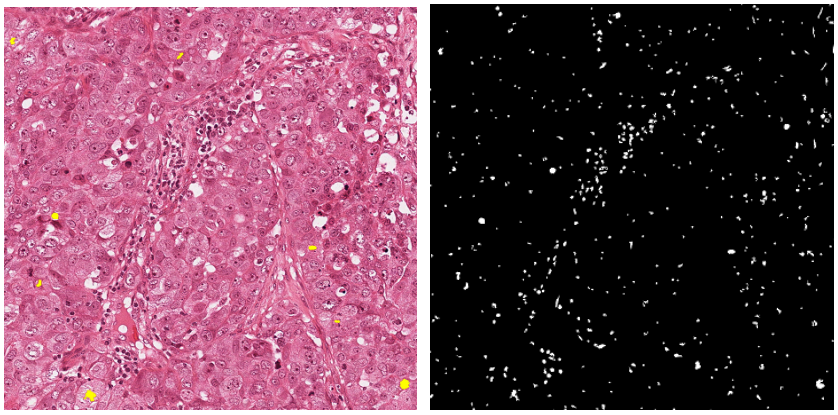
$$IoU = \frac{|A \cap E|}{|A \cup E|} \in [0, 1] \tag{1}$$

the higher the IoU, the better. You can take the mean IoU across all detections with your test images, so as to obtain a single value to quantify performance. And the second will be the F1-score, due to the imbalance between the classes.

4 Task

4.1 Stage 1 - 5 points

I recommend starting with visualizing where the mitosis cases take place within a given image; for this you'll need to color the coordinates contained in the csv file, and display them, so as to see which are mitosis instances, and which not; since it is not at all straightforward (Fig. 2a).



(a) Colored sample (b) Background separation

Figure 2: Colored sample and background separation example (your output doesn't have to be exactly like it; it's only provided as a guideline.)

4.2 Stage 2 - 30 points

The images are complex, with many cells, fat tissue and other histological content. I recommend separating the background from the foreground. Ideally the cells should be set to white, and the background to black. The cells and especially their nuclei are mostly darker w.r.t. their surrounding. Mathematical morphology can help in this regard as it is only dependent on relative intensities. After which you should extract all the connected components. Take precautions against noise, if necessary. Ideally every cell should consist of one connected component (Fig. 2b).

4.3 Stage 3 - 30 points

Now, you need to figure out a way to describe these cells; i.e. convert each of them into a numerical feature vector of a fixed length. You can exploit their color, their texture and even their shape. You have complete freedom in this regard. You can use any approach you consider suitable (except for deep learning, we are saving this for later on). You can use global as well as local descriptors. If you use global descriptors you can get at most 15 points from this question. If you use the bag of visual words paradigm you get 30 points.

4.4 Stage 4 - 35 points

Now, it's time to put all pieces together. Use your training images to a) preprocess them if necessary b) extract the cells's coordinates c) compute the feature vector of every cell d) then use either a Support Vector Machine or a Random Forest to train a classifier that will learn to perform binary classification for a

given cell's feature vector: as either mitosis or not mitosis.

Use your validation images to fine tune the parameters of your solution (preprocessing, background separation, description, classifier hyperparameters). Report in a table how each setting affected performance in terms of IoU in the validation set. Once you have established the optimal pipeline, then you are ready to test it with your test images. If your method is robust, then the performance between validation and test images will be relatively similar. Report your final result. You are free to explore multiple approaches so as to determine the optimal overall strategy.

If your method works with a decent mean IoU across the test set, then congratulations, you have just developed a piece of software with a market value of tens of thousands of USD as of 2025 [2].

5 What to submit

You are expected to submit

- the notebook with your code and
- a PDF report prepared in LaTeX of at most 2 pages (IEEE conference format[3])

explaining your pipeline and the results that you have obtained.

6 Rules

In addition to the assignment and plagiarism rules outlined in the course syllabus:

- the notebook must be ready to run, with no errors. The teaching assistant will not solve your compilation/syntax/... errors. If it doesn't run, it will get a grade of ZERO.
- in order to get the grades of each stage, you need to provide both working code, and the respective explanations of how you solved it in your report.

[2] https://www.aiosyn.com/news/aiosyn-mitosis-breast-becomes-the-first-ai-powered-mitosis-detection-solution-to-achieve-ce-mark-certification-under-ivdr/

[3] https://www.ieee.org/conferences/publishing/templates.html