

Segmentation and Classification of glioma cells

Tosato Lucrezia

Abstract— Biomedical imaging is an important source of information in cancer research, it deals with characterizations of cancer morphology, at initiation, during progression, and in response to treatment. It provides complementary information to that obtained from genomics and clinical data. In this paper we will try to make an overview regarding segmentation, classification, prediction and analysis of gliomas in the context of biomedical imaging. Regarding the first part we will focus on cell segmentation in multiplexed images, for this purpose new data augmentation techniques developed to improve segmentation will be presented, followed by segmentation techniques using deep learning, newly developed pipelines and morphological pre-processing techniques. In the second part a list of classification techniques using SVM and deep learning and their description will be presented. Eventually new applications concerning the study of the stem niche, and the structure of the tissue surrounding cancer will be introduced.

INTRODUCTION

Cancer is one of the most unresolved problems in the world, indeed, in the paper Sung H. (2021), «Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries», is stated that one in 5 people on average will develop cancer in their lifetime.

In particular, brain cancer is one of the most deadly, with a very low survival rate for both sexes, which is why it is widely studied and investigated in the scientific field in different areas.

Besides arousing medical interest, imaging also seeks to make its contribution in this field to find more innovative solutions and expand research in a completely non-intrusive manner.

In particular, interest in the field is aimed at understanding what can trigger cancer, at identifying and predicting tumor progression, in order to prevent and anticipate possible worsening.

Biomedical imaging has therefore evolved into a critically important tool for doctors and researchers to extract, analyze and interpret complicated medical situations and to be able to make an accurate diagnosis and prognosis.

Therefore, measuring objects and their intensities in images is a key step in many quantitative tissue image analysis processes.

Tissues are very complex assemblages of different cells that interact with each other and with other membranes and connective tissues. When a cancer arises, disruption of regulatory mechanisms intrinsic to the cells and the environment occurs, causing morphological and structural changes on spatial scales ranging from sub-cellular organelles to the entire tissue.

Again, classical histology provides insufficient molecular information to identify precise sub-types of cells, evolution of mechanisms and characterization of malignant genes.

Particularly, profiling the molecular features associated with the morphological landscape of tissues is critical for investigating the structural and spatial patterns that underlie the

biological function of tissues.

Fortunately, however, in recent years, several methods for multiplexed tissue imaging have been implemented.

SEGMENTATION

Single-cell analysis of imaging-acquired data requires segmentation, a technique that assigns class labels to an image to subdivide the image, after which quantification by marker is applied on a per-cell or per-organelle basis.

Segmentation plays an important role in computer-aided analysis of pathological images by automating the process of delineating nuclei and cell locations and regions. To achieve this purpose several methods have been reported in the literature. These methods can be classified into three categories:

1. The intensity-based approach: the most common approach to separate the foreground from the background. This method relies on discontinuities in the gray level intensity of pixels to identify the boundaries of nuclei.
2. The mathematical morphology-based approach: is a technique for processing geometrical structures based on set theory. It provides a powerful tool for extracting essential shape characteristics and eliminating irrelevancies.
3. The deformable model approach: this approach outline object contours using closed parametric curves that can deform under the influence of internal and external forces. The internal force will keep the contour smooth during deformation and the external force will move the contour to a target shape. To find the contour of an object in an image, a closed curve must be placed near the desired boundary and then begin an iterative relaxation process.

A great deal of work has been done in expanding segmentation techniques, but in the specific area of tissues the task becomes very complex. Cell types, and consequently cell sizes and shapes, are more assorted in tissues; cells are often tightly packed and could only be captured halfway by any single area. Cell division traditionally relies on traditional

methodologies such as watershed calculations, whose parameters are physically defined for various cell types.

To make a good segmentation a good starting point is the location of the nuclei for a variety of reasons, including:

- many cells have only one nucleus, and
- nuclei with a very high signal-to-background ratio are widely available
- the nucleus is generally quite large, making it easy to detect.
- it is often in the center of the cell

Nuclei can be identified in the absence of staining (used to increase the contrast between the organisms and the background so that they are more readily seen in the light microscope.), but for fluorescence imaging, labeling with DAPI or Hoechst 33342 is common.

More and more recently, AI strategies have become standard, particularly in general imaging there has been increasing use of Convolutional Neural Networks (CNNs), for the purposes of image recognition, object detection, and synthetic image generation. The use of CNNs is widely used because it is able to automatically extract all the features of the image, for a large number of images.

A disadvantage of CNNs is that the accuracy of the results depends directly on the quality and quantity of the images given as input to the training, and most of the images have artifacts especially in focus and saturation.

For this reason it is important to explore and use this "problem" to train in robust way our network, which must be ready to deal with such aberrations.

For this purpose data augmentation is used, data augmentation is the creation, starting from the images that are given, of other useful images for the training, and to make more homogeneous the set of images.

The data augmentation includes in the pre-processing random rotations, cuts, overturning, etc..

Another advantageous aspect using the images obtained in this way, is that the algorithms avoid learning irrelevant aspects of an image such as, for example, the orientation.

UnMICST

As previously mentioned, focus problems are very common in images, in the recent research Yapp C. (2021), «UnMICST: Deep Learning with Real Augmentation for Robust Segmentation of Highly Multiplexed Images of Human Tissues», this phenomenon was imitated during data augmentation to improve the results of the network during segmentation. The results obtained were then compared to the standard method used, that is the application of a Gaussian blur to the training data.

The purpose was to add more contrast and blur variations in the dataset, few studies have been done by adding brightness and contrast variations.

The UnMICST method indeed, investigate ways to maximize the accuracy of image segmentation on multiplexed tissue data by including different types of information in the images and by augmenting the training data, particularly adding

blurred and oversaturated images to the training data to make the models more robust to the types of artifacts encountered in real tissue images.

A training and testing dataset with ground-truth annotations is generated through human inspection of lung adenocarcinoma, prostate-derived fibroblasts, normal tonsil, non-neoplastic small intestine, colon adenocarcinoma, glioblastoma, and non-neoplastic ovary.

These data are used to evaluate the segmentation accuracy obtained using three deep learning algorithms: UNet, Mask R-CNN, and Pyramid Scene Parsing Network (PSPNet), at this point a family of models called Universal Models for Identifying Cells and Segmenting Tissue (UnMICST) is introduced.

The three proposed approaches are referred to by the algorithms they use as UnMICST-U, UnMICST-M, and UnMICST-P, respectively. Specifically, the choice of the 3 different architectures depends on unique characteristics of the architectures:

- UNet was selected for its previous success in the biomedical domain
- Mask R-CNN was selected for its ability to perform both object detection and mask generation
- PSPNet was selected for its ability to integrate image features from multiple spatial scales.

Training, validation, and testing data were derived from 12 cores in 7 tissues and a total of 10,359 nuclei in the composition of colon:

- colon: 1,142;
- glioblastoma: 675;
- glioblastoma: 1735;
- ovary: 956;
- fibroblast: 922;
- small intestine: 1677;
- tonsil: 3252;

Each of the three architectures was then trained differently.

1. UnMICST-U model training: A UNet model (network and training strategy that relies on the strong use of data augmentation to use the available annotated samples more efficiently presented by Ronneberger O. (2015), «U-Net: Convolutional Networks for Biomedical Image Segmentation») with three classes was trained based on annotation of nuclei centers, nuclei contours, and background. The neural network is comprised of 4 layers and 80 input features. Training was performed using a batch size of 32 with the Adam Optimizer and a learning rate of 0.00005 with a decay rate of 0.98 every 5,000 steps until there was no improvement in accuracy or 100 epochs had been reached.
2. UnMICST-M model training: Mask R-CNN (a flexible and general framework for object instance segmentation presented by He K. (2018), «Mask R-CNN») begins by

detecting bounding boxes of nuclei and subsequently performs segmentation within each box. This approach eliminates the need for an intermediate watershed, or equivalent, segmentation step. Thus, Mask R-CNN directly calculates a segmentation mask, significantly reducing the overhead in traditional segmentation pipelines. ResNet50 was then adopted and the original input images were unsampled to 800 x 800 pixels and trained a model for 24 epochs using a batch size of 8. The Adam optimizer, with a weight decay of 0.0001 to prevent overfitting, was exploited with a variable learning rate, initially set to 0.01 and decreased by a factor of 0.1 at epochs 16 and 22.

3. UnMICST-P model training: A three class PSPNet model (convolutional neural network to segment the prostate automatically presented by Tian Z. (2018), «PSNet: prostate segmentation on MRI based on a convolutional neural network») was trained to extract cell nuclei centers, nuclei contours, and background from a wide variety of tissue types. The network employs a so-called pyramid pooling module whose purpose is to learn global as well as local features. The additional contextual information used by PSPNet allowed the segmentation algorithm to produce realistic probability maps with greater confidence. ResNet101 was used, the training of the network was performed using a batch size of 8 with an image size of 256 x 256-pixels for 15,000 iterations or until the minimum loss model was obtained. A standard cross entropy loss function was used during training. Gradient descent was performed using the Adam optimizer with a learning rate of 0.0001 and a weight decay parameter of 0.005 via L2 regularization. Batch normalization was employed for faster convergence, and a dropout probability of 0.5 was used in the final network layer to mitigate overfitting.

The accuracy of each model was determined by comparing the number of successfully segmented cells to the total number using a variable intersection over union (IoU) threshold ranging from 0.55 to 0.8. The IoU (the Jaccard Index) is computed by calculating the overlap between the ground truth annotation and the prediction using a ratio of the intersection to the union of pixels in both masks. The bigger the IoU, the greater the accuracy, with a value of 1 being optimal (although this is very rarely achieved). IoU is not sensitive to class-imbalance, unlike the traditional pixel accuracy statistic (the fraction of pixels in a picture that were correctly categorized).

The approach of comparing test images on trained models with Gaussian blurred or defocused image data is depicted in Fig.1. Higher contrast probability maps show greater confidence, while red arrows highlight regions of importance. Corresponding probability maps show that a model trained with defocused photos outperforms a Gaussian-blurred model, on defocused test images (the scale bar represents 20 micrometers).

It's visible from Fig.2 that incorporating real augmentations (red curve) into the training set is statistically significantly superior to training sets with Gaussian blur (yellow curve) and without real augmentations (blue curve) for UnMICST-U, UnMICST-M, and UnMICST-P.

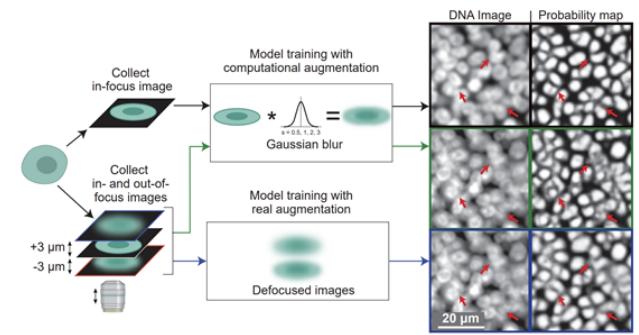


Fig. 1: Computation of Real Augmentation and Computational Augmentation.

Simulating defocused images with Gaussian blur is superior only to not enhancing the training data at all. In the final graph, we can also see how increasing by defocusing produces better results than a simple rotation, which is an easily deducible phenomena.

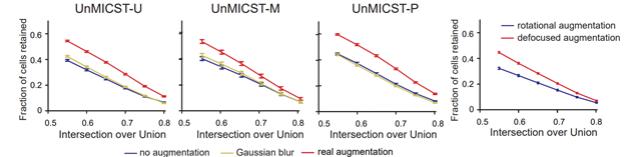


Fig. 2: Plots of three results using different Augmentation and Training Model.

MASK-NMS

Another segmentation method that uses deep learning, particularly for semantic segmentation and segmentation of object instances is preset by Kurt T. (2020), «Segmentation and Classification in Digital Pathology for Glioma Research: Challenges and Deep Learning Approaches» and, specifically, employs an adaptation of the MaskRCNN algorithm. The authors of this approach created pre- and post-processing processes to boost the algorithm's performance, notably morphological operations to patch holes in the masks. Morphological erosion and dilatation are employed to break fused nucleic masks.

In the post-processing, the predictions of five cross-training models are combined; this approach is known as Mask-NMS.

MASK-NMS takes unions of masks with maximum overlap and removes false-positive masks with a small overlap. It starts with a set of segmentation results, this set is called I. Each result in set I is assigned a score S, which is the value of the classification probability from the Mask-RCNN module and corresponds to the confidence level of the segmentation result. After selecting the segmentation with the maximum score M (among scores S), MASK-NMS removes it from the set I and appends it to the final segmentation set D. D is initialized to an empty set. It also removes any segmentation with an overlap greater than a threshold N in the set I, where the intersection over union (IoU) is used as the overlap metric.

As we already saw previously, IoU measures the similarity between finite sample sets. It is defined as the size of the intersection between two sets divided by the size of the union of the sets.

The selection process repeats until set I becomes empty. Finally, the segmentation results in set D is obtained.

To summarize, the MASK-NMS module assembles multiple results together and reduces false positives and false negatives. Regarding the network and its training modes,

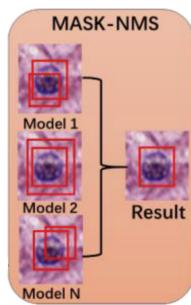


Fig. 3: Tissue image segmentation mode: Focus on MASK-NMS.

ResNet-101 is used to build a Mask-RCNN pyramid network backbone for nuclei segmentation, starting from the implementation of Matterport, the anchor sizes were reduced and the number of anchor was increased to be used because the nuclei are small objects and can be found anywhere in a tissue image.

In this model also the maximum number of predicted objects was increased because even a small image tile from a tissue slide can contain 1000 or more nuclei.

The dataset used was MSCOCO and the network layers were trained in multiple steps, starting from the heads up to the top layers. After this, the learning rate was reduced by a factor of 10 and the network was re-trained for a number of epochs of approximately 300, stochastic gradient descent.

Data augmentation was applied also in this model, to avoid overfitting, but the methods used are standard ones: rotation, clipping and Gaussian noise application. The method achieved a similarity coefficient compared to manual segmentation of 0.868.

Ilastik and CellProler, an easy pipeline

All of these methods, however precise, are not easy to use for doctors and biologists who do not have computer skills. For this reason, pipelines using user friendly and free software have been proposed, an example is a flexible and scalable image processing pipeline tailored for highly multiplexed images. It has been presented by Zanotelli V. and Damond N. (2019), «A flexible image segmentation pipeline based on pixel classification for heterogeneous multiplexed tissue images».

This pipeline allows single cell and image structure segmentation of hundreds of images. It is based on supervised pixel classification using Ilastik to distill the relevant segmentation information from multiplexed images in a semi-supervised and automated manner, followed by standard image segmentation using CellProler.

Ilastik is a user-friendly interactive program that enables

machine-learning-based (bio)image analysis for users who lack significant computer competence. It includes procedures for picture segmentation, item classification, counting, and tracking. Users customize the workflows to the problem at hand by supplying sparse training annotations for a nonlinear classifier interactively. Ilastik is capable of processing data in up to five dimensions (3D, time and number of channels). Wherever possible, its computational back end conducts operations on-demand, allowing for interactive prediction on data larger than RAM. After training the classifiers, Ilastik workflows can be applied to new data from the command line with no additional user intervention.

CellProfiler is a free, open-source software that allows scientists with no experience in computer vision or programming to automatically assess phenotypes from thousands of photos. Also, advanced image analysis techniques are available as distinct modules that can be linked together in a pipeline to recognize and measure biological objects and properties in images, particularly those obtained by fluorescence microscopy.

A clear representation of the overall process is shown in Fig.4: Ilastick aims to classify pixels into 3 classes: core (red), membrane (yellow) and background (green).

To do that n channels are used, but in the figure only three can be seen: (a), (b), (c).

At this point CellProler is used to segment the probability maps, class to obtain segmentation masks at nuclear and cellular level, as shown in (e) and (f). An example of a probability map is present in (g), and its final segmentation in (h).

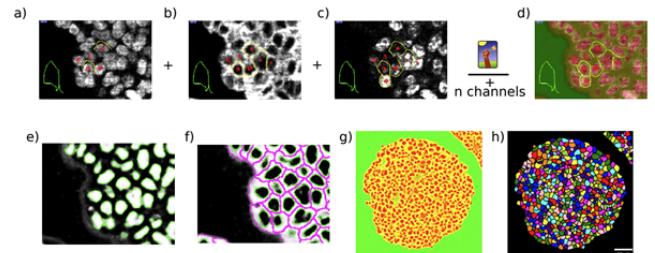


Fig. 4: Pipeline for heterogeneous multiplexed tissue images based on pixel classification.

Overall, the presented pipeline enables high-throughput processing of hundreds of multiplexed tissue images and thus forms a solid foundation for standardized, open source data analysis, useful even for those who do not have much experience with image processing, and therefore with great potential in the world of biologists and doctors.

Mathematical morphology

Another segmentation method to remove noise and accentuate nuclear shapes, based on gray scale reconstruction, was proposed by Po-Whei H. and Lai Y. (2010), «Effective Segmentation and Classification for HCC Biopsy Images». In Fig.5 there is the flowchart of dual morphological gray scale reconstruction method, that is based on erosion, dilation and reconstructions.

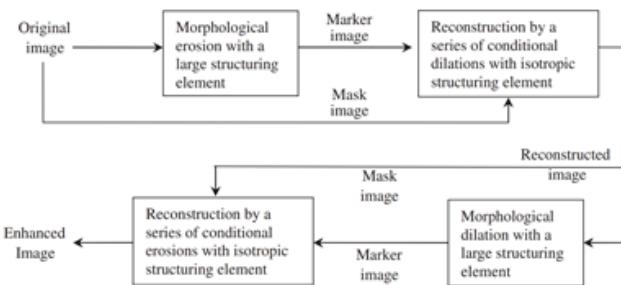


Fig. 5: The flowchart of dual morphological gray scale reconstruction method.

The result of the method is shown in Fig.6, we can see that starting from (a), the original RGB color, we will use only (b), the grey scale image obtained from the red plane of (a). At this point in (c) an erosion is applied and in (d) a morphological reconstruction using (b) as a mask and (c) as marker is applied. In (e) there is a dilation followed by (f), a morphological reconstruction using (d) as the mask and (e) as marker.

The result is satisfactory, core shapes are well preserved and other irrelevant objects are removed.

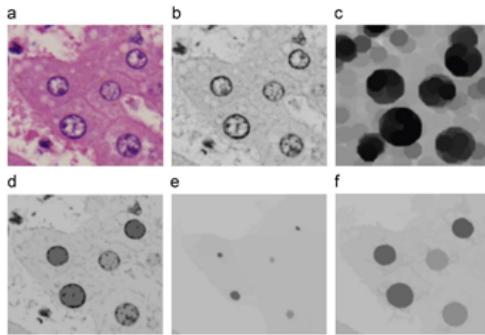


Fig. 6: An example of dual morphological gray scale reconstruction for HCC biopsy images.

Nuclei were segmented from the images with the pre-processing just shown, using marker-controlled watershed transformation and GVF active contour method following the steps shown in Fig.7.

Going into more detail, starting from (a), the image ob-

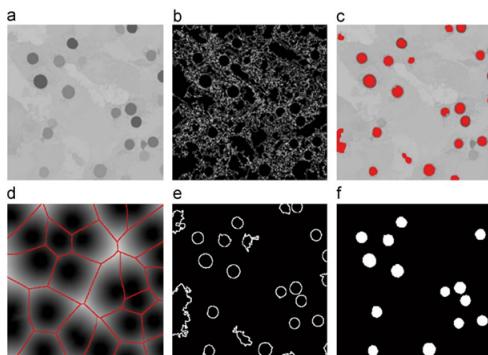


Fig. 7: An example of nuclei segmentation by using marker controlled watershed transform.

tained from pre-processing step, watershed transform is applied obtaining (b), the ridge lines, it's clearly visible that

regular over-segmentation problem occurs. In (c) the foreground markers associated with nuclei are find and in (d) the background markers are indicated by ridge lines. In (e) we can find the ridge lines obtained from marker-controlled watershed transform and finally in (f) the segmentation result after removing undesirable elements by using heuristic rules. At this point a gradient vector flow (GVF) method is applied for refining the contours of nuclei, the final result of the overall process is shown in the examples in the Fig.8.

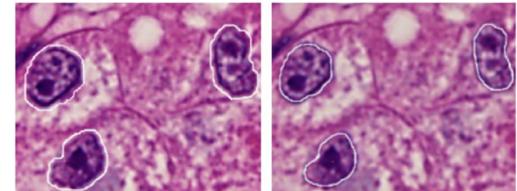


Fig. 8: An example of refining contours of nuclei by using GVF active contour method.

This hybrid technique is strong in terms of noise removal and nucleus shape preservation in biopsy.

CLASSIFICATION

SVM

Although the work regarding pre-processing and segmentation in the paper of Po-Whei H. and Lai Y. (2010), «Effective Segmentation and Classification for HCC Biopsy Images», is of considerable importance, the main contribution is to develop an efficient and effective automated grading system for biopsy images, this goal clearly requires also a part of classification, for this reason a feature extractor was used.

Particularly 14 features were extracted from segmented biopsy images according to six types of characteristics including nuclear size, nucleocytoplasmic ratio, irregularity, hyperchromatism, anisonucleosis, and nuclear texture. These criteria include both local and global aspects, allowing benignancy and various degrees of malignancy to be separated efficiently.

To select an appropriate feature subset from the 14 features for the support vector machine associated with each decision node of the decision-graph classifier, the sequential floating forward selection approach was utilized.

The system is efficient because an HCC biopsy image can be classified into a correct grade through four decision steps and the system is effective, experimental results show that 94.54 % of accuracy can be achieved, it was also verified that using an SVM that directly uses all 14 features has a lower accuracy, precisely 92.88 % as shown in Fig.9 and Fig.10.

Grades	Visual grading	Classification result					Accuracy (%)
		G ₀	G ₁	G ₂	G ₃	G ₄	
G ₀	58	58	0	0	0	0	100
G ₁	26	0	26	0	0	0	100
G ₂	75	0	0	72	2	1	97.33
G ₃	220	0	1	0	201	18	91.36
G ₄	225	0	0	6	5	214	95.11
Total	604	58	27	78	208	231	94.54

Fig. 9: The performance of HCC classification by using our decision-graph SVM classifier with an optimal subset of features.

Even if SVM can work nicely, in recent years, convolutional neural networks (CNNs) have become the state of the art for

Grades	Visual grading	Classification result					Accuracy (%)
		G ₀	G ₁	G ₂	G ₃	G ₄	
G ₀	58	58	0	0	0	0	100
G ₁	26	0	26	0	0	0	100
G ₂	75	0	0	70	3	2	93.33
G ₃	220	0	2	2	197	19	89.55
G ₄	225	0	0	8	7	210	93.33
Total	604	58	28	80	207	231	92.88

Fig. 10: The performance of HCC classification by using SVM and all 14 features.

object recognition in computer vision, most methods in fact use this type of classifier.

FCNN and XGBoosst

An example is presented by Solorzano L. (2019), «Machine Learning for Cell Classification and Neighborhood Analysis in Glioma Tissue» through a methodology for cell classification using two popular machine learning architectures FCNN and XGBoosst.

The classifier ensembles arrive at similar conclusions as the two expert annotators, which is a good reason to believe that this method is a good way to classify larger cohorts of tissue microarrays (TMA) cores and avoid manual thresholding.

The FCNN architecture for cell classification consisting of an input layer with the size of the cell feature vector (size 68) and as output a feature vector of size six, corresponding to the six different cell types.

The model had three hidden fully connected layers in the middle, of sizes 100, 200, and 300 with rectified linear unit (ReLU) activation. Batch normalization was used after the second and third hidden layer, before the activation functions, to reduce internal covariate shift and avoid convergence problems. To train the network, AdaBelief was used as an optimizer, this optimizer arrived to the solution twice as fast as a stochastic gradient descent optimizer while providing similar classification accuracy.

Each FCNN was trained for 15 epochs and all networks were trained using PyTorch 1.7.1.

For XGBoost (a scalable end-to-end tree boosting system) most of the default parameters were used, this resulted in an ensemble of 1200 trees (given by the product of the number of classes and the number of rounds) that are weak classifiers. Each tree is more specialized on hard training examples than the previous trees.

To obtain a confidence measure for the XGBoost method, similar to FCNN, an ensemble of models is required. For this purpose, 100 XGBoost models were trained and the mode of their votes was used as the assigned class and the percentage of votes for the assigned class as the confidence measure.

All XGBoosts were trained using the XGBoost Python API implementation version 1.1.1.

The results of the two methos are shown in the Fig.11 .

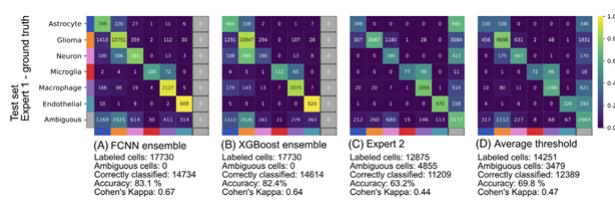


Fig. 11: Confusion matrices comparing ground truth.

Another contribution of the paper is to have investigated the benefits of including the differences of 90th percentiles of intensity, simply called "d90s features", in relation to the performance of both machine learning models.

100 components per model were trained, one that included d90s features and one that included only common marker intensity features.

Each ensemble architecture used the same parameters for both feature sets, both of them increased in accuracy when d90s features were presented as can be verified in Fig.12.

	Model	Precision (%)	Precision in maximum confidence (%)
with d90s	FCNN ensemble	83.1	94.47
	XGBoost	82.4	88.3
without d90s	FCNN ensemble	80.8	-
	XGBoost	82.0	-

Fig. 12: Effect of d90s features.

Combination of radiologic and histologic data for CNN

Some methods attempt to combine radiologic data with histologic and tissue data to improve the quality of classifier results. Although having both types of images available is not always feasible, the approaches are very interesting and worth mentioning.

Citing again the paper done by Kurt T. (2020), «Segmentation and Classification in Digital Pathology for Glioma Research: Challenges and Deep Learning Approaches», we can have an overview of 3 different classification methods of lower grade glioma (LGG) cases into oligodendrogloma and astrocytoma subtypes using both radiological and histopathological images.

The first method combines the predictions of a tissue image model to refine the low confidence predictions of a radiological image model.

The second method implements two separate classification models for radiographic and histological images and combines them through a dropout learning ensemble.

The third method employs multiple deep learning models, one for classifying tissue pictures and two for segmenting and classifying radiological images. To assign a class label to each case, a weighted average operation is used to the classification results of the tissue and radiological images. The approaches yielded accuracy values of 0.90, 0.80, and 0.75, respectively accuracy was calculated by dividing the number of correctly classified cases by the total number of cases.

Because the focus of this document is on histology and tissue images rather than radiological images, we shall concentrate on the description of image processing, since, the procedures carried out could be taken as a starting point for future developments and applications. Going more in deep in the three approaches: in the first approach called, "An Approach for Classification of Low-Grade Gliomas Using Combined Radiology and Pathology Image Data", in order to detect and segment tissue regions and remove regions corresponding to

the glass background, a tissue image is first converted from RGB color space to HSV color space.

Then, lower and upper thresholds are applied on the color intensities to obtain a binary mask. The binary mask is processed to fill small holes and remove clusters from foreground pixels. At this point, bounding boxes around all discrete contours are obtained. The bounding boxes serve as the outline for the patch extraction process.

The patch extraction process partitions the segmented tissue region into 224×224 pixel patches, which are normalized by color and assigned the same label as the whole slide image (WSI). A subset of distinct patches is filtered using an outlier detection technique called Isolation Forest.

The filtering step is performed by training an autoencoder with a pixel-wise reconstruction loss, to generate vector representations of patch features from the input image. The isolation forest method then runs with these feature vectors to find the outlier patches. The remaining patches after the outlier detection step are used to refine a DenseNet-161 network, which has been pre-trained on ImageNet. The binary cross entropy is used as a loss function. During the prediction phase, test patches extracted from a WSI are ranked using the trained model, and a probability score is assigned to the image based on the grade of the predicted classes for individual patches.

The overall process is shown in Fig.13.

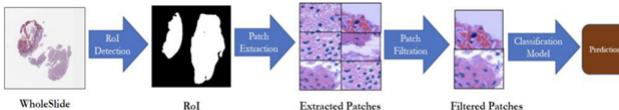


Fig. 13: Pathology image analysis in "An Approach for Classification of Low-Grade Gliomas Using Combined Radiology and Pathology Image Data".

At this point there is the combination of the predictions from the pathology and radiology models. The results from the two models are processed in a confidence-based voting step, which chooses the class with the highest prediction probability value.

The Second method, "Dropout-Enabled Ensemble Learning for Multi-Scale Biomedical Image Classification", is a multiple instance learning approach, Fig14 shows the implementation for histopathological images.

The learning phase is performed after a pre-processing phase consisting of tissue detection, color normalization, and tiling. Tissue detection is done with Otsu thresholding (a thresholding method that uses image histogram analysis techniques and is particularly suitable for bimodal images, that are images with histograms with a clear separation between two main peaks) to detect and segment only tissue regions, eliminating regions that have a glass background. A simple histogram equalization algorithm is used for color normalization prior to tiling.

The tiling step extracts 20 patches of 448×448 pixels with uniform random sampling.

Once the image patches have been extracted, a pre-trained DenseNet network on ImageNet is tuned, after removing its last fully connected layer. The rest of the network is used as a fixed feature extractor for tissue images, and two fully connected dropout layers are used for classification. Also in

this model, classical data augmentation techniques were used along with dropout to eliminate overfitting.

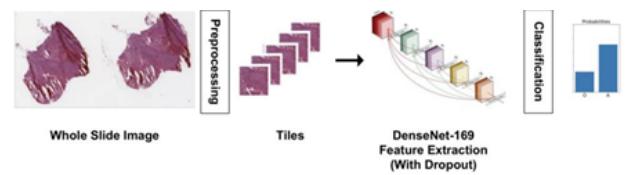


Fig. 14: Histopathology image analysis pipeline in "Dropout-Enabled Ensemble Learning for Multi-Scale Biomedical Image Classification"

An ensemble model combining classifications from the radiology and histopathology image analysis pipelines is applied at this point.

The third method, "A Weighted Average-Based Classification Method", analyzes each imaging modality (radiological images and pathological images) separately and combines the prediction results by a weighted average operation. Pathologic image classification is performed by identifying tissue features that differentiate oligodendrogloma from astrocytoma. In the proposed method, each histological image is divided into 512×512 patches.

A sample set is created to identify typical samples of both sub-types of diffuse gliomas of the brain to evaluate the imbalance in the data.

In order to prevent classification error caused by data imbalance, the method expands the sample set using data augmentation, specifically by rotating the original image in symmetric and asymmetric directions, dropout layers or batch normalization layers are also added to the classification model to reduce the risk of overfitting.

The balanced samples are sent to a CNN classifier network, which is trained to fully recognize the tissue and cell characteristics of oligodendrogloma and astrocytoma as shown in Fig.15. The CNN network used is VGG16.

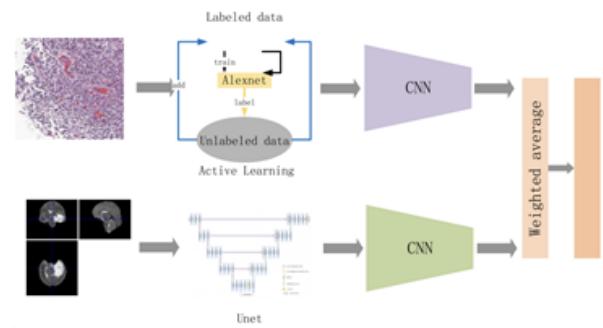


Fig. 15: The entire flow of "Weighted Average-Based Classification Method".

The Accuracy scores of the classification methods presented are:

ANALYSIS AND PREDICTION

Since cancer is characterized by loss of control of the stem population, the study of the stem niche, which is the specialized micro environment surrounding a stem cell population, is crucial.

Method	Score
Section "An Approach for Classification Of Low-Grade Gliomas Using Combined Radiology and Pathology Image Data"	0.90
Section "Dropout-Enabled Ensemble Learning for Multi-Scale Biomedical Image Classification"	0.80
Section "A Weighted Average-Based Classification Method"	0.75

Fig. 16: Accuracy scores of the classification methods that combine radiologic data with histologic and tissue data.

Spage2vec

To be able to do analysis on the stem niche and tissue architecture, classification is necessary, followed by other methods of analysis such as Spage2vec, a method for analysis of tissue spatial cellular composition on classification results, presented by Solorzano L. (2019), «Machine Learning for Cell Classification and Neighborhood Analysis in Glioma Tissue».

Spage2vec is an unsupervised segmentation-free method for deciphering the spatial transcriptome heterogeneity of complex tissues at the sub-cellular level. Spage2vec represents the spatial transcriptomic landscape of tissue samples as a graph and leverages a powerful machine learning graph representation technique to create a lower dimensional representation of local spatial gene expression, it was firstly proposed by Partel G. and Wählby C (2020), «Spage2vec: Unsupervised representation of localized spatial gene expression signatures».

Spage2vec provides several novel characteristics: it associates a value with a cell based on its neighborhood, meaning that it characterizes each neighborhood, offering a measure of the difference between neighborhoods.

To verify Spage2vec as a method for detecting cell niches an exploratory analysis and comparison with a previously study of tissue architecture from CODEX data was performed.

Since Spage2vec in gene expression can be associated with spatial compartments and cell types, the authors were able to observe that Spage2vec applied to cell types within the tissue has a relationship with spatial compartments, or niches, confirming that Spage2vec is applicable to single cell data.

For this reason, the method was used with the output of the classification proposed by Solorzano L. (2019), «Machine Learning for Cell Classification and Neighborhood Analysis in Glioma Tissue» to find cell niches in the TMA nuclei. The results can be seen in Fig.17A, where similar colors correspond to similar niches. The angular histogram of Spage2vec values in Fig.17B shows the frequency (x-axis) of different niches (y-axis), making it clear that some niches are more common than others (log scale was used). Fig.17C shows the cell-vicinity composition of each niche, and the peaks in the Spage2vec histogram coincide with the peaks in the average class representation in the niche histogram.

The colors represent:

- Blue: Astrocyte
- Orange: Glioma
- Pink: Neuron
- Red: Microglia

- Purple: Macrophages
- Navy-green: Endothelial

From the results, Spage2vec seems able to describe a niche with a single value and is able to organize niches based on similarity, capturing distinct class compositions. High values in the average cell composition indicate niche compositions with low Standard Deviation and thus homogeneous niche cell composition, while low values indicate very heterogeneous niches.

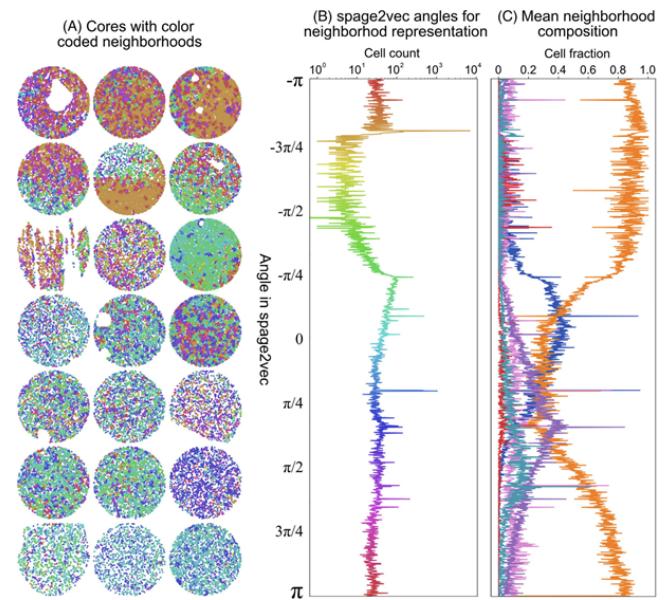


Fig. 17: Cell niches defined by Spage2vec on glioma TMA cores.

MULTIPLAI

Since cancer progression and response to treatment largely depend on the structure of the tissue surrounding cancer cells in a tumor, known as the tumor microenvironment (TME). Recent technological advancements have resulted in the creation of highly multiplexed imaging techniques like Imaging Mass Cytometry (IMC), which capture the intricacy of the TME by providing spatial tissue maps of dozens of proteins. Combining these multidimensional cellular phenotypes with their spatial organization to predict clinically relevant information is a computationally challenging task. Till now, not many method has directly addressed it.

In the paper Martin-Gonzalez P. (2021), «Predictive Modelling of Highly Multiplexed Tumour Tissue Images by Graph Neural Networks», MULTIPLAI, a framework for predicting clinical biomarkers from IMC data, was proposed and evaluated.

The method is based on attention-based graph neural networks (GNNs) that integrate both phenotypic and spatial dimensions of Imaging Mass Cytometry (IMC). MULTIPLAI was utilized in this proof-of-concept study to predict estrogen receptor (ER) status, an important clinical characteristic for breast cancer patients. On 240 samples, various framework designs were trained and benchmarked with graph learning utilizing graph kernels.

When the optimal mix of graph convolution and pooling layers was utilized, propagation attribute graph kernels achieved a balanced class accuracy of 66.18 % on the development set, whereas GNNs achieved a balanced class accuracy of 90.00 % on the same set. This design was validated further in internal and external test sets from several institutions, confirming the method's generalizability.

Going into more detail about the method, and the choice of architecture parameters, for MULTIPLAI, six GNN architectures were designed with different numbers of graphics convolution layers and different pooling methods.

The graph-level features obtained after pooling in all cases were fed into two fully connected layers and a sigmoid function, whose outputs were rounded to obtain binary predictions. Binary Cross Entropy (BCE) Loss and the Adam optimizer were used. The architectures were implemented using Deep Graph Library (version 0.4.3) and Pytorch (version 1.5.1). Each architecture was trained on the training set using mini-batches, and for each epoch the loss and accuracy balanced by class in the development set was evaluated. Early stopping was implemented by finding the minimum loss in the development set and allowing 20 epochs of patience before stopping. The grid search for hyper-parameter optimization was performed by batch sizes of 10, 20, and 30, hidden sizes of fully connected layers 50, 100, and 150, dropouts of 0.3 and 0.5, learning rates of 0.01 and 0.001, decay weights of 0.1, 0.01, and 0.01, and hidden graph sizes of 50, 100, and 150 in cases with two graph convolutions.

Each hyperparameter was optimized to give a lower loss and lower standard deviation in the development set. Once switched to GNNs, even the shallow architectures were able to capture TME differences in ER+ and ER- patients with the lowest balanced accuracy of 78.40. This suggests the potential of graph representation learning for this problem, and overall, GNN models do significantly better than Graph Kernel.

Furthermore, the losses in models with two layers are significantly lower than those with only one, for this reason the authors decided to select the combination of two graph convolutions and attention pooling as the best model for the next sections since it achieved the lowest binary cross entropy loss (0.159) on the development set.

Although this method could be improved both because of some choices made by the authors and because there is currently little literature available on data augmentation for graphs, GNN-based methods appear adequate for spatially integrating the different phenotypes of TME that could be used to make data-driven clinical decisions.

To summarize, the data appear to indicate that this technique catches essential TME traits of clinical significance. MULTIPLAI is the first GNN application to this type of data, opening up new possibilities for predictive modeling of highly multiplexed images.

The scheme of the overall process is shown in Fig.18.

SPADE

The last method that profile molecular characteristics linked with the morphological landscape of tissue for the aim of investigate the structural and spatial patterns that underpin the biological function of tissues is SPADE, introduced in

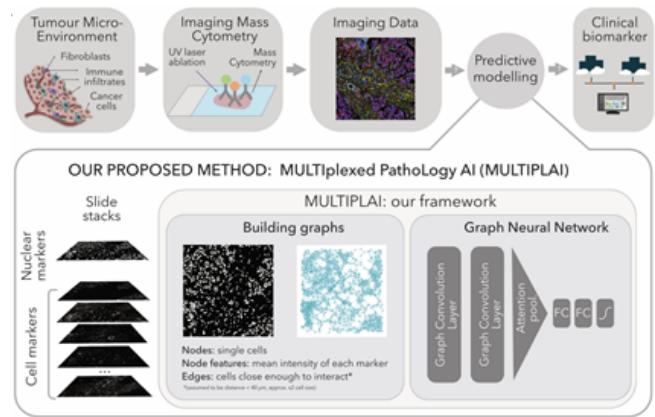


Fig. 18: Overview of the proposed method.

Bae S. (2021), «Discovery of molecular features underlying the morphological landscape by integrating spatial transcriptomic data with deep features of tissue images».

SPADE is a method for detecting spatial gene expression patterns in tissue pictures using deep learning. It identifies gene expression indicators by merging morphological features from an image patch surrounding each place holding transcriptomic data.

To define visual latent features associated to gene expression, a convolutional neural network (CNN) was used. Molecular markers from various tissues associated with the morphological landscape were provided in order to discover not only a spatial pattern of gene expression in tissues, but also biological processes related to histological architecture.

Significant gene markers were discovered that were linked to picture information acquired by a CNN.

SPADE genes, which are gene expression markers linked to the morphological landscape, were discovered using five publicly available datasets.

A pre-trained CNN method discovered picture latent characteristics represented by 512D vectors from image patches surrounding transcriptome data sites. To generate highly variable picture latent features, PCA was applied to the CNN output for all patches corresponding to spots. SPADE genes were identified using linear regression analysis with image feature PCs.

SPADE focuses on the relationship between morphology and gene features and can be used to examine molecular profiles responsible for the tissue morphological landscape by listing relevant genes and biological processes.

The combination of various types of data, images, and spatially resolved transcriptomes may aid in revealing the strong relationship between structure and molecular processes, eventually leading to a complete understanding of disease pathophysiology.

SPADE can also be used on fluorescence pictures, as illustrated in Fig.19 and Fig.20.

Fig.19A shows the spatial mapping of image latents PC1, PC2 and PC3. PC values in each spot are displayed using colormaps.

The maximum and minimum colormaps values represent two standard deviations above and below the mean value, respectively. Spatial mapping of the top 5 genes showing the greatest contrast in Fig.19B PC1, Fig.19C PC2 and Fig.19D

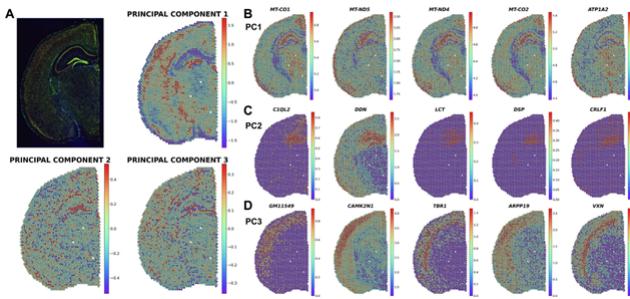


Fig. 19: Application of SPADE in immunofluorescence staining of mouse brain.

PC3 latent image space. The top genes are presented in descending order. The normalized gene expression level at each point is displayed using a color map. The maximum and minimum values in the color map represent two standard deviations above and below the mean expression, respectively. In Fig. 20 the spatial distribution of SPADE and highly variable genes (HVG) based on cluster spots mapped to mouse brain tissue.

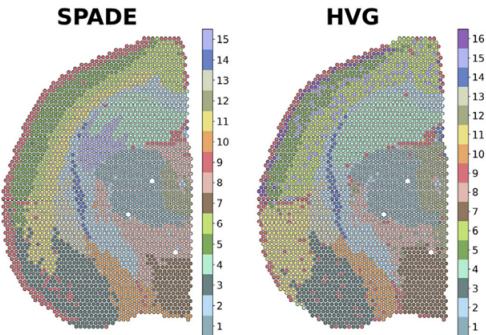


Fig. 20: Application of SPADE in immunofluorescence staining of mouse brain.

The immunofluorescence image background has been removed so that the colors of each cluster are more clearly displayed. The cluster numbers for the SPADE or HVG-based cluster are displayed in the right panel.

CONCLUSIONS ABOUT THE STATE OF ART METHODS

During this research, several methods of segmentation, classification, and experimental approaches to the study of cancer cell density and prediction in tissues were presented. In all three areas, we started from methods considered to be state of the art, moving to more innovative methods.

Each of them seems to have good performances, regarding the segmentation part the first approach described is difficult to implement as it directly depends on the dataset that is used. While the other three methods are more accessible as they do not depend on the dataset, they are pipelines that use free-source software, deep learning and morphological methods. Regarding the classification, the three methods presented that use combined results between radiological and histological data of tissues, are very interesting and innovative but in our specific case they are not methods that can be used because the images in our possession are only histological. The other

two approaches presented that involve SVM or CNN are more appropriate to our data. Regarding the study of the stem niche, and the structure of the tissue surrounding cancer, all are promising methods, still experimental, the idea is to use them to identify any links between structure and molecular processes and analysis on the stem niche.

PRACTICAL WORK

Segmentation

Starting from the image provided (Fig. 21), the first part implemented was the detection of the cells in the image, to do this the histogram (Fig. 22) of the available image was examined, in particular the aim was to find a value to set as a threshold that could distinguish the cells from the background.

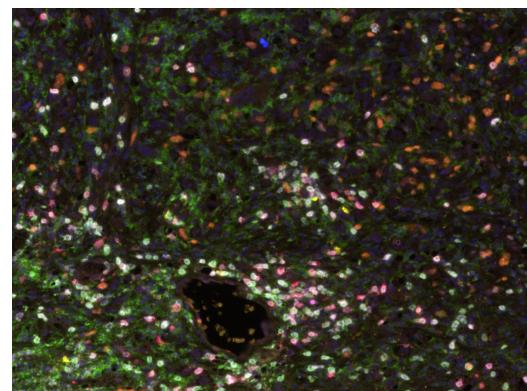


Fig. 21: Initial Prepared Image

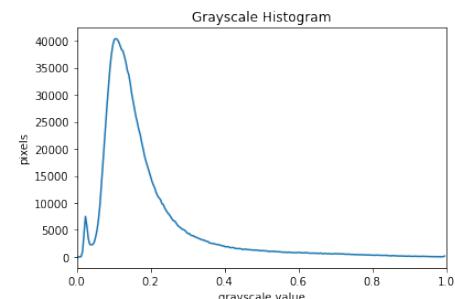


Fig. 22: Histogram of the image

This threshold was set at $t = 0.3$, and only the part of the histogram above this threshold was considered. The result obtained is shown in Fig. 23, where it can be seen only the areas belonging to the cells highlighted. The watershed technique was also initially tested, but, as this method has the disadvantage that it is highly sensitive to local minima, since at each minima a watershed is created, there were many errors even using the markers. The errors were certainly accentuated by the proximity of the cells.

Classification

At this point, using the images characterising the cells activated by the chemical markers, it was possible to begin the cell classification phase.

To do this, a function called **mask** was implemented:

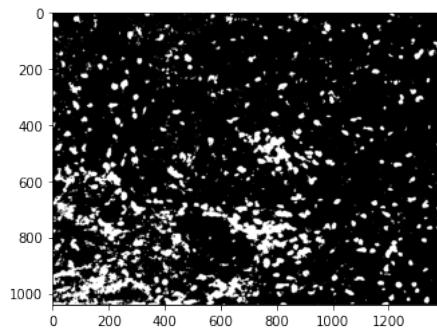


Fig. 23: Thresholded image

```
mask(im , th , op , er , color , plot=False )
```

The marker image is given as input to the function (**im**) with a threshold to transform it into a binary one (**th**), along with the size of the aperture and erosion structuring element (respectively, **op** and **er**), finally the colour that will characterise the chemical marker is given (**color**) and whether the graph is to be displayed, which by default is set to False (**plot**).

One by one, all marker images are fed into the function, each with specific parameters set to improve the performance of the algorithm, all of them are shown in Tabular 1.

Table	Threshold	Opening	Erosion
CD3	0.8	4	2
CD20	0.8	4	2
KI67	0.75	3	2
CD4+	0.73	1	1
CD8+	0.75	2	2
FOXP3	0.75	1	0

TABLE 1: SETTED PARAMETERS

At this stage, all the markers were characterized by a specific color: pink for CD3 cells, yellow for CD20, orange for KI67, green for CD4, red for CD8 and light blue for FOXP3.

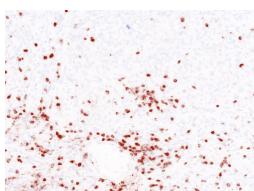


Fig. 24: CD3 Before Mask application

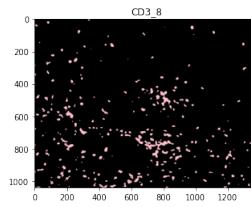


Fig. 25: CD3 After Mask Application

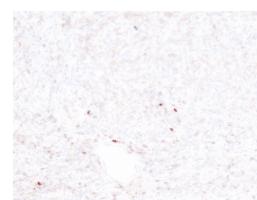


Fig. 26: CD20 Before Mask application

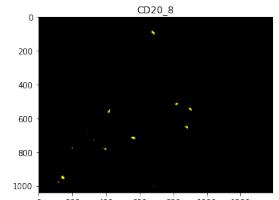


Fig. 27: CD20 After Mask Application

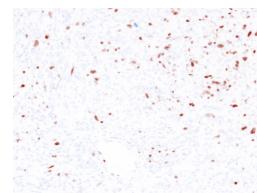


Fig. 28: KI67 Before Mask application

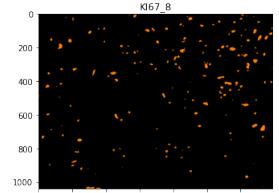


Fig. 29: KI67 After Mask Application

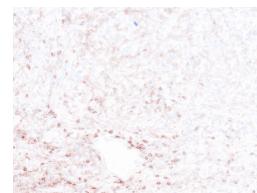


Fig. 30: CD4 Before Mask application

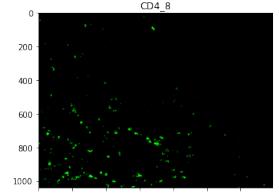


Fig. 31: CD4 After Mask Application

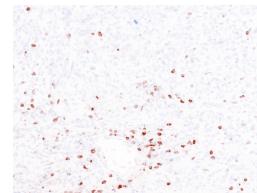


Fig. 32: CD8 Before Mask application

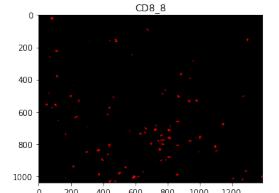


Fig. 33: CD8 After Mask Application



Fig. 34: FOXP3 Before Mask application

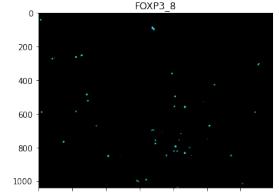


Fig. 35: FOXP3 After Mask Application

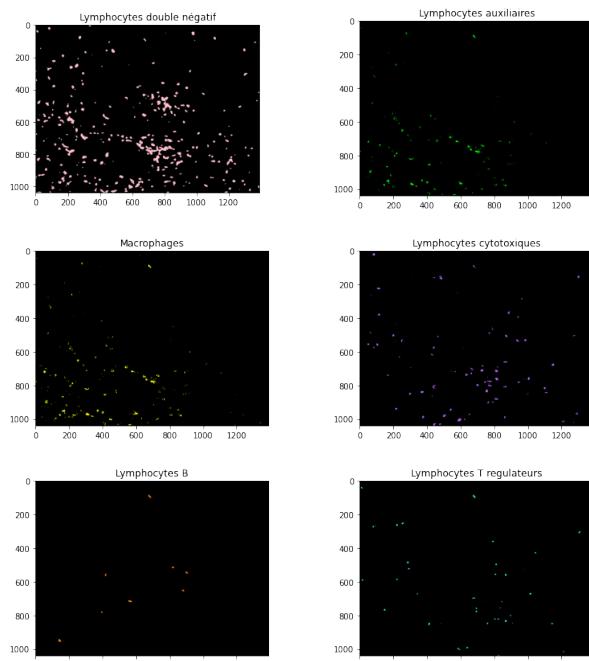
At this point we have highlighted all the cells activated by the various markers in the image, the next step is to combine the various markers to classify the different types of cells in the image.

To do this we have used the table provided by the medical experts shown in Fig. 36.

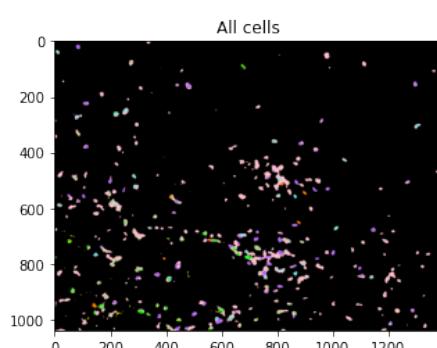
Combinations	Combinaisons de marquages panel lymphocyte (1=présence d'immunomarqueur ; 0=absence d'immunomarqueur)						Types cellulaires
	CD3+	CD4+	CD8+	CD20+	FOXP3+	EGFR+	
1	1	0	0	0	0	0	Lymphocytes double négatif CD3+ CD4- CD8-
2	1	1	0	0	0	0	Lymphocytes auxiliaires CD3+ CD4+
3	1	1	1	0	0	0	Lymphocytes cytotoxiques CD3+ CD8+
4	1	0	1	0	0	0	Lymphocytes régulateurs CD3+ FOXP3+
5	0	1	0	0	0	0	Macrophages
6	0	0	0	1	0	0	Lymphocytes B
7	0	0	0	0	1	0	Lymphocytes T régulateurs
8	1	0	0	0	0	1	Lymphocytes T helper
9	1	1	0	0	0	1	Lymphocytes T killer
10	1	1	1	0	0	1	Lymphocytes T naïfs
11	1	1	0	0	0	0	Lymphocytes naïfs
12	1	1	0	0	1	1	Lymphocytes naïfs régulateurs
13	1	1	0	0	0	0	Lymphocytes naïfs killer
14	1	0	0	0	1	1	Lymphocytes naïfs T helper

Fig. 36: Histogram of the image

The strategy used was to check the colours present in a certain area for all the various agents used, if the combination of colours in a certain cell was consistent with the classification table given to me, the cell would be classified into one of six possible types (Fig. 37). The color chosen are coherent with the one chosen for highlight different types of cells in the Excel file. One thing to bear in mind is that due to the frequent

**Fig. 37:** Different types of cells

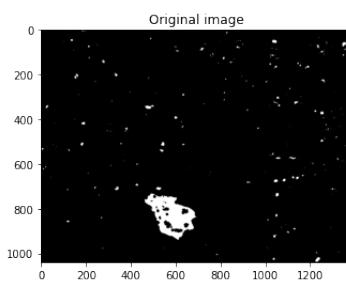
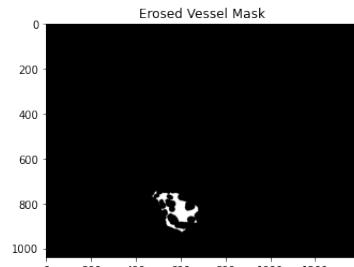
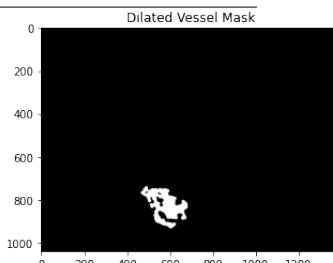
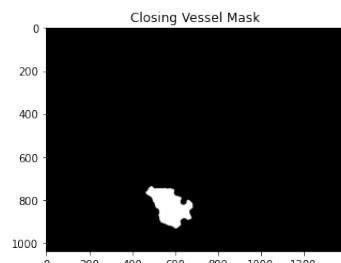
overlapping in the case of the Double negative lymphocytes, caused by its high density, erosion (structuring element was a disk of size 3) was used to try to divide the cells. This step is crucial for when the centroids will be calculated. Combining all the possible cell founded in the image, we obtain an image with cells of different colors as showed in Figure 38:

**Fig. 38:** All the Cells

Analysis

In order to carry out a spatial analysis of cell positioning, the initial image was used again, this time with the aim of highlighting the blood vessel. This is because it is known that glioma cells migrate to different regions of the brain guided by the extension of blood vessels, colonising the healthy adjacent tissue, as explained by Dubois L. G. (2014) in the paper «Gliomas and the vascular fragility of the blood brain barrier».

This time the threshold used was $t=0.05$, and the part of the histogram under this threshold was taken. The result (Fig. 39) necessitated the use of erosion to eliminate all the white dots that did not belong to the vessel (Fig. 40), a dilation, applied with a structuring element of a different size (Fig. 41), otherwise an opening would have sufficed. Finally a closing was applied to make the result as precise as possible (Fig. 42).

**Fig. 39:** Thresholded image**Fig. 40:** Erosion**Fig. 41:** Dilatation**Fig. 42:** Closing

The structuring elements were:

Table	Dilation	Erosion	Closing
SE	disk, size=8	disk, size=7	disk, size=13

TABLE 2: SETTED PARAMETERS

The closed image was used in the function:

```
imagecreation(dil ,n)
```

were, **dil** was the size of the dilation that will be used and which will characterise the width of the area to be studied during the cell density analysis, **n** is the number of iterations. We iteratively loop **n** times dilating and subtracting the previous result of the dilation. We have 21 iteration instead of 20 because in the first one, we just dilate the closing result of the Vessel.

The images are saved in a list, the expansion is shown in Figure 43, but can also be displayed with the animation at this link.

At this point a function called areaofinterest was created, in

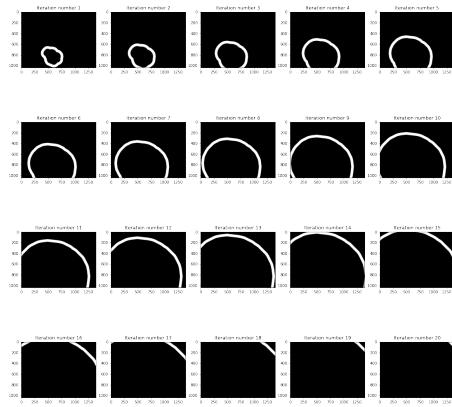


Fig. 43: Loop of dilation and subtraction

which, taking as input one of the images representing a type of cell, it would subdivide the image into the different areas just created.

```
areaofinterest(exp , cells , plot=False )
```

The list with the successive expansions is given as an input in **exp**, **cells** represent a list containing all the images of the various types of cells previously classified and **plot** represent whether the graph is to be displayed, which by default is set to False. At this point we have a list of 6 elements which contains for each of the 6 elements 20 images, representative of the cells present for the 20 rays created.

A visualisation to ideally understand the process can be seen in Figure 44 or at this link. In order to carry out a spatial

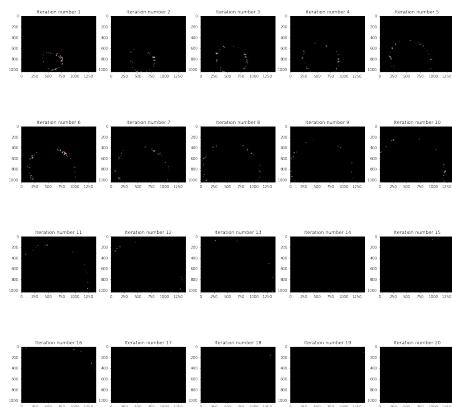


Fig. 44: Mask of dilation and subtraction on the cell image

analysis, it is necessary to obtain the centroids of the cells, both in order to count them and also to obtain statistics on their distances, which is why the propertyacquisition function was implemented, using the skimage.measure label and regionprops functions.

```
propacq( cell , cntrlist , name , plot=False )
```

cell is the image of which the centroids will be calculated, in this specific case, we are going to use as input image the result of the function ".readafinteres", so we will consider only the area in a specific radius. **cntrlist** it's the list where the centroids will be saved, **name** is just used to print the correct name in the plot and **plot** represent whether the graph is to be displayed, which by default is set to False.

At this point we have a list of lists with all the coordinates of the centroids, particularly we will have a list of 6 elements that will be 6 lists with 20 lists inside each of them, in each of the 20 lists we will have the coordinates of the centroids that were found.

At this stage we have all the data necessary in order to be able to start making statistics and graphs about the spatial arrangement, the first thing that was done was a graph describing the distribution of cells divided into each class in each radius. The result is shown in Figure 45 This graph is very informa-

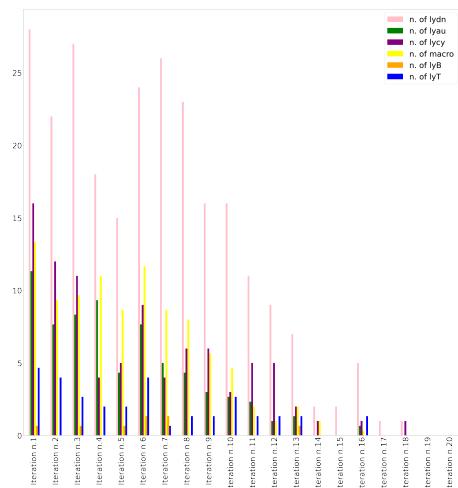


Fig. 45: Number of cells per radius

tive, but another one with the same information in percentage would give us further and more objective information about variations in distribution, the graph is show in Figure 46. From these two graphs we can clearly see that Double negative lymphocytes are a large in number for every radius, but the most interesting thing is that they tend to increase in percentage moving away from the vessel. This is due to the intrinsic nature of Double negative lymphocytes, they are immature cells, mainly responsible for proliferation. We can see them as cells that for the moment do not have a specific task, they need to be activated to be able to start playing a specific role in the fight against glioma. For the activation they need another type of cell, Macrophages.

Macrophages are a cell population that has two activities: one is antineoplastic, they do direct phagocytosis of tumoral cells, another function is the activation of Double negative lymphocytes. These two jobs being carried out by Macrophages explain their behaviour in the graphs, we can indeed see that they are a lot in the area close to the vessels, because they are tumor killers, but also, we can see that in the suburbs

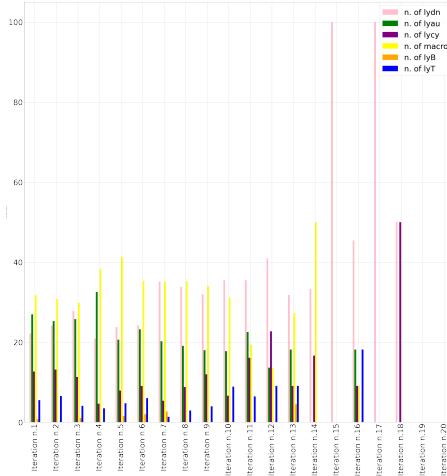


Fig. 46: Number of cells per radius in percentage

they try to follow the same distributions as the Double negative lymphocytes, this is because they will activate them, so they must be in percentages a similar number.

From the graphs we can also see that Lymphocytes auxiliaries tend to decrease going far away from the vessel, both in absolute number but also in percentage, this is due to their function, they are responsible of the genesis of the pro-inflammatory environment against cancer, so they must be close to the vessel.

Another analysis that was done was to see the trend in the number of total cells by radius, and to compare this with the number of total cells activated by KI67.

KI67 is a marker of cell involved in the cellular cycle, yielding to a division into two cells. The fact that a division occurs or not is a sign of activity. In Figure 47 we can clearly see that the further away we go from the vessel the lower is the number of cells, and the bigger is the number of activated cells by marker KI67, especially between iteration 9-15. This information related with the one that we obtained from the previous results, leads us to suppose that these cells are mostly Double negative lymphocytes, and this result is concrete, since these cells tend to proliferate, so they will be the most likely to divide and indeed be active.

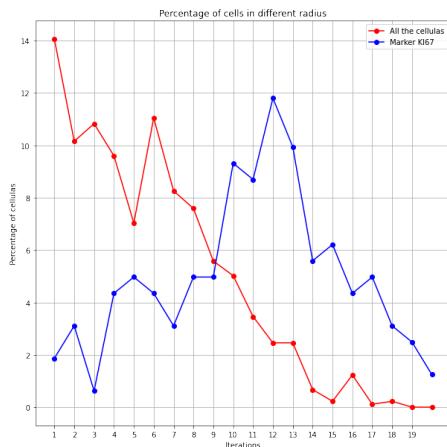


Fig. 47: Comparison between total number of cells and KI67

Another interesting data for our analysis discovered by Han, S. (2014) in «Tumour-Infiltrating CD4(+) and CD8(+)

Lymphocytes as Predictors of Clinical Outcome in Glioma», is the number of the cells activated by CD4+ and CD8+, indeed they are associated with the survival probability of the patient. The presence of high CD4+ and low CD8+ levels was an independent predictor of poor progress-free survival, the combination of CD4+ and CD8+ is a predictor of clinical outcome in glioblastoma patients, and a high level of CD4+ combined with low CD8+ was associated with unfavourable prognosis.

After counting the cells activated by the two markers in our image, we see that:

Table	CD4+	CD8+
Total number	207	104

TABLE 3: SETTED PARAMETERS

CONCLUSIONS AND CRITICAL ANALYSIS

From our image we can clearly see that there is an immune response is underway against the glioma, the most obvious indications of these are that Double negative lymphocytes are the most active cells, Macrophages follows Double negative lymphocytes distribution and Auxiliares lymphocytes decrease in module and percentage! A more detailed picture of this scenario, containing numerical values, is shown in Figure 48.

Iteration	% Lymphocytes double-negative	% Lymphocytes auxiliaries	% Lymphocytes cytotoxic	% Macrophages	% Lymphocyte B	% Lymphocyte T
1	21.18210180210383	27.44027402740273	31.08610001000101	31.37172102172102	0.8100010001000100	1.4900010001000100
2	20.8	26.8	31.14200001000100	30.14200001000100	0.8100010001000100	1.4900010001000100
3	20.50000000000000	27.33000000000000	31.07100001000100	30.17100001000100	0.8090010001000100	1.4890010001000100
4	20.30000000000000	27.13000000000000	31.05100001000100	30.15100001000100	0.8080010001000100	1.4880010001000100
5	20.10000000000000	26.93000000000000	31.03100001000100	30.13100001000100	0.8070010001000100	1.4870010001000100
6	19.90000000000000	26.73000000000000	31.01100001000100	30.11100001000100	0.8060010001000100	1.4860010001000100
7	19.70000000000000	26.53000000000000	31.00100001000100	30.09100001000100	0.8050010001000100	1.4850010001000100
8	19.50000000000000	26.33000000000000	30.98100001000100	30.07100001000100	0.8040010001000100	1.4840010001000100
9	19.30000000000000	26.13000000000000	30.96100001000100	30.05100001000100	0.8030010001000100	1.4830010001000100
10	19.10000000000000	25.93000000000000	30.94100001000100	30.03100001000100	0.8020010001000100	1.4820010001000100
11	18.90000000000000	25.73000000000000	30.92100001000100	30.01100001000100	0.8010010001000100	1.4810010001000100
12	18.70000000000000	25.53000000000000	30.90100001000100	29.99100001000100	0.8000010001000100	1.4800010001000100
13	18.50000000000000	25.33000000000000	30.88100001000100	29.97100001000100	0.7990010001000100	1.4790010001000100
14	18.30000000000000	25.13000000000000	30.86100001000100	29.95100001000100	0.7980010001000100	1.4780010001000100
15	18.10000000000000	24.93000000000000	30.84100001000100	29.93100001000100	0.7970010001000100	1.4770010001000100
16	17.90000000000000	24.73000000000000	30.82100001000100	29.91100001000100	0.7960010001000100	1.4760010001000100
17	17.70000000000000	24.53000000000000	30.80100001000100	29.89100001000100	0.7950010001000100	1.4750010001000100
18	17.50000000000000	24.33000000000000	30.78100001000100	29.87100001000100	0.7940010001000100	1.4740010001000100
19	17.30000000000000	24.13000000000000	30.76100001000100	29.85100001000100	0.7930010001000100	1.4730010001000100
20	17.10000000000000	23.93000000000000	30.74100001000100	29.83100001000100	0.7920010001000100	1.4720010001000100

Fig. 48: Total percentage

On the other hand, the presence of high CD4+ and low CD8+ levels is associated with unfavourable prognosis, this is probably due to an immune response that is not efficient enough.

Although the conclusions are consistent with the results and the literature on the subject, they cannot be relied upon completely, firstly the analysis is limited by the fact that we have only one image, it would be necessary to have a series of images captured at a certain time interval from each other in order to draw more concrete conclusions. Secondly, conclusions drawn by doctors or specialists would certainly be more reliable.

For better and more general results the best method would be to use CNN for segmentation and classification, unfortunately with a single image this was not possible, UNET could have been used, but datasets of fluorescent images of brain gliomas are not common. Other networks that are widely used for finding the centroids of cells are FCRN-A and FCRN-B, unfortunately also for these two network the dataset available (with source and target) are only of histological image. For these reasons only mathematical morphology was used.

The use of specific software would also have improved performance, as using centroids can lead to some errors in the calculation of cells due to cells that are too close together being counted as one. A check concerning the correctness of the calculated centroids with respect to the real ones was made with CD4+ and CD8+ and led to these results:

Table	CD4+	CD8+
Estimated	242	94
Real	207	104

TABLE 4: CHECK IN THE COUNTED CENTROIDS

It should be noted that, as expected, the results can also be wrong by quite a bit, as in the case of CD4+, depending on how close the cells are.

Such software could for example be Icy or Ilastick, some tests were made for cell detection with Icy and they have shown greater accuracy, the results are visible in Figure 49.

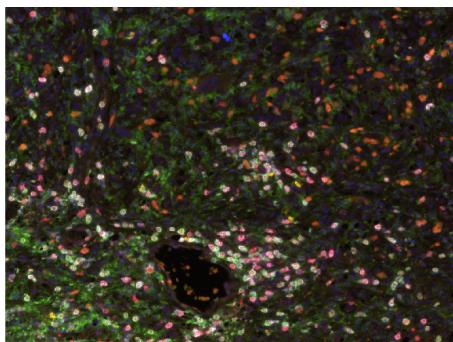


Fig. 49: Detection with Icy

Finally, also the use of stain separation may lead to better result in the thresholding.

The distances of the centroids of the cells to those of other classes were also calculated, making averages and variances per radius. However, the values did not help to conclude anything as there were comparison metres and most of the outcomes can be seen more easily graphically than numerically. An example for the first iteration is shown in Figure 50

Mean Values at Iteration 0	Lymphocytes double negative	Lymphocytes available	Lymphocytes cytoxenotes	Neutrophages	Lymphocytes B	Lymphocytes T
Lymphocytes double negative	0	196.305	197.430	187.237	299.818	207.371
Lymphocytes available	196.305	0	177.614	179.912	281.810	198.328
Lymphocytes cytoxenotes	197.430	177.614	0	189.314	388.936	198.393
Neutrophages	187.237	179.912	189.314	0	271.257	206.820
Lymphocytes B	299.818	281.810	388.936	271.257	0	279.818
Lymphocytes T	207.371	198.328	198.393	279.818	271.257	0

Var Values at Iteration 0	Lymphocytes double negative	Lymphocytes available	Lymphocytes cytoxenotes	Neutrophages	Lymphocytes B	Lymphocytes T
Lymphocytes double negative	11379.0	11071.2	11527.7	9780.8	11371.9	11371.9
Lymphocytes available	11379.0	11168.0	11477.3	10479.7	11371.9	11371.9
Lymphocytes cytoxenotes	11071.2	11290.8	0	11071.2	11370.8	11370.8
Neutrophages	11527.7	11077.3	11209.7	0	11008.5	11008.5
Lymphocytes B	9780.8	10379.7	10576.0	11848.5	0	11714.7
Lymphocytes T	11371.9	10876.0	11161.4	10874.7	11054.8	0

Fig. 50: Mean and Variance of distances for iteration 1

REFERENCES

- [1] Bae S. (2021). «Discovery of molecular features underlying the morphological landscape by integrating spatial transcriptomic data with deep features of tissue images». Nucleic Acids Research, vol. 49, n. 10, pagg. e55–e55..
- [2] Dubois L. G. (2014). «Gliomas and the vascular fragility of the blood brain barrier». Frontiers in Cellular Neuroscience, vol. 8, 2014. Frontiers.
- [3] Han, S. (2014). «Tumour-Infiltrating CD4(+) and CD8(+) Lymphocytes as Predictors of Clinical Outcome in Glioma». British Journal of Cancer, vol. 110, n. 10, pagg. 2560–68.
- [4] He K. (2018). «Mask R-CNN». arXiv.org.
- [5] Kurt T. (2020). «Segmentation and Classification in Digital Pathology for Glioma Research: Challenges and Deep Learning Approaches». Frontiers in Neuroscience, vol. 14, Frontiers.
- [6] Martin-Gonzalez P. (2021). «Predictive Modelling of Highly Multiplexed Tumour Tissue Images by Graph Neural Networks». medRxiv.
- [7] Partel G. and Wählby C (2020). «Spage2vec: Unsupervised representation of localized spatial gene expression signatures». The FEBS Journal, vol. 288, n. 6, pagg. 1859–70. Wiley Online Library.
- [8] Po-Whei H. and Lai Y. (2010). «Effective Segmentation and Classification for HCC Biopsy Images». Pattern Recognition, vol. 43, n. 4, pagg. 1550–63. ScienceDirect.
- [9] Ronneberger O. (2015). «U-Net: Convolutional Networks for Biomedical Image Segmentation». Medical Image Computing and Computer-Assisted Intervention, Springer International Publishing, pagg. 234–41. Springer Link.
- [10] Solorzano L. (2019). «Machine Learning for Cell Classification and Neighborhood Analysis in Glioma Tissue». Cytometry Part A, vol. 99, n. 12, pagg. 1176–86. Wiley Online Library.
- [11] Sung H. (2021). «Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries». A Cancer Journal for Clinicians, vol. 71, n. 3, pagg. 209–49. PubMed.
- [12] Tian Z. (2018). «PSNet: prostate segmentation on MRI based on a convolutional neural network». Journal of Medical Imaging, vol. 5, n. 2, PubMed Central.
- [13] Yapp C. (2021). «UnMICST: Deep Learning with Real Augmentation for Robust Segmentation of Highly Multiplexed Images of Human Tissues». bioRxiv.
- [14] Zanotelli V. and Damond N. (2019). «A flexible image segmentation pipeline based on pixel classification for heterogenous multiplexed tissue images». Portal Universitad Bern.