Implementing Deep Learning in Cell Cycle Regulation

Cell Cycle Regulation – a project for Instituto Gulbenkian Ciência

Ricardo Fraga Simões | 93674

Vicente Sobral | 102134

Applications of Data Science and Engineering

Instituto Superior Técnico

February 2022

Table of Contents

Introduction

Nowadays, with the proliferation of data science among most scientific fields, its contribution to medical sciences is becoming crucial to the development of some areas of study. Instituto Gulbenkian de Ciência (IGC) has been, since its foundation, a major hub of scientific investigation, dealing in their research programs with relevant biological and biomedical topics.

This project integrates the lab of Cell Cycle Regulation, an IGC program that analyses variations in the stages of development, evolution and disease of cells. One of the major steps within the Cell Cycle Regulation lab deals with the detection and counting of centrosomes, that are clusters of centrioles, small microtubular structures present in vertebrate cells. Centrioles, responsible for cellular division, exist in an abnormally high number on cancer cells. To obtain relevant results, centrioles must be detected, counted and have its intensity measured from immunofluorescence microscopy images, a process which, due to the scale and resolution of the image, is done manually, thus being very time consuming for the investigators.

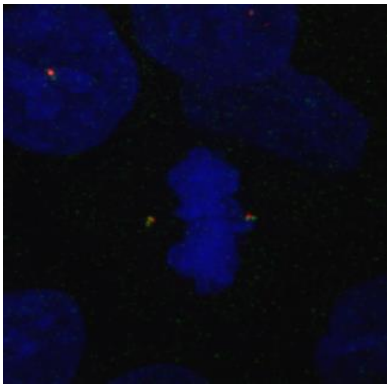
Therefore, our project has the goal of partially or totally automatizing, by using state-of-the-art object detection machine and deep learning tools, this identification and count of centrosomes and centrioles, a process that contributes to an investigation that is already having some results towards the early detection and medication of some forms of cancer.

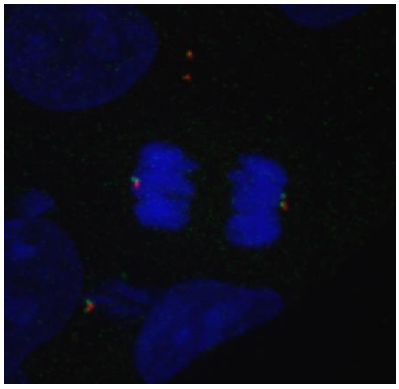
Materials and Methods

Images and traditional approach

Immunofluorescence imaging is a method that is frequently used in microbiological research and that consists in the highlighting of specific targets within a cell with the use of antibodies. Both sets of images that we worked with received the same fluorescence treatment: cells' nuclei were marked by DAPI, the blue stains that can be seen in the images below, whereas green stains and magenta stains marked respectively the centrin protein and the protein CEP135, both obtained using antibodies.

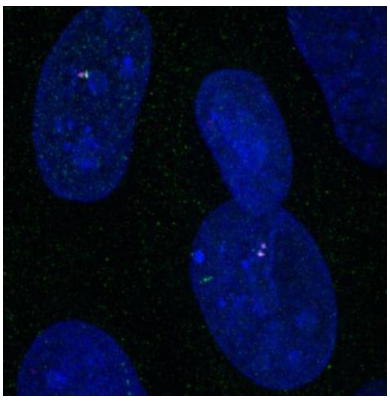
Our research focused on two types of cells: mitotic and interphase cells. On one hand, mitotic cells are very specific in its nature because they are in a state of mitosis, which represents its division into two identical cells. Therefore, their centrosomes are generally arranged in four different states, with the two most common being present in the image details below (Fig. 1 and 2). One of the two sets of images contains samples with at least one mitotic cell, but they may also contain some neighbor cells in a non-mitotic state, and in this type of images we are only interested in detecting the mitotic cells' nuclei and their centrioles.

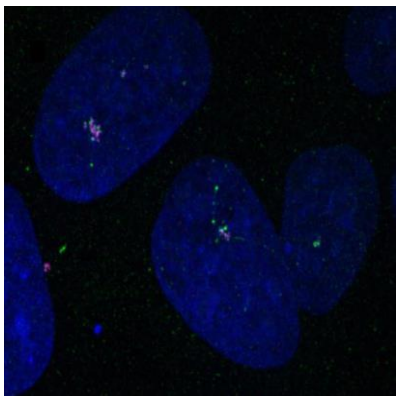




Figures 1 and 2: Mitotic cells in different stages of division. Details from the Instituto Gulbenkian Ciência images

On the other hand, interphase cells are cells in a state where they spend most of their lifetime. The other set of images contains cells only in an interphase stage, so for this set of images we will detect all the cells' nuclei and centrioles. An example of an image of this set can be seen in Fig. 3 and 4.





Figures 3 and 4: Interfase cell example. Details from the Instituto Gulbenkian Ciência images

The images obtained from the IGC immunofluorescence process have multiple nuclei each. These images are produced from the microscope in batches of 30, with each image coming in a TIFF format. The colour channels come separately in stacks of 81 images each, with the stack corresponding to the detailed vertical slicing along the cells. This is done to mitigate the loss of information due to the superposition of some of the cell elements. From there, a composite with the sum of the channels can be produced with the help of the ImageJ/Fiji software, which is currently used by IGC. After this process, the investigators routinely run a macro from the said software that creates a Max Projection – an image depicting the maximum intensity values from each pixel. The previous images are a result of this process.

Generally, IGC investigators resort to a naked-eye and totally manual analysis of the Max Projection image, from where they count the centrioles. However, for the first set of cells that we

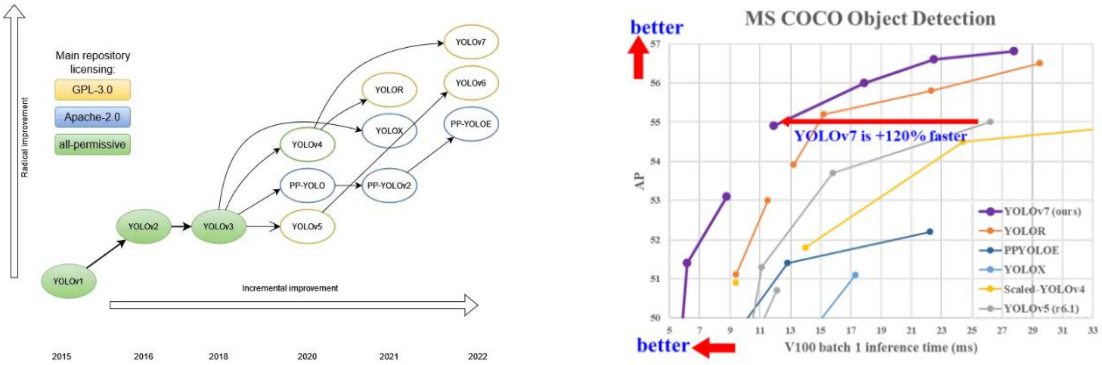
analyzed, each Max Projection can have more than a hundred centrioles, with multiple being mildly discernible due to the superpositions and the image quality, which makes this process very time consuming. For the mitotic cells' set, however, the number of centrioles subject to analysis is much smaller, since they only wanted to count the few centrioles that belong to the cells in the mitotic phase. Overlaps in these types of cells tend to be rarer, something that worked in favor of the approach we used.

For our project, we were initially given a small number of images: 10 images of cells in interphase and 25 images of mitotic cells. The numbers and image quality themselves present a limitation, but since this microscopy is very costly and the images belong to the investigators' direct line of work, that was the biggest amount that we could work with.

Computer Vision and state-of-art solutions in similar problems

The Computer Vision field is one of the most active research fields for deep learning applications. It is a very broad field that compresses many different tasks [1]. One particular task solved with such methods is the detection of objects (which can be humans, animals, objects). Therefore, the automatic identification of cellular structures can be tackled with Deep Learning methods suited for that purpose.

The YOLO family (You Only Look Once) is a class of models suitable for object detection, widely used nowadays for both academic and real-commercial projects of different industries. Such class of models surged in 2015, rapidly generating huge popularity. Since that year, multiple new versions have been released with the latest one being released in mid-2022 (see Figs. 5 and 6.)



Figures 5 and 6: YOLO model's family evolution, YOLO performance for the MS COCO dataset

The new models were the result of extensive research and experiments, which contributed to better results in object detection tasks and similar others. In particular, the performance of these models is generally measured by two factors: detection accuracy and inference times. The first one is related to mean average precision (mAP), which is a metric commonly used in object detection. See [2] for a good explanation of this metric. The last one is related to the amount of time it takes for a model to process test data and make a prediction. The image of Fig. 5 shows the values of these factors for some models of the YOLO family.

As one can see, the different YOLOv7 versions outperforms any previous YOLO version when trained in the MS COCO dataset [3].

Being the state-of-art model at the time of this project, we decided to use YOLOv7 for the experiments. Not only the mAP and inference times are better, but the number of parameters of the model is low - roughly 37 million - when compared to previous YOLO versions (like YOLOv4). More info about the model can be found in [4]. All experiments were done using Google Colab, which provides a free GPU in a cloud environment, which is an important feature for training Deep Learning models faster. The YOLOv7 implementation used is based in the original one, and it can be found in [5].

The YOLOv7 approach (Input Construction, Labelling, Evaluation Metrics, Transfer Learning)

The first thing we needed to do was to decide how to build the inputs for our YOLOv7 model, since we received 81 z-stacks for every colour channel for both mitotic and non-mitotic cells. To build an image out of the stacks we decided to apply a Max Projection to the stacks of each channel and merge them together, like the researchers at IGC usually do. Before providing any model with the necessary images and information for training and the consequent detection, we first needed to annotate some of the images with the labels of interest. Since the images of mitotic and non-mitotic cells are quite different, we trained two models in separate.

For the annotation, we used Roboflow, a computer vision website and app that features an annotation section, where we decided to work with our training material [6]. The process is one of drawing bounding boxes around the centrioles, which are the objects of interest. However, we did this only for the set of mitotic cells. For the non-mitotic cells, instead of training our models to detect centrioles, we trained it to detect centrosomes, which are sets of centrioles. The reason behind this will be explained in the results section. Therefore, for our images we added bounding boxes around each object of interest, and the process was not exhaustive because the number of images was small.

For mitotic cells we ended up labelling the centrioles of mitotic cells as “centriole” or “1” and some the cells’ nuclei as “cell” or “0”. For the mitotic cells, we only annotated the nuclei of the mitotic cells, while for the images of interphase cells we annotated all the nuclei because we only have interphase cells there.

Due to the microscopy image obtention concerns, we did not have too many images at our disposal. Therefore, a viable solution was doing data augmentation on the training, a process where we performed operations like rotation, crop, and add noise. Rotation allowed us to get centrioles (and cells’ nuclei) in many different positions, cropping allowed us to obtain images where centrioles occupy a greater area in the image. Noise was added because sometimes neural networks are very sensitive to pixel alterations or imperfections in images. For more info on this check [7],[8] and [9], respectively. The labelling software Roboflow allows us to create augmented images with their respective image label, but the number of images generated is limited since we only have access to the free version. Therefore, for the non-mitotic cells we left 6 images for training, 2 for validation and 2 for testing. Performing data augmentation allowed us to expand the training set to 18 images. For the mitotic cells, we left 17 images for training, 5 for validation

and 3 for testing. Since we had many more images, we performed cropping manually and labelled those cropped images on roboflow. That process allowed us to obtain more augmented examples of rotated images and images with noise. In the end, we obtained a training set with 51 images.

A common pre-processing procedure in Computer Vision applications such as object detection is to apply standardization or normalization on the images, thus converting the images to greyscale [10]. We experimentally verified that this procedure was not useful for our application due to the colour importance of our images. In fact, it is important for our models to know that cell nuclei is blue (and not any other colour), or that centrioles and centrosomes must have red and green proteins (centrins and the protein CEP135, respectively). Losing this type of information may result in worse results, which is something we encountered and that will be exemplified in the results section.

Instead of training the model from scratch, we trained over the weights of a YOLOv7 model already trained on the MS COCO dataset. This technique is known as Transfer Learning. Although this dataset does not contain centrioles or cells, the weights are already good for detecting shapes of many different objects of that dataset. Essentially, during training we fine-tuned the weights for detecting centrioles and cells.

First Experiment

Training Results

Interphase Cells

As previously mentioned, our initial objective was to detect centrioles individually for both types of cells. However, when analysing the images, we saw that there were many overlaps, and it would be impossible for the model to learn how to generalize with only 10 images (or 18 training samples after augmentation process). We found the model incapable of learning how to detect centrioles individually, as explained in the experience presented in the appendix. Therefore, here we simplified our approach by considering the detection of centrosomes instead of individual centrioles. Centrosomes are defined as a set of centrioles, and they are commonly found near the nuclei of the cells. To complement this analysis, we also computed the Raw Integrated Density (RID), which is a measure of the sum of all the pixel intensities for each colour channel (RGB).

Our results were, as requested by IGC, exported into an Excel file that depicts each cell and centrosome, establishing a relationship between both based on the distance to the centre of the nucleus of each cell. Alongside these values, we presented the RID of the relevant colour channels in the centrosome regions, computed in Python. This is a measure that sums the value from the intensities of each pixel comprised in a specific area, which provides the possibility to infer how many centrioles are comprised in each centrosome.

The training moment is the most relevant step of the whole process. Fig. 7 shows the training results obtained for the first experiment with cells in interphase. This image was automatically produced after training finished and it should be read as follows: The first row is related to values obtained in the training set, and the bottom row is related to validation set. The first three

columns are related to the YOLO loss function, which has three components – box (or regression) loss, objectness loss and classification loss. The four remaining graphs are the Precision, Recall, $\mathrm{Map}_{0.5}$ and $\mathrm{Map}_{0.5:0.95}$ on the validation set; for a more detailed explanation see [4]. The results show that the model converged to a fairly good solution around the 1000th epoch. Precision and recall stabilized around 86% and 84%, respectively. The values of $\mathrm{Map}_{0.5}$ and $\mathrm{Map}_{0.5:0.95}$ stabilized around 83% and 46%, respectively. Furthermore, the loss components achieved minimal values both on the training and validation sets.

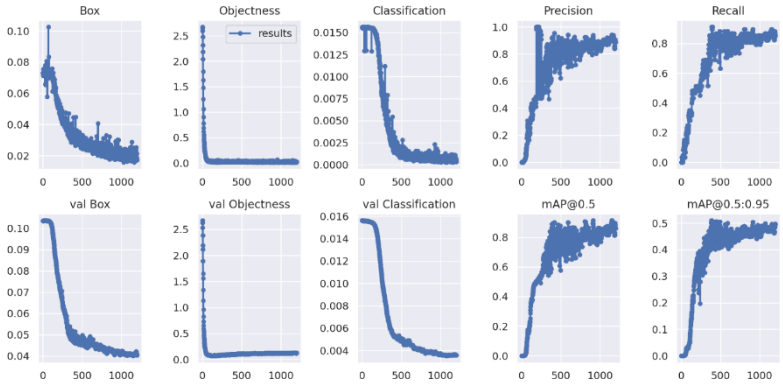


Figure 7: Training results for Non-Mitotic Cells

The results of the testing (using the model that essentially maximizes mAP0.5:0.95) are displayed in the appendix section.

Regarding the decision of not applying standardization or normalization, below we present a detection example where we see that the lack of colour information led to bad detection cases (see Figs. 8 and 9).

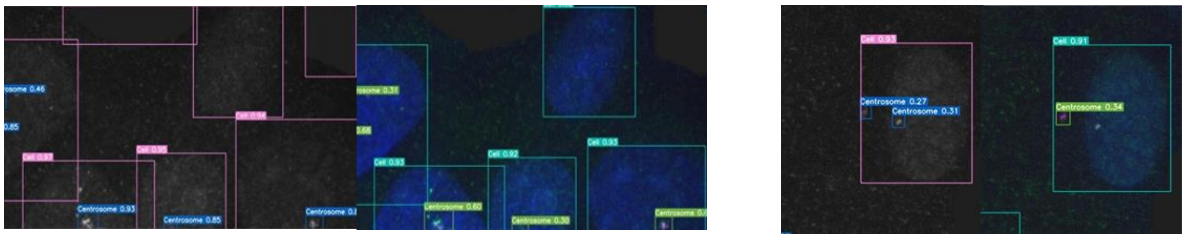


Figure 8 and 9: Detection example of Nuclei and Centrosomes – No colour Vs. Colour (for Non-Mitotic Cells)

Mitotic Cells

Since we noticed that there weren't many overlaps between centrioles of mitotic cells, we decided to detect centrioles individually, instead of centrosomes. The model was also trained to detect the nuclei of mitotic cells only. Below, in Fig. 10, we present the training results obtained.

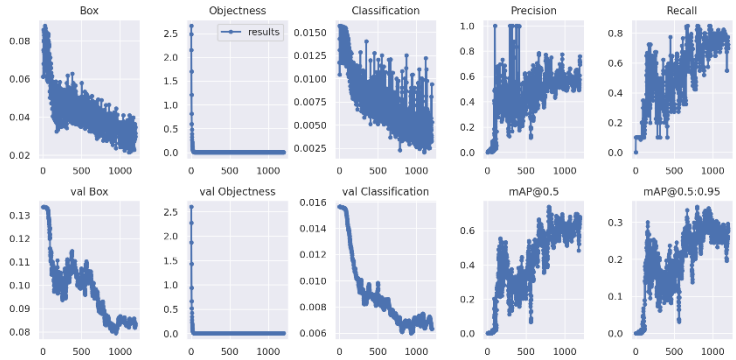


Figure 10: Training Results for Mitotic Cells

As depicted in the image, the loss components achieved minimal values both on the training and validation sets around the 1000th epoch, where the Precision was 71%, the Recall 74% and mAP 0.5 64%. Although the accuracy metrics of Precision, Recall and mAP values increased with training, they did not converge to peak values, contrarily to what happened with the interphase cells. Furthermore, we can see that the plots are quite "spikey", and there are many factors that might have influenced this. One possible way to tackle this problem would be training the model for a few more epochs and tweak the hyperparameters. A smaller learning rate and a smaller momentum factor may lead to less "spikey" values of the already mentioned metrics. Of course, these values can't be too small, otherwise we might end up in a local minima loss value (see [11]). At the same time, having more images for training would allow the model to learn more different examples and allow him to generalize better for new samples.

It is important to mention that although we could have trained the model with the default hyperparameters for a few more epochs, we were not able to do it because of Google Collab limitation for non-premium users. The model was picked using the same previous criteria – the testing results are present in the appendix.

Second Experiment

Training Results

Interphase Cells

Following the first approach, we requested more images from IGC and we were given another batch of 10 images for the interphase cells. Through pre-processing, augmentation, and by joining the images previously used for training, we ended up with 56 images. We used the same images for validation and testing than we had used in the first experiment.

Running this experiment for 600 epochs proved to be fruitful since the metrics registered a slight increase. Precision stabilized around 87%, Recall stayed at 83% and the $\mathrm{Map}_{0.5}$ was of 83.5%. Google Colab's GPU limitations prevented us to go farther than 600 epochs, which normally would be a point where the loss component was still improving significatively. However, this experiment was a promising achievement: considering the increase in the number of images affected positively the performance of the model, as expected, we are led to think that had more images been available and correctly annotated, the results could have improved marginally. No other test

that we performed – or that is available in the computer vision literature for centrioles and centrosomes – had a better result, so far.

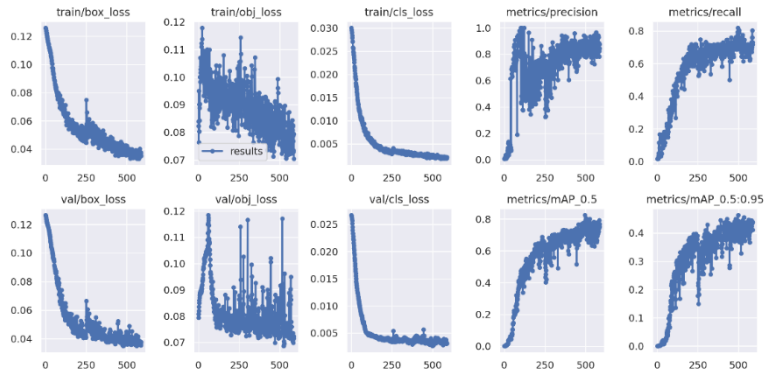


Figure 11: Second Experiment Training results for Interphase Cells

We confirmed that the only downside in the increase of the number of images is that the training takes longer to run. However, since the best weights are automatically registered by the model, that process would happen only once.

Conclusion and Future Research

Throughout this extensive process, we faced numerous drawbacks, namely the lack of a significant number of images for training (1), the quality of the images (2), the fact that in interphase cells centrioles tend to overlap (3) and GPU constraints (4) already mentioned. However, the knowledge that we now have on these matters, allows us to propose alternative paths for future research. Therefore, we suggest that the process for obtaining the images is changed: although Max Projection is not a bad approach for a naked-eye view (as used by IGC), it does not provide the model enough information. Nevertheless, the use of a 3D model or another transformation of such images would tackle the problems (2) and (3). Upgrading the GPU would provide an easy solution to issue (4), whereas the use of more images for training (1) – a hundred or so - would be crucial to obtaining better results.

Regarding the experience, this project was a great opportunity to contact and learn from a professional work environment in a specialized field. We were able to delve into state-of-the-art deep learning models such as YOLOv7 and to apply a similar reasoning and thought processes that we will need for our upcoming Master's Thesis.

In what concerns the results of the project, we provided a pioneering trial on mitotic cells and went off the beaten track by avoiding to count centrioles for our interphase cells' experiments, ending up with satisfactory results for the first ones – a consequence of the material we had – and great results for the latter, that wee believe to be numerically the best yet in the slim centrosome literature available.

Finally, by providing a final product for Instituto Gulbenkian Ciência in the form of an annotated easy-to-use Python notebook that generates an Excel File like the ones currently computed manually, thus contributing to research that is having relevant results in the early detection of some types of cancer, we believe that our endeavour constitutes a relevant contribution to these topics.

References

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. The MIT Press, pp.452-453.

[2] Mao, L. (2021, April 12). *Mean average precision map for object detection*. Lei Mao's Log Book. Retrieved January 1, 2023, from https://leimao.github.io/blog/Object-Detection-Mean-Average-Precision-mAP/

[3] Boesch, G. (2023, January 1). *Yolov7: The most powerful object detection algorithm (2023 guide)*. viso.ai. Retrieved November 12, 2022, from https://viso.ai/deep-learning/yolov7-guide/

[4] Hughes, C. (2022, November 25). *Yolov7: A deep dive into the current state-of-the-art for object detection*. Medium. Retrieved October 22, 2023, from https://towardsdatascience.com/yolov7-a-deep-dive-into-the-current-state-of-the-art-for-object-detection-ce3ffedeeaeb

[5] WongKinYiu. (n.d.). *Wongkinyiu/Yolov7: Implementation of paper - yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors*. GitHub. Retrieved October 25, 2022, from https://github.com/augmentedstartups/yolov7

[6] *Give your software the power to see objects in images and video*. Roboflow. (n.d.). Retrieved November 12, 2022, from https://roboflow.com/

[7] Solawetz, J. (2020, August 20). *Why and how to implement random rotate data augmentation*. Roboflow Blog. Retrieved December 11, 2022, from https://blog.roboflow.com/why-and-how-to-implement-random-rotate-data-augmentation/

[8] Nelson, J. (2020, October 5). *Why and how to implement random crop data augmentation*. Roboflow Blog. Retrieved Retrieved October 23, 2022, from https://blog.roboflow.com/why-and-how-to-implement-random-crop-data-augmentation/

[9] Nelson, J. (2020, October 5). *Why to add noise to images for machine learning*. Roboflow Blog. Retrieved October 23, 2022, from https://blog.roboflow.com/why-to-add-noise-to-images-for-machine-learning/

[10] Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. The MIT Press, p.676.

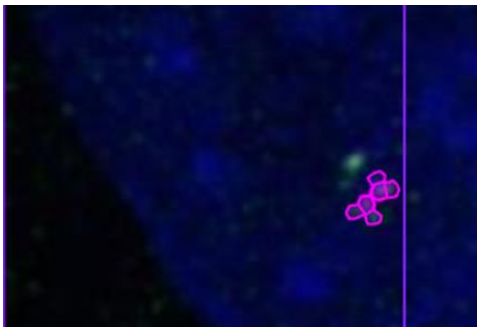
[11] Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. The MIT Press, p.296.

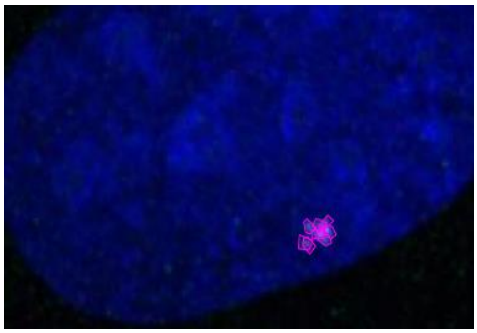
Appendix

Here we provide simple and concise information complementing the one of the report.

Individual centrioles detection on interphase cells

The images of figs. X and X show, respectively, a labelled training with four and nine centrioles.





Figures 12 and 13: Details of Training for individual centrioles on interphase cells

Loss values dropped to minimal values, but evaluation metrics are showing spikey results without converging. This is, most likely, a consequence of having a very difficult detection problems with less than 20 images for training.

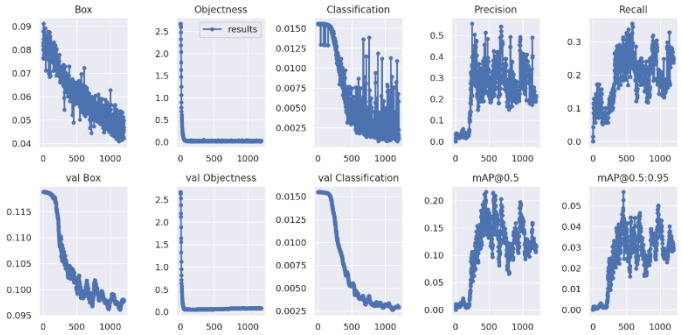


Figure 14: Results of Training for individual centrioles on interphase cells

Mitotic Cells in the Second Experiment – no improvement

Similarly to the process used for the interphase cells, a new batch of 15 images was requested for training. Through pre-processing and augmentation, we doubled this value. However, from the beginning it was noticeable the low image quality of this batch.

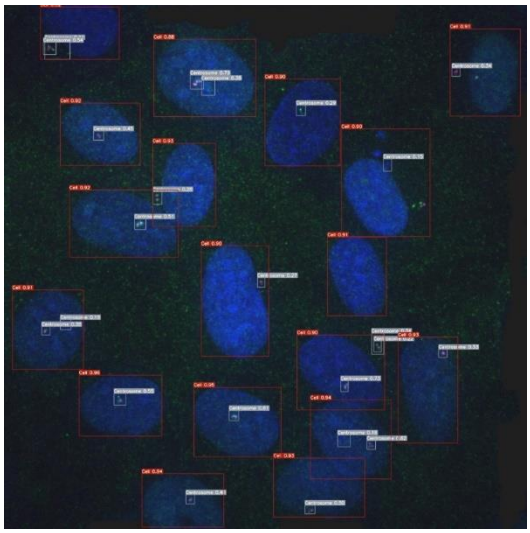
This time, on a first test, we used the best weights from the first approach, allocating only the new images for training and the same images for validation and testing that we had used on the previous trials. The images previously used for testing were left untouched. We concluded that although the weights contained some useful information, the lack of information caused by the absence of the first training images was fatal to the process. Besides, the model was overfitting due to the lack of more training images and the disproportionately high number of epochs.

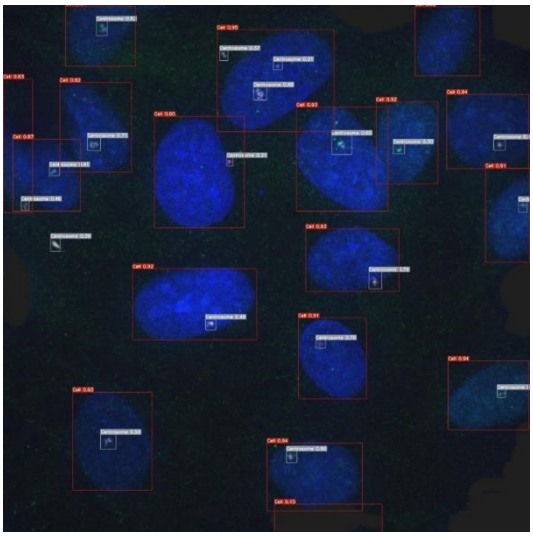
Therefore, we disregarded the previous trial and ran a final experiment with all images from all batches for training, with the same four images used for validation and testing.

By running this experiment, also with 1200 epochs, we were hoping to find better results. However, the approach was slightly worse than the best result already obtained, which should be caused by the lower quality of the new images. These results are not presented since they do not constitute any improvement.

Testing Results

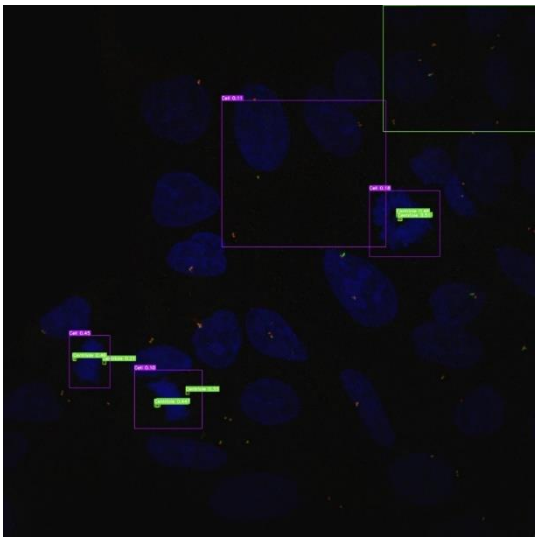
Interphase Cells

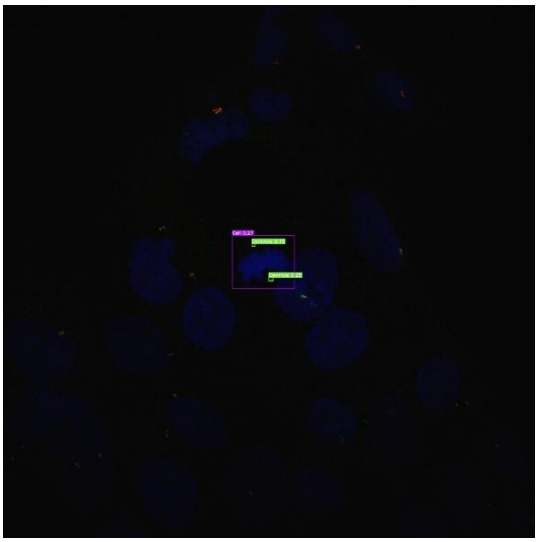




Figures 14 and 15: Testing for Interphase Cells

Mitotic Cells





Figures 16 and 17: Testing for Mitotic Cells