Cytology Challenge:

Segmentation of Multiple Myeloma Cells

Technical report

Serena Donadio, Ilaria Facchi, Giuseppe Parisi, Burçak Carlo Pasqua

1. Abstract

Multiple Myeloma (MM) is a type of cancer that produces an abnormal growth of plasma cells in the bone marrow. Early-stage detection is fundamental to have a highly effective treatment. The golden standard for detecting MM cells is the visual microscope inspection executed by an expert, but this method is time-consuming and subject to human errors. In this project, a fully automated, deep learning-based system for the segmentation of MM plasma cells was developed. In particular, the system consists of an ensemble of three U-net deep neural networks' predictions. The system's performance was evaluated using the Intersection over Union (IoU) metric. Our system obtains an IoU on the training set of 0.84 (nuclei), 0.73 (cytoplasm) and 0.82 (whole cell) while on the validation set 0.67 (nuclei), 0.59 (cytoplasm) and 0.67 (whole cell). As expected, the segmentation of the cytoplasm was the more challenging, and thus no notable results were obtained.

2. Introduction

Multiple Myeloma is a type of cancer characterized by the accumulation of plasma cells in the bone marrow, and the treatment could be very effective if the diagnosis is made early. The most frequent method used to diagnosticate the MM is the bone marrow aspiration which permits the successive visual inspection at the microscope, in which an expert pathologist estimates the percentage of MM plasma cells. This technique is still considered the golden standard for MM cells detection, even though it is time-consuming and subject to human errors. An automated system that reduces time and improves detection accuracy could assist the pathologist's work[1].

A deep learning neural network that implements the segmentation of plasma cells based on pixel-level segmentation can be used. To detect plasma cells, both the nucleus and cytoplasm must be segmented. Achieving accurate and robust segmentation of myeloma cells is though still challenging; there are indeed several problems: the cytoplasm is not well distinguishable from the background, the cells are sometimes isolated and sometimes clustered, there is more than one type of unstained or stained cells, and there are different levels of clustering (nucleus-nucleus cluster, nucleus-cytoplasm cluster, and cytoplasm-cytoplasm cluster).

Our proposed method is an automated detection system based on an ensemble deep learning neural network. The system combines the probability maps obtained by three U-net models.

3. Methods

3.1 Dataset

The developed system uses a dataset obtained from the cytology challenge archive. The dataset consists of 450 RGB images, divided as follow:

- 300 RGB images for the training dataset;
- 50 RGB images for the validation dataset;
- 100 RGB images for the test dataset.

These images were captured from bone marrow aspirate slides of patients with MM. Slides were stained using Jenner Giemsa stain. For train and validation, ground truth masks were provided. The masks were obtained from an expert pathologist's manual segmentation that identified and marked the nucleus and cytoplasm of MM cells. The images and manual segmentations are provided in TIF format encoded in uint8. Images can have a size equal to 1536x2040x3 or 1920x2560x3.

3.2 Pre-processing

The first step of pre-processing was resizing the images to 512X512. This step was necessary because the limitations of computational resources of the used tools. The change in input resolutions increases the model speed and can be done automatically, but it may distort the results and decrease the accuracy. Moreover the resizing step is a quick method to deal with the difference in images dimensions.

Secondly, the images were color normalized. The main objective of color normalization in histopathology images is to reduce color variation among a set of source images transferring desirable color from the reference image to the source images. We used the Reinhard method, which is a global color normalization method[2]. The color normalization was implemented using algorithms present in the *HistomicsTK* package[3]. Specifically, we chose "104. tif" from the training set as target image. This image showed a good contrast for visually distinguishing the three components of the patch image: cytoplasm, nucleus, and background. We calculated the mean and the standard deviation values of the reference images in the lab space. The transformation of the images from the RGB space to the lab space is necessary because there is a considerable correlation among R, G, and B channels in RGB space, while in the lab space, there will not be any color mixing. After that, we normalized the images (figure 1). The source image is firstly converted to Ruderman's LAB space and then the LAB channels are centered and scaled to zero-mean unit variance and then rescaled and shifted to match the target image statistics[4]. The result is that the background color and luminance of source images are replaced with that of the target image. The main drawback is that this method sometimes produces a poorer contrast color normalized image if the target image contrast is lesser than that of the source image.

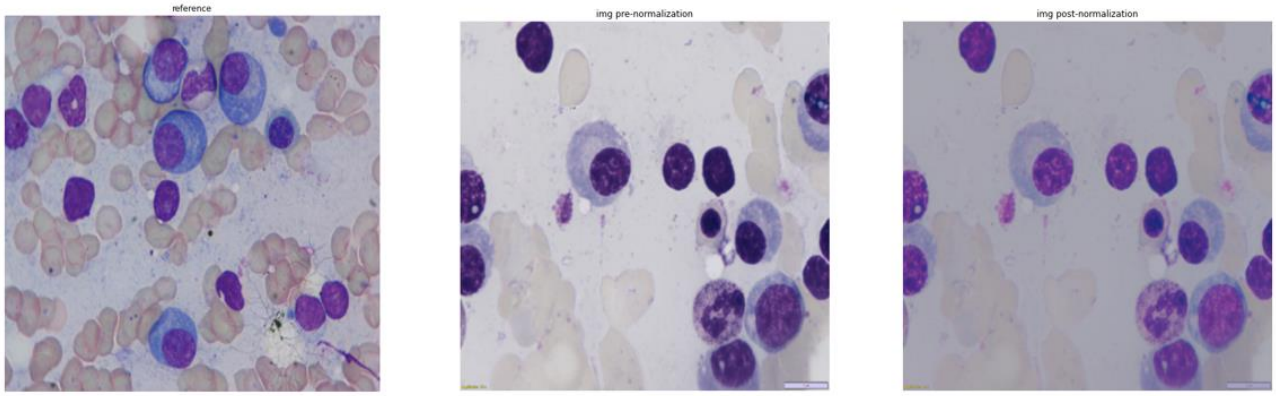

Figure 1: reference image 104.tif on the left, 1743.tif at the center, and normalized 1743.tif on the right

Regarding the masks, as can be observed in figure 2 and figure 3, some ground truth masks included single pixels or little spots erroneously annotated or were characterized by small holes in the nucleus that do not reflect the cell's proper aspect. These mis-segmentations were thought to possibly decrease the quality of the training, therefore they have been removed using morphological opening and closing with a defined area threshold of 0.0010*total mask area.

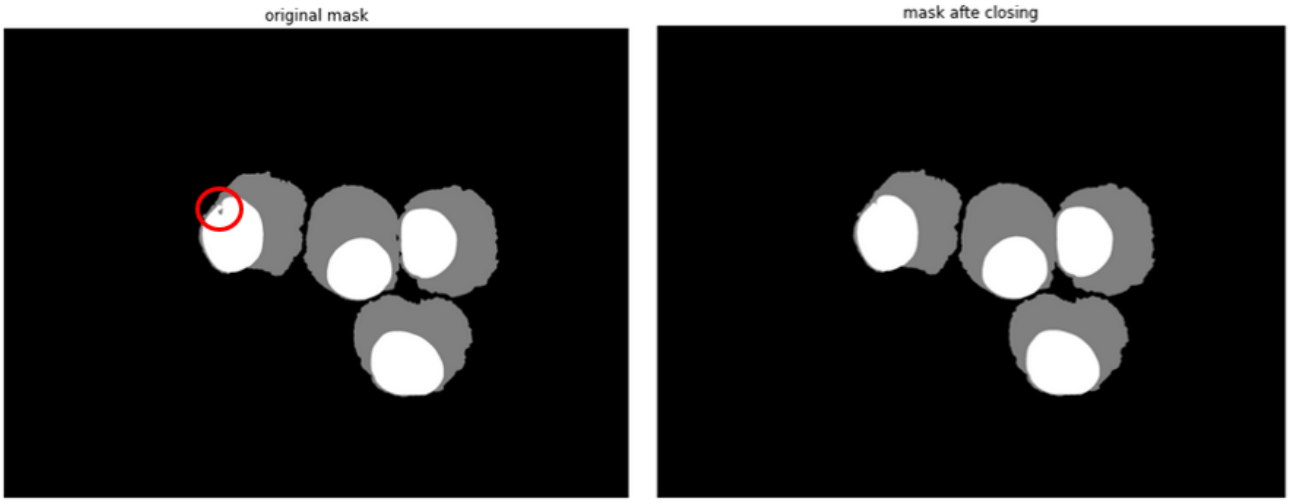

Figure 3: result of morphology closing on image 2352.tif

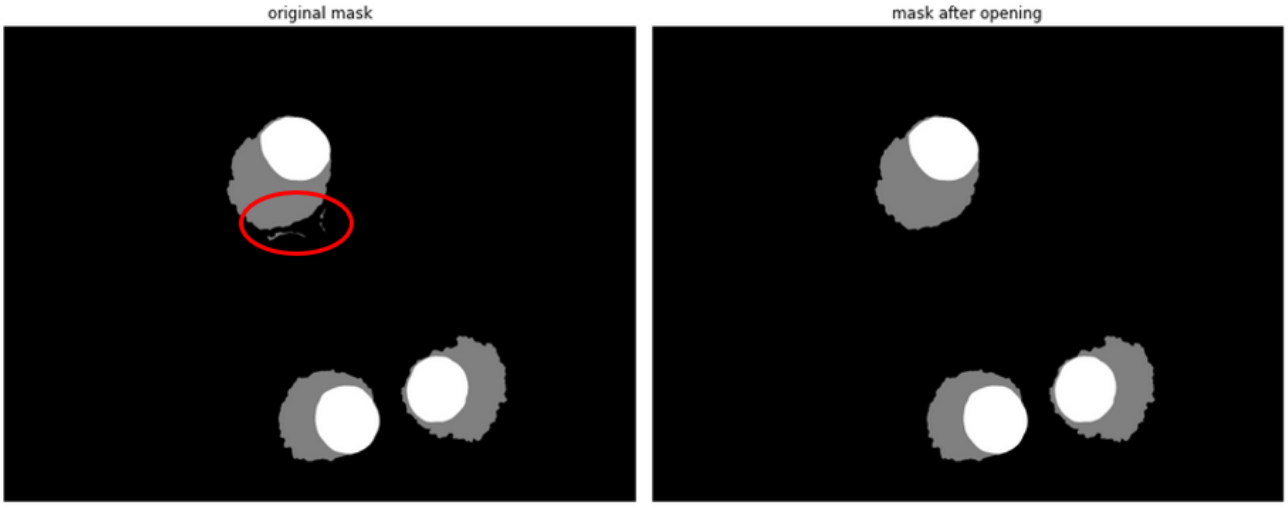

Figure 2: result of morphology opening on image 2306.tif

The provided dataset was unbalanced in terms of classes representation (5.7% nucleus, 7.6% cytoplasm, and 86.7% background), but this was not thought to be a problem because it reflects the intrinsic unbalancing of classes in this kind of images, therefore the bias introduced in the pixel-wise classification as one of the three classes (nucleus, cytoplasm and background) has not a bad effect on the performances of the system (the risk of balancing the classes is to perform an over-segmentation finding too many erroneous cells).

3.3 Training of the model

The U-net model is a symmetric network with skip connections between down-sampling and up-sampling paths with concatenation (figure 4). The skip connections help in passing the local information from down-sampling stages to global information in the up-sampling stages, giving a combination of local and contextual information to better perform segmentation. Making different tests we found that the densenet201 backbone was the one generally giving better performance. In order to obtain a faster and more accurate model's convergence, we used a transfer learning approach loading in our models the pre-trained weights of the ImageNet database.

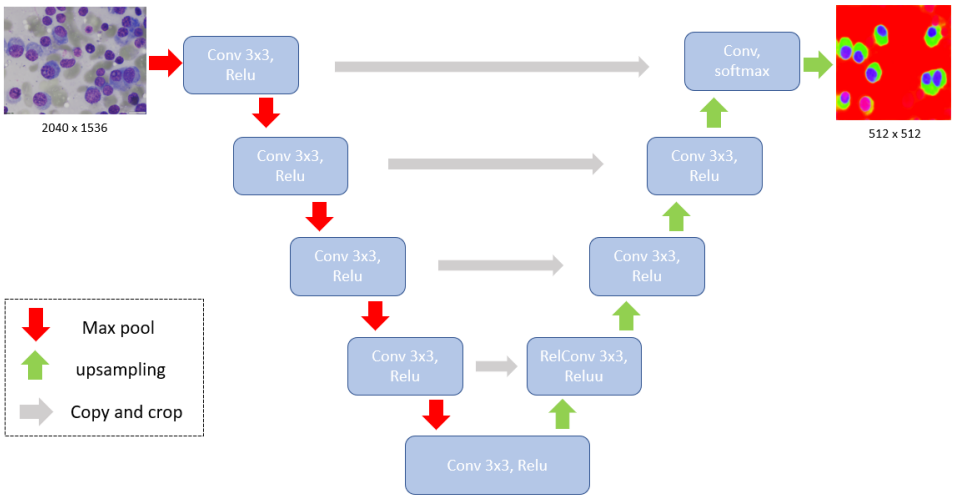

Figure 4: UNet model

A multi-start optimization approach has been used to obtain better results and test the robustness of pre-processing and training pipeline. The models were trained for 20 epochs using the Adam optimization with a learning rate of 1e−3. All training trials have been done using as loss function the cross-entropy and as evaluation metric the categorical accuracy. During training, an early stopping criterion was implemented to monitor the validation metric over the epochs and to stop the training in case of premature convergence. In particular, the training stopped if the categorical accuracy did not show any improvement in the range of 1e-3 in 10 epochs. At the end of the training, the best epoch weights were restored. In order to solve the overfitting problem, image augmentation has been carried out. Different kinds of data augmentation were combined: geometric (vertical-flip, horizontal-flip, rotation, zoom), color (brightness), and fine-tuning on these augmentation parameters have been performed. Data augmentation aims to introduce variability in the training dataset and thus improves the generalization capabilities of the trained system. Also, trials without any data augmentation have been tried, obtaining in general very good performances on the training set (IoU scores always greater than 0.8) but these trained models showed a very poor generalization capability, therefore the performance on validation dataset showed an important decrease with respect to those on the training data (IoU scores nearly around 0.5).

3.4 Semantic prediction of net and post-processing

To obtain an initial semantic segmentation, starting from the heatmaps associated with each class predicted by the neural network, simple thresholding has been used for each conceptual class of interest (i.e. 'nucleus' and 'cytoplasm,' while the background has been automatically assigned to pixels not identified as such two previous classes), in particular threshold value has been varied in [0.30, 0.60] range in a paired fashion (i.e., $threshold_{nuclei}$ = 0.3 - $threshold_{cytoplasm}$ = 0.3; $threshold_{nuclei}$ = 0.35 - $threshold_{cytoplasm}$ = 0.35; etc.) at steps of 0.05

and the value giving the best performances was chosen. Notice that in this thresholding phase, even though the softmax function was set as the activation function of the model, since threshold values smaller than 0.5 have been tried (for each pixel the prediction of the class is not mutually exclusive in these cases), priority on the semantic segmentation was given to the 'nucleus' class, having observed that in general all the constructed nets performed better on this one.

An ad hoc post-processing (for the best-obtained net at the thresholding step) has been implemented starting from this initial mask. The post-processing consisted of an initial area closing (with an area threshold value of 200 pixels) followed by an area opening (with an area threshold value of 100 pixels) in order to both delete small nuclei and cytoplasm objects (having observed that these were associated with erroneous predictions) and fill small nuclei and cytoplasm holes. Subsequently resizing to original image shape has been performed- Notice that performing resizing at this stage is the best choice with respect to doing it on probability maps (before thresholding) since up-sampling through interpolation on few possible data values (conceptually categorical data: 0 → 'background', 128 → 'cytoplasm', 255 → 'nucleus') is likely to not degrade predictive information returned by the net; while this could more likely occur when up-sampling probability maps. Lastly, marker-based watershed by flooding has been applied in case of detection of clustered cells. For each connected object in the mask, the number of nuclei objects has been calculated, and if more than one was present, then this meant that the object was a cluster of cells and not a single cell; thus a marker in correspondence of the local maxima of the distance transform of each nucleus contained in the connected object was set, and watershed by flooding algorithm using these markers was applied on the negated distance transform of the connected object to separate cells of the cluster. Observing that the watershed sometimes cut the nuclei of the watersheded object, a restoring nuclei stage has been queued in the post-processing pipeline. Lastly all masks containing only cytoplasm were deleted by the moment that the minimal requirement for a cell to be segmented as such is to have a nucleus, moreover we found that these predictions (masks with only cytoplasm) were always wrong. In figure 5 a schematic flowchart of the post-processing pipeline is shown.

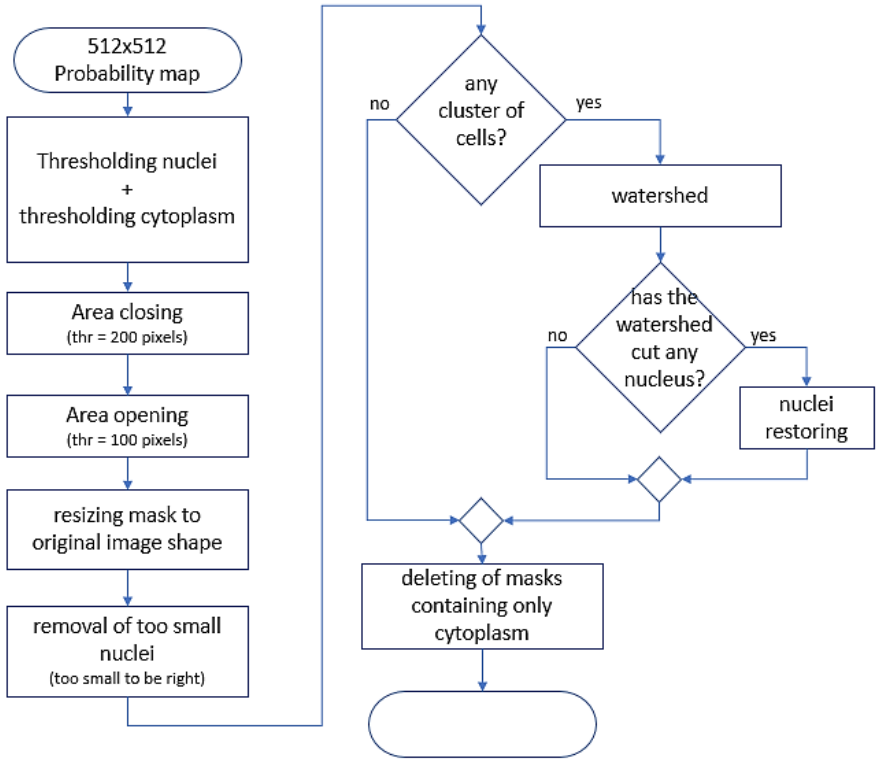

Figure 5: post-processing pipeline

3.5 Ensembling

In order to try to further improve the performance of the system, an ensembling approach has been tried. We chose three best-performing trained neural networks and combined their predictions at probability map level (probability maps associated with each class have been simply summed). The best thresholding value has been then evaluated similarly to the case of the single net prediction, but now varying the value in the range [1.00, 1.70] (always at steps of 0.05). As post-processing stage, we maintained the very same pipeline as in the case of the single neural network prediction.

The three nets to ensemble has been selected as the two best performing nets both on training and validation data trained on augmented data and as the best performing net both on training and validation data trained on non-augmented data. As said before these nets trained on non-augmented data generally perform very well on training data and much worse on validation data, being overfitted on the former; but we decided to choose also such a net trained on non-augmented data because with multi-start optimization approach we managed to obtain a net performing very well on training data and also quite well on validation data. Thus aware that acceptable performance on validation data could possibly be owed to a sort of "overfitting" also on this division, but that on test data this could perform much worse than the generally expected decrease of performance between validation data and test data, we accepted to use such a chancy net only in the context of ensembling, which by nature could eventually mitigate this possible drawback.

We found ensembling slightly improving performances on the validation set with respect to the best single net prediction. On the training set, this did not occur, indeed performances on this division had a slight decrease due to the high performance of the net trained on non-augmented images on training data. Nevertheless in this context we considered more important the performances on the validation set, being these more representative of the real case usage of the system (performing well on already known masks is not of practical interest); furthermore using only a net trained on non-augmented data, even if excellently performing on these and acceptably performing on validation data, was not a considerable choice to us for the previously mentioned reasons. In addition ensembling approach is intrinsically likely to provide more robustness to the performance of the system. For all these reasons we decided to build such an ensemble system.

From a computational load point of view during the usage (assessed in terms of computational time), the implemented kind of ensemble only minimally overloaded the system by the moment that the heaviest part of the system pipeline is the post-processing.

4. Results

We evaluated the performance of the system at a single-cell level (in order to have the most complete and restrictive assessment → also the goodness of the watersheding is evaluated) using the intersection over union score

$$IoU(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

(being X the single-cell manual mask and Y its respective single-cell predicted)

separately on the three objects nucleus (IoU_{nuclei}), cytoplasm ($IoU_{cytoplasm}$), and whole cell (IoU_{cell}). To do this, the matching between single-cell manual annotation and single-cell automatic mask has been made through the IoU_{cell} score. Moreover, we used three other metrics concerning the difference in the number of identified MM cells between the two types of masks, both to have a better understanding of the behavior of the system and to more smartly perform tuning of training, threshold value and post-processing parameters. For each predicted mask, we calculated the number of "invented cells" (i.e., the number of cells segmented by the system that were actually not segmented in the manual annotations), the number of "missed cells" (i.e., the number of MM cells that were segmented in the manual annotations but that were not found by our system), and the difference in the number of MM segmented cells (simply obtainable as the absolute value of the difference of the previous two). The number of "invented cells" and "missed cells", conjunctly with the observation of the IoU scores at single-cell level (pre-grand average across all cells), permitted us to see that, in general, the major problem in the predictions using a U-net model is that some cells not to be segmented are segmented and others to be segmented are not segmented at all. In general, when an MM cell is correctly identified, the IoU scores on the cell are quite good. The number of "invented cells" and "missed cells" give more accurate information with respect to the simple difference in the number of MM segmented cells being of major interest in evaluating the system performance from an absolute point of view and in case the segmentation of MM cells is done to subsequently perform further kind of assessment only on these

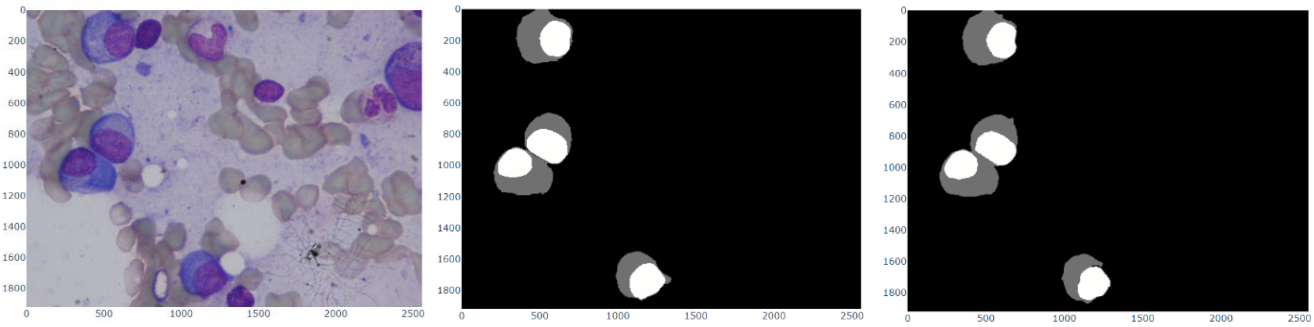

Figure 6: original mask (left), manual mask (center), predicted mask (right) of image 110.tif

segmented cells (i.e. there will be a further system processing only portion of images segmented from this system). However the coarse difference of found MM cells can also be of practical interest when the assessment of the patient is done directly on the result of this kind of system (for example assessing the percentage of MM cells on a whole slide of dissected tissue), indeed in this case "invented" and "missed" can balance each other.

The performances of the three ensembled components (prior to ensembling) and of the ensembled system are shown in tables 1, 2, 3, and 4 as mean ± standard deviation for IoU scores and as only mean for number of cell scores. To obtain system 1, no data augmentation was applied during the training while in the other two, two different types of data augmentation were applied. For system 2, the data augmentation consisted of horizontal flip, vertical flip, and zoom range of [0.8,1,2], while for the system 3 horizontal flip, vertical flip, and brightness range [0.9,1.5] were applied.

SYSTEM 1	IoU nuclei	IoU cytoplasm	IoU whole cell	invented cells	missed cells	# of diff *
training set	0,88 ± 0,25	0,79 ± 0,24	0,86 ± 0,25	0,14	0,24	0,31
validation set	0,65 ± 0,43	0,57 ± 0,40	0,63 ± 0,43	1,34	0,64	1,30

Table 1: performances of system 1

SYSTEM 2	IoU nuclei	IoU cytoplasm	IoU whole cell	invented cells	missed cells	# of diff *
training set	0,71 ± 0,36	0,59 ± 0,36	0,69 ± 0,37	0,67	0,73	0,91
validation set	0,63 ± 0,40	0,57 ± 0,38	0,63 ± 0,40	0,92	0,88	1,00

Table 2: performances of system 2

SYSTEM 3	IoU nuclei	IoU cytoplasm	IoU whole cell	invented cells	missed cells	# of diff *
training set	0,69 ± 0,34	0,62 ± 0,32	0,72 ± 0,35	0,64	0,60	0,82
validation set	0,6 ± 0,38	0,55 ± 0,37	0,64 ± 0,41	1,04	0,70	1,10

Table 4: performances of system 3

ENSEMBLE	IoU nuclei	IoU cytoplasm	IoU whole cell	invented cells	missed cells	# of diff *
training set	0,84 ± 0,28	0,73 ± 0,28	0,82 ± 0,28	0,30	0,30	0,46
validation set	0,67 ± 0,41	0,59 ± 0,39	0,67 ± 0,40	1,00	0,66	1,02

Table 3: performances of ensemble system

*Number of different segmented cells

5. Discussion

Generally speaking, our system does not perform well in the semantic segmentation task, and this due to sub-optimal, but kind of "forced", choices in our pipeline.

Firstly, we performed only a poor color normalization (only able to slightly standardize images between each other), and due to images of tissue possibly being stained with two or three stains (depending on case in case), color deconvolution was not easily doable in order to perform a better color normalization or to access to higher quality information contained on colour deconvoluted channels.

Secondly, as mentioned in the Result section, poor performances are mainly due to "invented" and "missed" instances of MM cells, respectively plausibly attributable to (at least in part) not so good manual mask (for example, in some cases, actual MM cells at the border of the patch have not been annotated, but the net recognize and segments them) for what concerns "invented" cells; and due to low spatial resolution of images fed to the net reasonably usable with available tools (starting images were of 1536x2040 pixels or 1920x2560 pixels and had to be resized to 512x512 pixels → nearly a four time worse spatial resolution) in order to obtain predictions. This information loss could have led to miss some MM cells. Thus, likely using more heavy models (allowing greater input size of images) could improve performances with respect to what we obtained by the moment that a greater spatial resolution could be maintained. Considering only correctly found MM cell (i.e. when a minimal matching between manual and predicted masks has been found) the system performance are IoU_{nuclei} = 0.93, $IoU_{cytoplasm}$ = 0.81, IoU_{cell} = 0.90 on training data and IoU_{nuclei} = 0.91, $IoU_{cytoplasm}$ = 0.80, IoU_{cell} = 0.90.

In addition our system systematically fails in segmenting overlapping cytoplasms due to how the problem has been faced, indeed in case of manual annotations it could happen that two or more cells have overlapping cytoplasms, while this is not possible a priori due to the very simple clustered cell separation performed through watershed. To overcome this problem a more structured solution should be implemented, using an additional class for overlapping cytoplasm areas.

Lastly, we have observed that in literature state-of-art approach to perform MM plasma cells segmentation is to use other models such as mask R-CNN (Recurrent Convolutional Neural Network), HTC (Hybrid Task Cascade) and SCNet (Semantic Correspondence), which lead to performance over the 90% in terms of IoU on the test set. [5]

[1] V. M. T, G. E. A, S. v Variyar V, V. Krishna Menon, and S. K. P, “Deep Learning Based Approach For Multiple Myeloma Detection.”

[2] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, Sep. 2001, doi: 10.1109/38.946629.

[3] “https://digitalslidearchive.github.io/HistomicsTK/.”

[4] S. Roy, S. Panda, and M. Jangid, “Modified Reinhard Algorithm for Color Normalization of Colorectal Cancer Histopathology Images,” in *European Signal Processing Conference*, 2021, vol. 2021-August, pp. 1231–1235. doi: 10.23919/EUSIPCO54536.2021.9616117.

[5] Á. G. Faura, D. Štepec, T. Martinčič, and D. Skočaj, “Segmentation of Multiple Myeloma Plasma Cells in Microscopy Images with Noisy Labels,” Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.05125