Cell type identification using Deep Learning Model trained with cell type information transferred from mIF to co-registered H&E images

Thesis by
Martin Beaussart

Carried out at the CHUV
Under the supervision of Dr. Andrew Janowczyk

UNDER THE DIRECTION OF PROF. PASCAL FROSSARD
FROM THE SIGNAL PROCESSING LABORATORY

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
Lausanne, Switzerland

2022
Defended August 25, 2022

Abstract

Research using whole slide images (WSI) of histopathology slides has grown exponentially in recent years. Different types of tissues are used. Most are stained with hematoxylin and eosin (H&E). Still, some use more expensive but more accurate staining techniques, such as immunofluorescence (IF), which give more information (e.g., cell types) compared to H&E. The analysis of these digital images then allows the application of various programs, such as those based on deep learning. These programs can then be used to develop image-based biomarkers to predict the patients' diagnosis, prognosis, or therapeutic response, which would facilitate precision medicine applications.

Unfortunately, deep learning models require numerous data. Many digitized histological slides are available to researchers, especially with H&E. However, the H&E does not contain cell type information, which must therefore be labeled. Sadly, manual labeling is laborious and time-consuming work that can only be done with the help of pathologists.

The approach we propose is based on mIF WSI because mIF WSI contains information about the cell type. The project aims to automatically label H&E cells by transferring the information from the mIF WSI. The process involves double staining the tissues, first in H&E and then in multiplexed immunofluorescence (mIF) after decoloring. After the double scan, a process matches each cell on H&E with the corresponding cell on mIF. The match between H&E and mIF allows us to transfer the cell type to the H&E cells and thus to create a massive dataset. Current datasets are composed of about 20,000 cells [3][6][11], while our method allowed us to extract nearly one million cells from three tissues. The dataset created allowed us to train different models, giving us an accuracy of around 93%.

TABLE OF CONTENTS

Introduction

1.1 Background

Digital pathology is a sub-field of pathology that processes digitized glass slides using a whole slide image scanner and then analyzing the digital images. Technological advances and increased interest in precision medicine have recently paved the way for the development of digital pathology. This growing interest has allowed many clinical researchers to build large databases of digital whole slide images (WSI).

Hematoxylin & Eosin Staining

In digital pathology, researchers widely use hematoxylin & eosin (H&E) images due to their clinical availability in both retrospective and prospective cohorts. Hematoxylin stains cell nuclei a purplish blue, eosin stains the extracellular matrix and cytoplasm pink, and other structures take on different shades, hues, and combinations of these colors. Thus, a pathologist can easily differentiate between a cell's nuclear and cytoplasmic parts. In addition, the general staining patterns of the stain show the general arrangement and distribution of the cells and provide an overview of the structure of a tissue sample.

While H&E images are a quick and cost-effective way of visualizing histological tissue, it remains challenging to infer cell types consistently (e.g., lymphocytes, macrophages, neutrophils) from H&E. Other staining techniques, such as immunohistochemistry (IHC) or immunofluorescence (IF), are more adequate for this task.

Immunohistochemistry and Immunofluorescence Staining

IHC is the most common application of immunostaining. It uses antibodies to check for specific antigens (markers) in a tissue sample. The antibodies are usually linked to an enzyme. After the antibodies bind to the antigen in the tissue sample, the enzyme or dye

is activated. As a result, the antigen can be seen under a microscope. On the other hand, IF images reveal cell type as it relies on fluorescence-tagged antibodies, which bind with cell-type specific antigens (e.g., CD20).

While IHC and IF use antibodies to target individual proteins, IHC employs reporter enzymes that produce a chromogenic signal to indicate the presence of target proteins, while IF uses fluorophore. Multiple fluorophores can be used to detect individual proteins, which enables the multiplex capabilities of multiplexed Immunofluorescence (mIF).

The benefits of IF and IHC are that each cell type can be identified when a specific biomarker is available. The main difference between IHC and mIF resides in the brightfield view of IHC and the fluorescence view of mIF. This difference has repercussions on the number of possible targets. Indeed, the brightfield view limits the number of potential targets to two or three when mIF can exhibit tight excitation and emission spectra that enable multiple fluorophores to be used simultaneously. For mIF, 5-8 targets can typically be detected simultaneously. For this reason, we have decided to use mIF in this work as it has good signal intensity and the number of markers we can operate simultaneously. Unfortunately, one downside is that mIF remains expensive in both time and cost, requiring a fluorescence-enabled slide scanning microscope.

Deep Learning

The growing number of these WSIs allows the application of various techniques, such as image processing or machine learning, to these tissues. These techniques are currently applied to help in diagnostic medicine, achieving an efficient and less costly diagnosis, prognosis, and disease prediction, thanks to the success of machine learning, and in particular deep learning.

Deep learning is a sub-field of machine learning methods based on artificial neural networks. These neural networks attempt to simulate the behavior of the human brain by allowing it to learn from large amounts of data. These neural networks are made up of a simple mathematical function that can be stacked on top of each other and arranged in layers, which gives them a sense of depth, hence the term Deep Learning. As information passes through a layer, each node in that layer performs simple operations on the data and selectively passes the results to other nodes.

A perceptron is a single-layer neural network. It consists of four main parts: input values, weights and bias, weight sum, and an activation function.

The operation starts by multiplying all the input values by their weights. Then all these

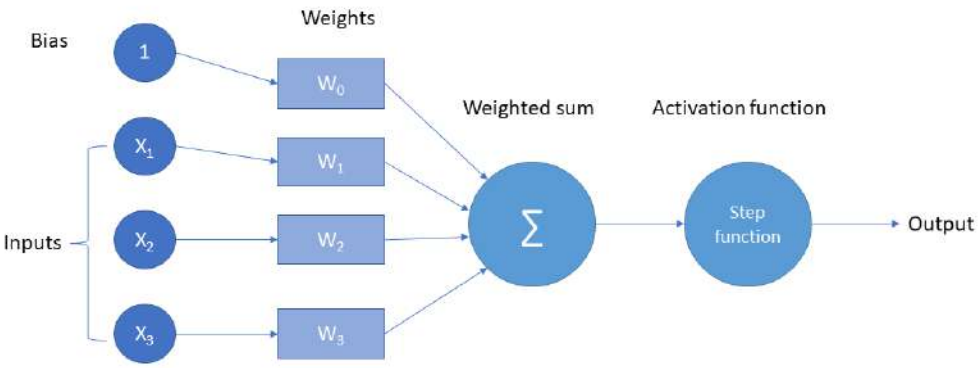

Figure 1.1: The perceptron has four essential elements: input, bias, weight, and activation function.

multiplied values are added together to create the weighted sum. The weighted sum is then applied to the activation function, producing the perceptron's output. The activation function matches the output to the required values, such as (0,1).

A basic neural network contains three layers: the input layer, the hidden layer, and the output layer, where each layer is composed of artificial neurons.

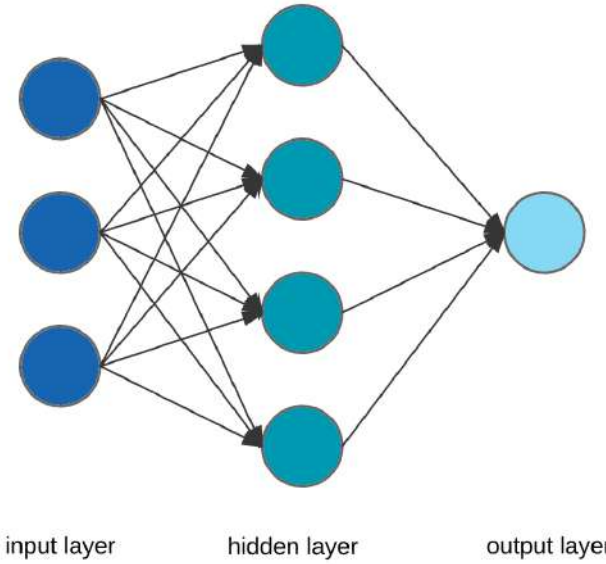

Figure 1.2: Deep neural network architecture with three layers.

The perceptron calculations occur for all neurons in a neural network, including the output layer, and are known as forward propagation. After a forward pass, the output layer compares its results to the ground truth and adjusts the weights according to the

differences between the ground truth and the predicted values. This process is called backpropagation.

Different types of neural network architecture exist. A famous one is called Convolutional Neural Network (CNN). It is a standard for image recognition. CNNs are designed to automatically and adaptively learn spatial hierarchies of features using a cascade of filters automatically tuned to extract meaningful information from the input image data.

Convolutional neural networks have three main types of layers, which are:

- Convolution layer: runs the input images through a set of convolution filters, each of which activates certain features of the images.
- Pooling layer: simplifies the output by performing non-linear downsampling, thus reducing the number of parameters the network has to learn.
- Fully connected layer: connecting each neuron in one layer to each neuron in another layer.

The convolutional layer is the first layer in a convolutional network. Although additional convolutional layers or pooling layers can follow convolutional layers, the fully connected layer is the last layer. The CNN network gains complexity with each layer and identifies more significant portions of the image. The first layers focus on simple features, such as colors and edges. As the image data progresses through the layers of the CNN, the latter begins to recognize more prominent features or shapes of the object until it finally identifies the desired object.

The convolutional layer is considered the central building block of a CNN, where most of the computations are performed. It requires a few components: input data, a filter, and a feature map. The process is simple. It takes a filter or kernel (in the example fig. 1.3, it is a 3×3 matrix) and applies it to the input image to get the feature map. More precisely, it computes a dot product between the weights of the kernels and each small region they are connected to in the input volume.

The pooling layer is responsible for reducing the spatial size of the convolved feature. This reduces the computational power required to process the data. Unlike the convolutional layer, the kernel applies an aggregation function to the values in the receptive field, feeding the output array. There are two main types of pooling:

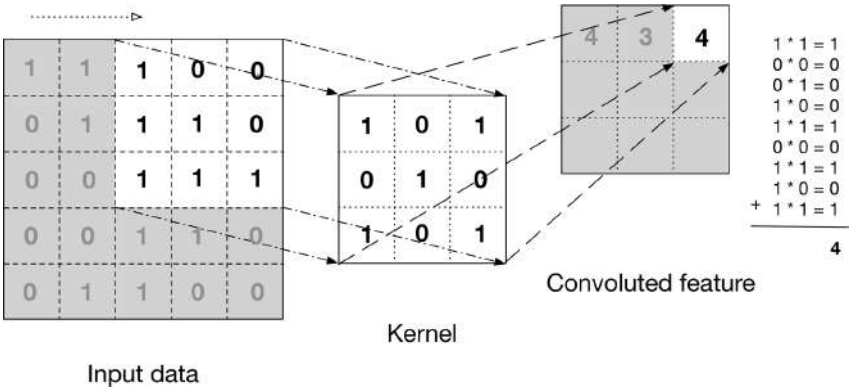

Figure 1.3: Example of a convolution with a filter/kernel and its application to the input image in order to get the convolved feature.

- Max pooling: As the filter moves across the input, it selects the pixel with the maximum value to send to the output array.
- Average pooling: As the filter moves across the input, it calculates the average value within the receptive field to send to the output array.

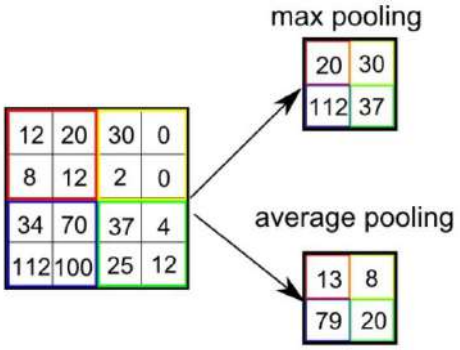

Figure 1.4: Example of the two main types of pooling.

There are many types of artificial neural networks used to approximate generally unknown functions of varying complexity. A simple approach to dealing with complex tasks is increasing the network's depth. As the hypothesis space of a learning algorithm increases, the algorithm can learn increasingly rich structures. For example, in a deep convolutional neural network classifying images, the first layer will be trained to recognize fundamental features such as edges, the next layer will train to recognize collections of edges such as shapes, and the next layer will even learn higher-order features such as smiles. A problem with deeper models is that they can overfit the training data, so they memorize and do not generalize. More importantly, deeper models face the vanishing gradient problem [2].

The vanishing gradient problem is when the gradients of the loss function approach zero because it has passed through many layers.

To overcome this problem, several methods were proposed, such as the use of batch normalization or the use of residual networks.

Current Methods in Digital Pathology

Deep learning models are mainly used on H&E WSI because H&E is the primary staining used in histology and has been performed since the 19th century. Indeed, as mentioned earlier, H&E is a simple and inexpensive staining technique. In addition, this staining allows the visualization of structural information of the tissue, such as nuclear morphology and extracellular matrix. The main limitation of H&E is the lack of information on cell type, so researchers combine the use of H&E with more detailed staining techniques such as IHC or mIF. Often, two sequential cuts of the same tissue block are taken, one with H&E staining and the other with IHC or mIF staining. The researcher can then translate the H&E information by looking at the IHC or mIF as a reference [7]. Another method is to use the same tissue for both H&E and IHC staining, starting with the H&E stain, then decolorizing the stain, and finally staining with IHC [8].

1.2 Goals and Challenges

While deep learning classifiers have been shown to be able to identify various cell types on H&E images [12] training these classifiers often require large amounts of labeled data. Manually annotating this training data on whole slide images (WSIs) is laborious, time-consuming, and nevertheless prone to errors. On the other hand, mIF images afford the ability to localize identifying proteins of cell types at scale, obviating the need for manual annotation. If these labels can be successfully transferred to H&E images, a deep learning classifier can be trained to label subsequent H&E WSI, which lack corresponding mIF WSIs. Once validated, this DL classifier could be employed to develop image-based biomarkers to predict the diagnosis, prognosis, or therapy response of patients, further helping facilitate precision medicine applications.

Therefore our goals and challenges can be summarized as follows:

- Correctly map an mIF WSI to its H&E WSI at the nucleus level.
- Generate a large H&E dataset with the corresponding label acquired from the mIF.

- Creation of three models to accurately identify lymphocytes, myeloid cells, and melanoma tumor cells
- Creation of sub-models to accurately identify the type of lymphocyte or myeloid cells.

1.3 Proposed Approach

Our approach is to use the information given on an mIF WSI and map it to the H&E WSI. The slide is first stained with H&E and then scanned. Afterward, a destaining protocol removes the H&E stain from the tissue before restaining with mIF and scanning. From the mIF WSI, targeted protein expression levels of all cells and their corresponding positions are obtained. This part is done with the inForm® analysis software [5], which gives us the centroid coordinate (x, y) of the nucleus for each identified protein. Due to tissue movement, as a result of restaining and coverslipping, cells often no longer remain aligned between mIF and H&E images. As a result, a WSI registration process must take place to identify the correct mapping from the mIF space to the H&E space.

Registration starts with a translation transform, followed by an affine transform on the region of interest. The next step of the registration is done on pairs of small tiles, where we apply non-rigid transforms. The registration process results in a transform which can then be applied to mIF cells labels to match them with the H&E cells. The H&E cells are localized by segmentation on the H&E image, and this process is done with the HoVer-Net architecture [3], a deep learning model using pre-trained weights.

Locating cells on H&E and comparing them against their mIF counterparts has the advantage of performing quality control and filtering to ensure high-quality information for DL training (e.g., detecting and removing missing cells or false positive areas). Using the updated positions of nuclei and their expressed proteins, we built a database containing 64 by 64 pixels patches (16 by 16 μm at 40X magnification) centered on the nuclei with their associated expressed proteins (labels). The identified proteins are then compared to a catalog containing different cell types with their corresponding proteins, from which we can identify the cell type.

The project aims to identify six different cell types. Three types of lymphocytes: T cells, B cells, and natural killer cells. Two types of myeloid cells: macrophages and dendritic cells. And finally, the melanoma tumors cells.

The large dataset created with the registration includes the six cell types of interest, plus the cells identified as "others", which are various cells present in the tissues that we do

Table 1.1: Cell type with their associated cell markers and fluorochrome

`Cell type`	Dendritic cell	Natural killer	T cells	Macrophage	B cells	Melanoma tumor
`Cell` marker	CD11C	CD56	CD3	CD68	CD20	SOX10
`Fluorochrome`	OPAL 620	OPAL 570	OPAL 520	OPAL 480	OPAL 780	OPAL 690

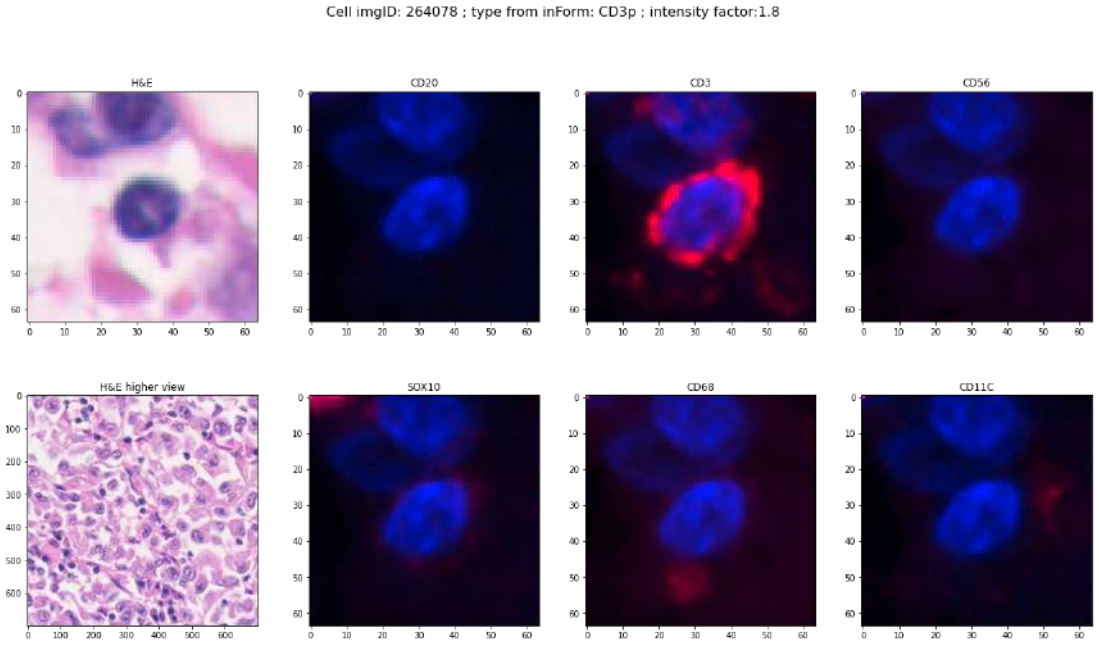

Figure 1.5: Example of a lymphocyte, the six images on the right are mIF images. The blue color represents the DAPI channel, and the red color represents the specified protein channel. More examples in the appendix [fig. .2 and fig. .1]

.

not want to identify but are necessary for training. This dataset is then used to train the DenseNet models. Three binary models are trained to determine whether the cell is a lymphocyte, a melanoma tumor cell, or a myeloid cell. In addition to these three models, four sub-models are trained to determine the type of lymphocyte or myeloid cell.

1.4 Contributions

Dr. Andrew Janowczyk, Alexandre Wicky and Prof. Olivier Michielin initiated the project. Under there supervision, Antoine Ribault-Gaillard and Justin Mapanao helped to create and improve the project. The project I inherited was already promising. The dataset created was composed of 36,000 patches, and the registration was on track with a valid

registration at the structure level. However, at the nucleus level, the registration was not sufficient. The dataset shows the presence of patches without nucleus in their center. As a result, the model trained using this dataset only reached 77.6% accuracy in their simplest experiment.

The contributions to this project were first to improve the registration. Multiple improvements were made to finally achieve accurate and robust registration at the nucleus level. These improvements include cell localization on the H&E with HoVer-Net and new image transformations. The improved registration allowed us to obtain a better, bigger, and higher quality dataset.

The second contribution concerns the scale of the registration. The inherited project worked efficiently on a section of a WSI but was not adequate for a larger scale. Therefore, modifications were made to optimize the process, resulting in a dataset of 965,169 patches.

Finally, the last contribution concerns the classification models. Different models were created and gave a better accuracy, even for complex tasks. These models were able to predict with an accuracy around 93% the label of nuclei as lymphocyte, myeloid cell or melanoma tumor cell. More specifically, our model identifying the cells as lymphocytes achieved 93% accuracy. The model identifying the cells as melanoma tumor cells achieved even a higher accuracy of 96%. Our worst model, the model identifying the cells as myeloid cells also obtained an accuracy of 93%, but this accuracy must be nuanced because myeloid cells represent only 5% of our dataset.

To obtain these models, one innovation, in particular, was the addition of the nucleus mask as input. Indeed, adding this information helps our model by helping it to identify the cell we want to label correctly. A second innovation that was lightly tested but would require more analysis was the addition of a low magnification image of the cell. The idea was to provide more information to the model by giving it the context of the cell, such as the general staining of the tissue or the density of the area around the cell. This information could potentially help the model on external tissues. However, this innovation was first tested to potentially improve the accuracy of our myeloid cell model on the test dataset, but the result was not convincing.

At that time, no external tissue was available to confirm an improvement. Due to time constraints, this feature has been removed because no improvement was found on the test dataset, and the computational time was doubled, but it could be tested again later to possibly improve our models.

Summary of contributions:

- Implementation of a registration accurate at the nucleus level.
- Registration works efficiently on the WSI, not just on one region.
- Creation of a large dataset of relatively good quality.
- Creation of cell type identification models with significantly improved accuracy compared to the inherited project.

Methods

2.1 H&E and mIF Stainings

The tissues utilized to create our dataset were obtained at the CHUV in Lausanne. Three sequential sections of a formalin fixed paraffin embedded (FFPE) block from melanoma have been used. Due to the loss of tissue during the restoration and discoloration process, the use of special slides was necessary, for which we decided to use Superfrost Plus slides.

Indeed, tests on standard archival H&E slides show tissue loss during mIF staining. After the first staining, the tissue is scanned to generate the digital H&E image before applying the discoloration protocol.

The discoloration protocol used is the following:

- Put the H&E slide in xylene for 3-5 days to remove the coverslip.
- Remove the coverslip and discard.
- Rinse with xylene to remove the remaining glue.
- Rinse the slide in 100 EtOH 100 (3 x 1 minute), 95 (2 x 1 minute) and 70 (1 x 1min) baths, sequentially.
- If the slide still contains impurities, continue to shake in 70 EtOH bath.
- Rinse the slide in 0.18 acid alcohol (0.18 ml HCl in 100 ml ethanol) for 10 minutes to remove most of eosin and hematoxylin.
- Rinse the slide in distilled water before use

After destaining, multiplexed immunofluorescence staining is performed with a Ventana BenchMark Special Stains system. The staining is performed using fluorescently labeled

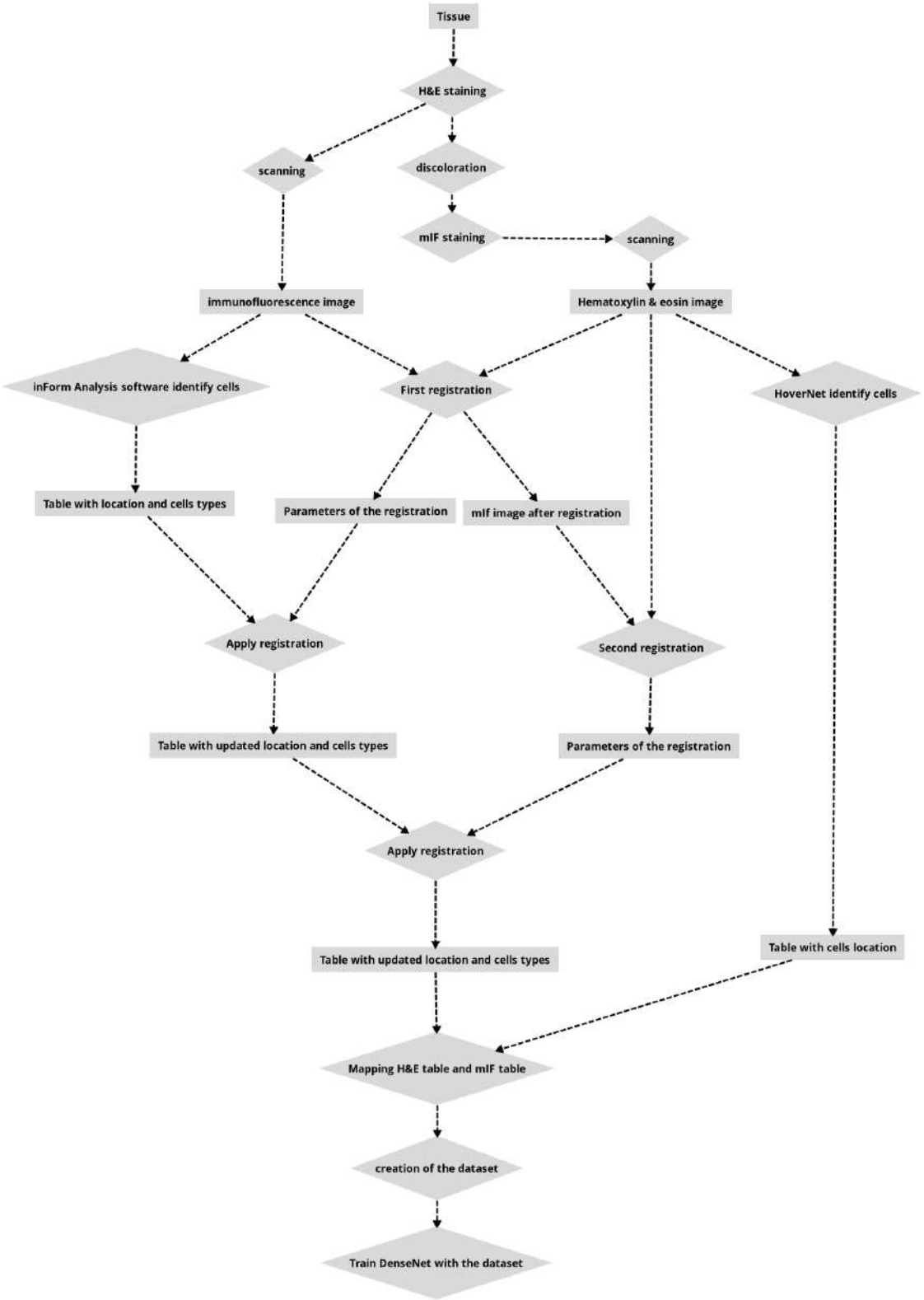

Figure 2.1: Flowchart of the overall process.

antibodies that bind to cell type-specific antigens. For scanning, H&E and multiplexed immunofluorescence images were digitized using a VectraR PolarisTM imaging system into a qptiff file.

2.2 Data Processing

InForm®

Cell type identification in the multiplexed immunofluorescence image is made using inForm® Analysis software. The configuration of the software is the following:

- Segment Tissue: Trainable Tissue Segmentation.
- Find Features: Adaptative Cell Segmentation.
- Phenotyping: Phenotype the cells.

The inForm® output provides a complete representation of each cell: phenotype, coordinates, and other various information.

To work, the project requires three pieces of information from the inForm® output to work, the coordinate, the phenotype, and the confidence value. The confidence value reported by inForm® is the probability of the winning phenotype, $P(i|X)$, expressed as a percentage.

The confidence value is critical because it helps us improve the dataset quality by preventing us from taking mislabeled cells. Indeed, labeling cells from an mIF image is prone to errors.

The problem of mIF image processing is that obtaining the cell data is a rather complex task. Therefore, different softwares give different results. In addition, a recurring problem when labeling cells from an mIF WSI is that of cells expressing multiple phenotypes. This ambiguity is partially solved by inForm® with the confidence value but is still subject to errors. For example, it happens that tumor cells expressing SOX10 also express CD56. In this case, the inForm® analysis could mistakenly identify the cell with CD56 and not SOX10. Therefore, we only select cells for which inForm® is sufficiently confident about their type.

In order to check the inter-software variability, another software was used on a WSI to compare it to the inForm® analysis software. IFQuant, new software developed in-house by the Centre Hospitalier Universitaire Vaudois (CHUV) for multiplex IF analysis was chosen for this purpose. Our comparison [fig. 2.3] was made on the test dataset of a WSI

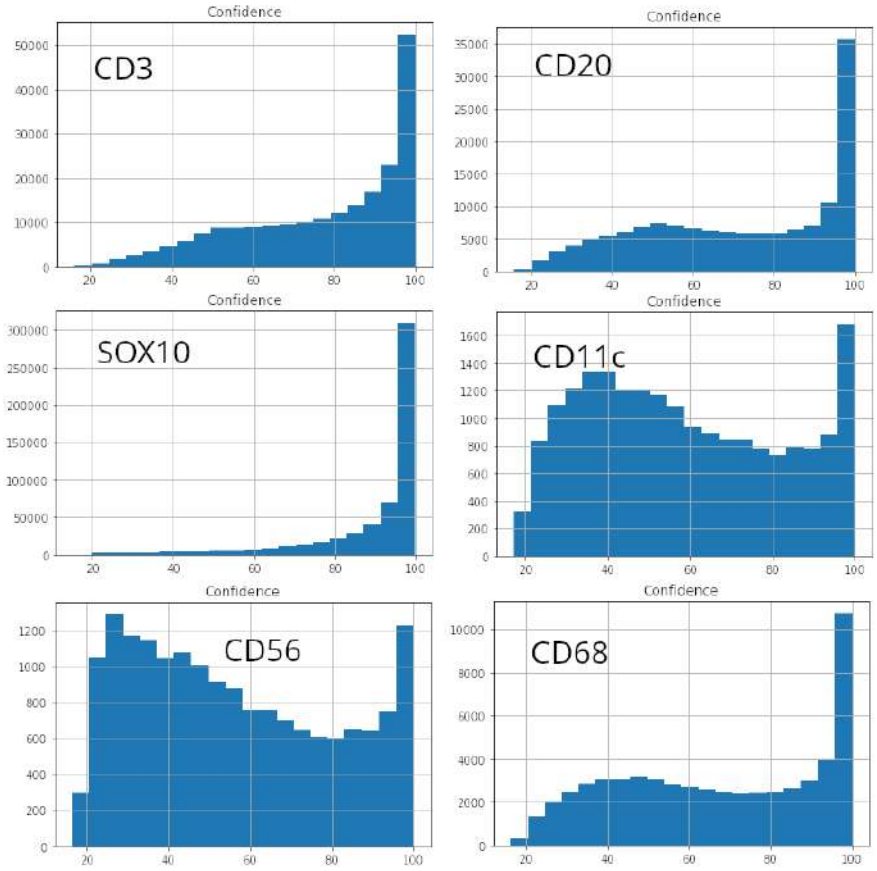

Figure 2.2: Distribution of the confidence value for each phenotype on one WSI.

with the output of models trained with a dataset composed of cells labeled with inForm®. The use of models helps to estimate the true nature of a nucleus when there is a divergence in labeling between inForm® and IFQuant.

IFQuant, unlike inForm®, does not exhibit a single cell marker but all markers expressed by the cell. Indeed, it does not use a confidence value to filter the cell markers. For simplicity's sake, we group these cell markers to create four categories. The lymphocyte category is validated when a cell expresses at least one of these markers (CD3, CD20, or CD56) but none of the others (CD68, CD11C, or SOX10). The melanoma category is validated when a cell expresses at least SOX10 but none of these markers (CD68, CD11C, CD3, CD20). The myeloid cell category is validated when a cell expresses at least one of these markers (CD11C or CD68) but none of the others (CD3, CD20, CD56, or SOX10). If none of these categories are validated by a cell, there is uncertainty, therefore it will be classified as 'uncertain'.

Multiple pieces of information can be highlighted from the distributions obtained in

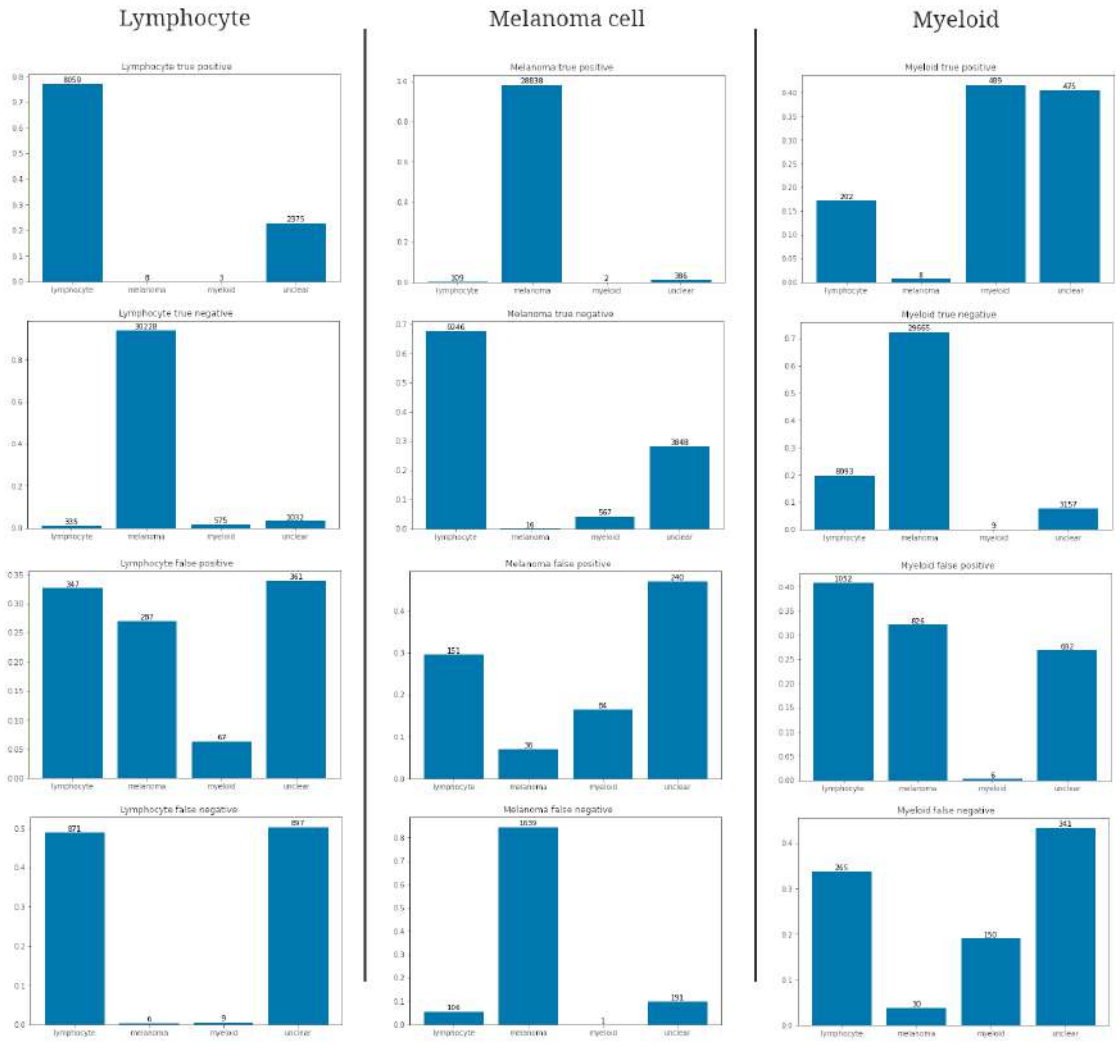

Figure 2.3: Distribution of each cell type identified by IFQuant on the results of models trained with a test dataset created using inForm®.

fig. 2.3. For lymphocytes, true positives, true negatives, and false negatives show a low variability between inForm® and IFQuant. The only problem could be the false positives because at least 33% of these cells are actual lymphocytes. The good thing is that our model can recognize them as lymphocytes because they are in the false positive category. Melanoma tumor cells show a good overall result. Few false positives are in fact true positives for IFQuant, 7%, or only 36 cells, and 6% are mislabeled for IFQuant in the false negatives. The myeloid cell category is where we have the biggest divergence between inForm® and IFQuant. Without the "unclear" label, inForm® finds 901 myeloid cells in the test data set, but IFQuant finds that 505 of these cells are lymphocytes or melanoma tumor cells.

Input for Registration

Registration is computationally expensive, and WSIs are large images (4-8 Giga per H&E WSI and 16-18 Giga per mIF WSI). We, therefore, want to apply the registration not to the entire image but only to regions of interest. Our regions of interest are designed to be 6,000 by 6,000 pixels images (1,500 by 1,500 µm). To select our regions of interest, we grid the entire H&E WSI image and filter it based on the proportion of tissue in each region.

The general approach of our method is to apply the registration on each selected ROI and then aggregate the results. The registration process uses two images as input, an H&E ROI and a region of the mIF where the corresponding mIF ROI is supposed to be located. Indeed, the coordinates of a point in the H&E are not the same in the mIF, so the mIF region is selected based on an initial registration with a low magnification translational transformation.

The initial registration gives us the parameters that best match the two WSI using a translation transform. The parameters are horizontal and vertical translation ($param_x$ and $param_y$). Using the translation parameters and the H&E coordinate, we can roughly guess that the point (x, y) in the H&E should be close to the point ($x + param_x$, $y + param_y$) in the mIF.

The parameters found after the registration suggest where to match the H&E coordinates on the mIF. However, the low magnification and the registration over the whole image imply approximation errors. The selected mIF region is made larger to overcome these errors. This ensures that the desired ROI lies within this region. Finally, the last input for our registration is the position of the cells and their labels obtained from inForm®.

2.3 Registration

Rigid and Affine Registration on ROIs

The registration works by matching the images with the information of the nuclei. For the H&E WSI, we use the hematoxylin channel, and for mIF WSI, the DAPI channel.

DAPI is a fluorescent dye that binds strongly to adenine-thymine rich regions of DNA, so it is widely used for nuclear staining. Based on the pre-registration parameters found during low-magnification registration ($param_x$ and $param_y$), each selected H&E ROI obtains a reasonable approximation of the coordinate of the corresponding supposed mIF ROI.

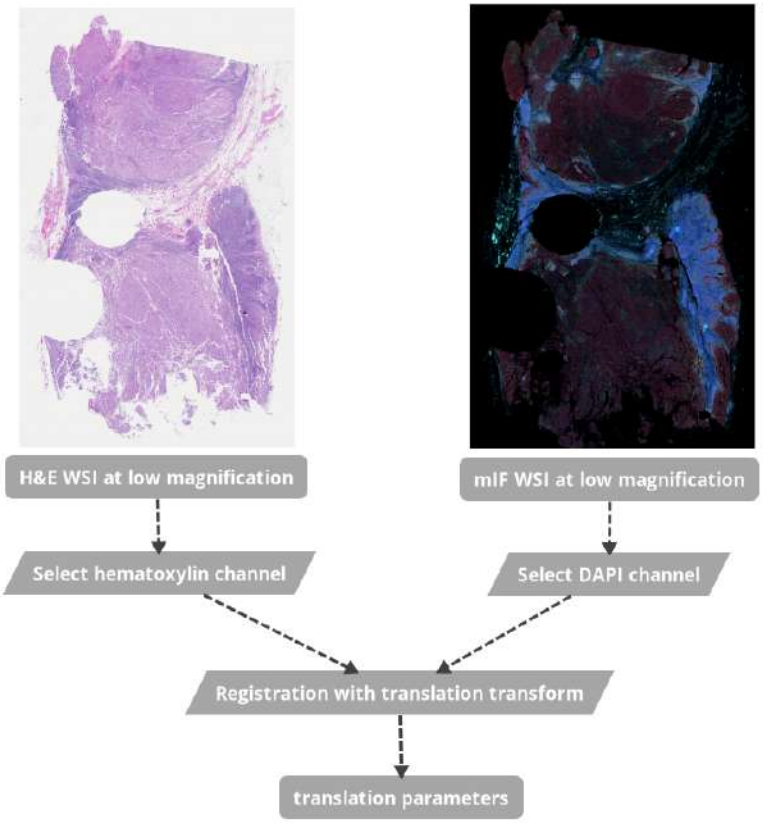

Figure 2.4: The pre-registration on the WSI is fundamental since it allows us to match the ROI in the H&E with the corresponding ROI in the mIF.

However, regions are not perfectly aligned, so we expand the mIF region to ensure we do not miss any part of the image after registration [fig. 2.6].

The registration process involves two images. One image, the moving image $I_M(x)$, is deformed to fit the other image, the fixed image $I_F(x)$. Registration is the problem of finding a transformation $T(x) = x + u(x)$ that makes $I_M(T(x))$ spatially aligned to $I_F(x)$. The transformation is defined as a mapping from the fixed image to the moving image. To evaluate the performance of the registration, a similarity metric (C) measures how well the fixed image matches the moving image. In this work, we use the Mattes mutual information (MI) metric [13].

The optimisation problem of the registration can thus be written as: $T^{opt} = \arg\min_{T} C(T; I_F; I_M)$

SimpleElastix [9] is a medical image registration library that we use in the project. Its mechanism can be found in fig. 2.5. The process is divided into different components.

The pyramid component is a multi-resolution strategy to improve the capture range and the

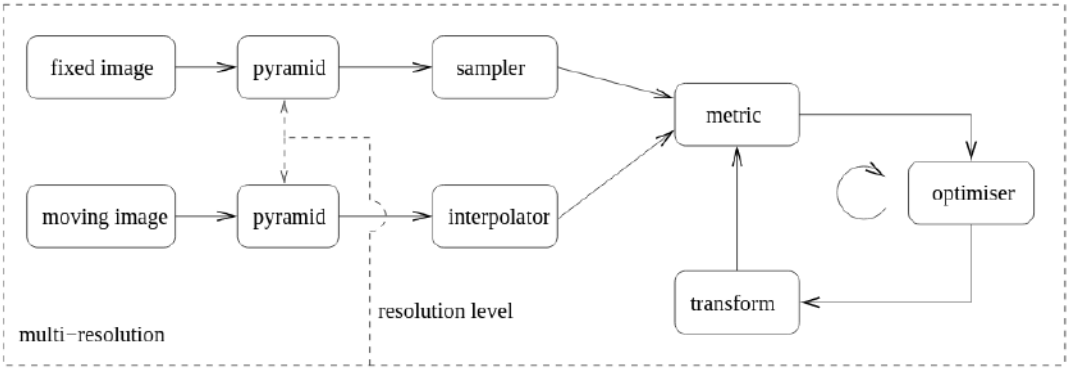

Figure 2.5: Mechanism of the registration.

robustness of the registration. The basic idea is to first estimate $T(x)$ on a low-resolution version of the images and then propagate the estimated deformation to higher resolutions. The sampler is responsible for selecting the locations in the input images for the metric. It is generally sufficient to evaluate a subset of randomly sampled voxels.

The similarity metric measures the degree of similarity between the moving and fixed images. The metric samples the intensity values of the fixed images and transformed moving images and evaluates the fitness value and derivatives, which are passed to the optimizer. When a point is transposed from one space to another by a transformation, it is usually transposed to an off-grid position. Interpolation is required to estimate the intensity of the image at the mapped off-grid position. Finally, the optimizer estimates the optimal parameters of the transformation.

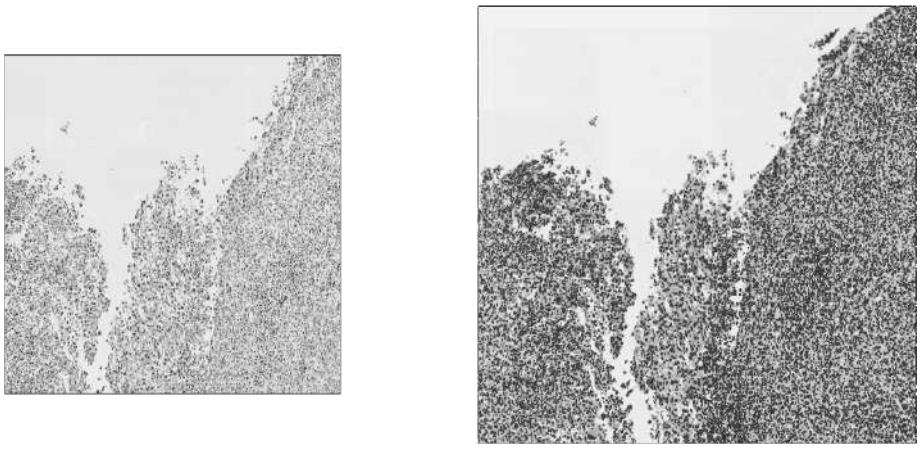

Figure 2.6: The image on the left is an H&E ROI of interest, and the image on the right is the extended mIF region.

In the project, the first registration used on ROIs consists of a translation transformation followed by an affine transformation. The objective is to align the two images at the global structure level. The affine transform allows for shearing, scaling, rotation, and translation. It is generally a good transform choice for initializing non-rigid transforms like the B-Spline transform that will be used later.

Once the rigid-affine registration is executed, we can apply the same registration on the set of cell positions obtained by inForm® for the assumed mIF ROI.

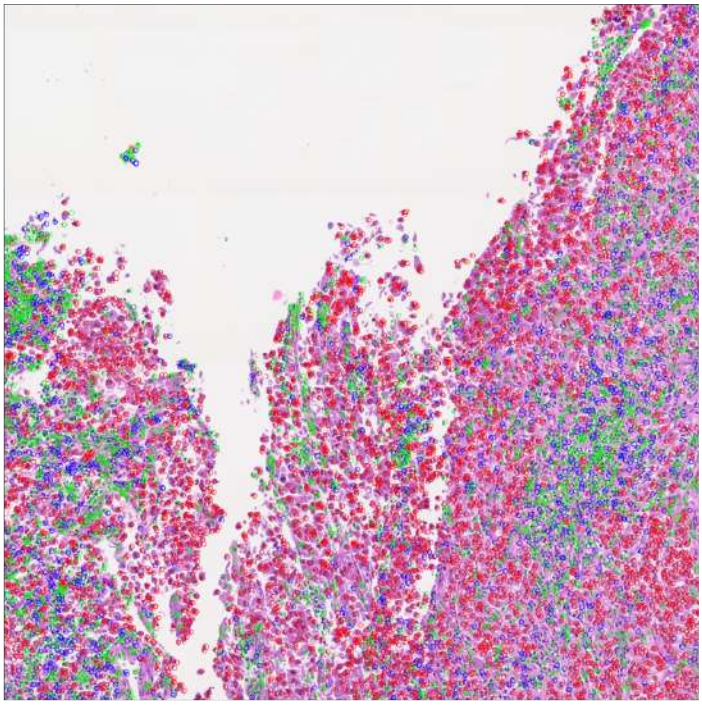

Figure 2.7: Cells position after the rigid-affine registration on the H&E. Red circle = SOX10p, Green circle = CD3p, CD20p and CD56p , Blue circle = CD68p and CD11Cp.

Non-rigid Registration on ROIs

The rigid-affine registration allows us to have a good alignment of the general structure of the two images, but as we want a registration at the level of the nuclei, the second type of registration is necessary to compensate for the more localized deformations. The non-rigid registration on ROIs uses B-Spline transforms. The B-spline transform is applied at each level of resolution, from the lowest to the highest.

The second registration can be a complex task depending on the size of the ROI, so to simplify the process, registration is done on pairs of small tiles (3,000 X 3,000 pixels).

The second advantage of using tiles is avoiding processing a region without tissue, which saves computational time.

The B-spline transform introduces a regular grid with N_x control points and 2 degrees of freedom for each of them. This gives high flexibility for an accurate alignment. The B-spline can be written:

$$T_\mu(x) = x + \sum_{x_k \in N_x} p_k \beta^3 (\frac{x - x_k}{\sigma})$$

Figure 2.8: B-Spline transformation: The equation describes the shifting of the current position x, towards the cubin interpolated B-spline control points x_k.

With x_k the control points, $\beta^3(x)$ the cubic multidimensional B-spline polynomial, p_k the B-spline coefficient vectors, α the B-spline control point spacing, and N_x the set of all control points within the compact support of the B-spline at x.

Once the non-rigid registration is executed, we can apply the same registration to all the cell positions obtained after the rigid-affine registration.

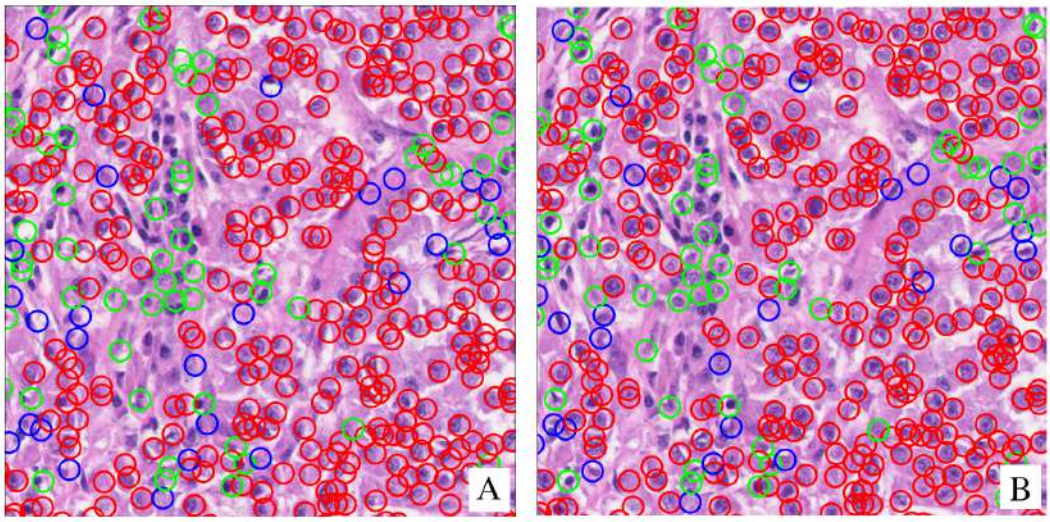

Figure 2.9: (A) Cells position after the rigid-affine registration, (B) Cells position after the non-rigid registration using B-Spline. Red = SOX10p, Green = CD3p, CD20p and CD56p , Blue = CD68p and CD11Cp.

HoVer-Net

The registration process results in assumed cell locations on the H&E. To perform quality control and filtering, we compare the cell locations resulting from the registration and

segmentation on the H&E. The segmentation on the H&E is achieved using HoVer-Net.

HoVer-Net is a model for simultaneous segmentation and classification in histology images. The network is based on the prediction of horizontal and vertical distances of nuclear pixels to their centers of mass, which are then leveraged to separate the clustered nuclei.

In addition to the cell location, we retrieve, from the HoVer-Net output, the binary mask of each nucleus. The binary mask will be helpful to the deep learning model to identify the cell it needs to predict correctly. We also use the binary mask to obtain other useful information for further filtering our dataset, such as the size of the nucleus.

The size is the primary filter of the dataset from the HoVer-Net output. Indeed, the nuclei in WSI are rarely perfectly cut in the center. Therefore, some nuclei are only partially visible. Removing these complex nuclei allows us to improve the quality of the dataset.

2.4 Network Architecture

Cell prediction is based on a combination of models. The cell is first sent to three binary models that will each identify whether the cell is a tumor cell, a lymphocyte, a myeloid cell, or another type of cell. Furthermore, if in the first step it has been identified as a lymphocyte, three models will determine whether the lymphocyte is a T cell, a B cell, or a natural killer. It will be done the same way in the case of a myeloid cell. Indeed, one model will try to specify the myeloid cell type of the cell, between dendritic and macrophage.

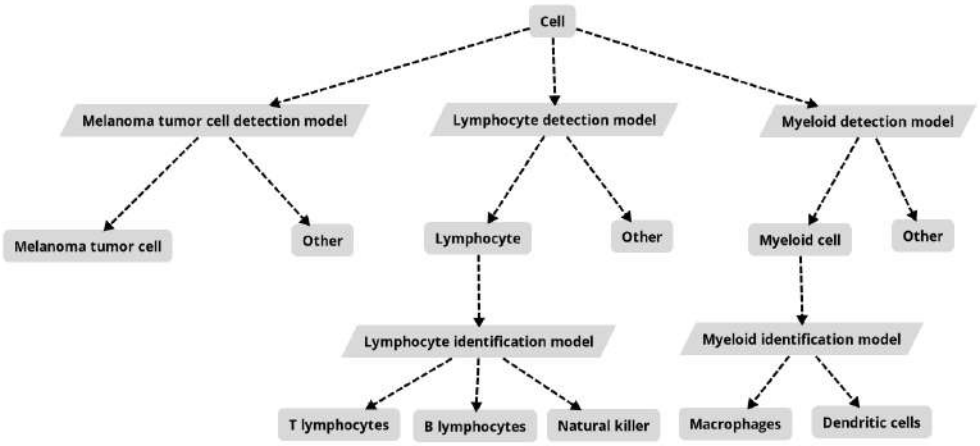

Figure 2.10: Global architecture with our different models

DenseNet

The architecture of each model used in the project is a Densenet [4]. DenseNet is an architecture proposed in 2017 and has been widely used for its performance.

DenseNet is a network architecture in which each layer is directly connected to all other layers in a feed-forward method in each dense block. For each layer, the feature maps of all previous layers are treated as separate inputs, while their own feature maps are passed as inputs to all subsequent layers.

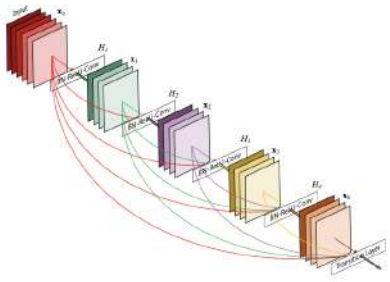

Figure 2.11: A 5-layer dense block. Each layer takes all preceding feature-maps as input.

As each layer receives the feature maps of all previous layers, the network can be thinner and more compact. Therefore, the number of channels can be reduced.

DenseNets have several clear advantages: they mitigate the vanishing-gradient problem, enhance feature propagation, encourage feature reuse, and significantly reduce the number of parameters.

RGB channels

The input feed to the model is composed of multiple channels of 64 by 64 pixels. The first three channels are the RGB channels of the patch centered on a cell at the higher magnification (40X), which correspond to 0.25 µm per pixel.

A common technique used in deep learning to improve a model's performance is data augmentation. The idea of data augmentation is to increase the amount of data by adding slightly modified copies of already existing data. The ability of the model to recognize cells in different contexts is therefore improved, which is an essential aspect of our model as we want our models to recognize cells on different WSIs.

Different transformations have been selected to improve our dataset:

- Horizontal and vertical flip.

- Gaussian blur.
- Adjustment of the sharpness of the image.
- Change of the brightness, contrast, saturation, and hue of an image.

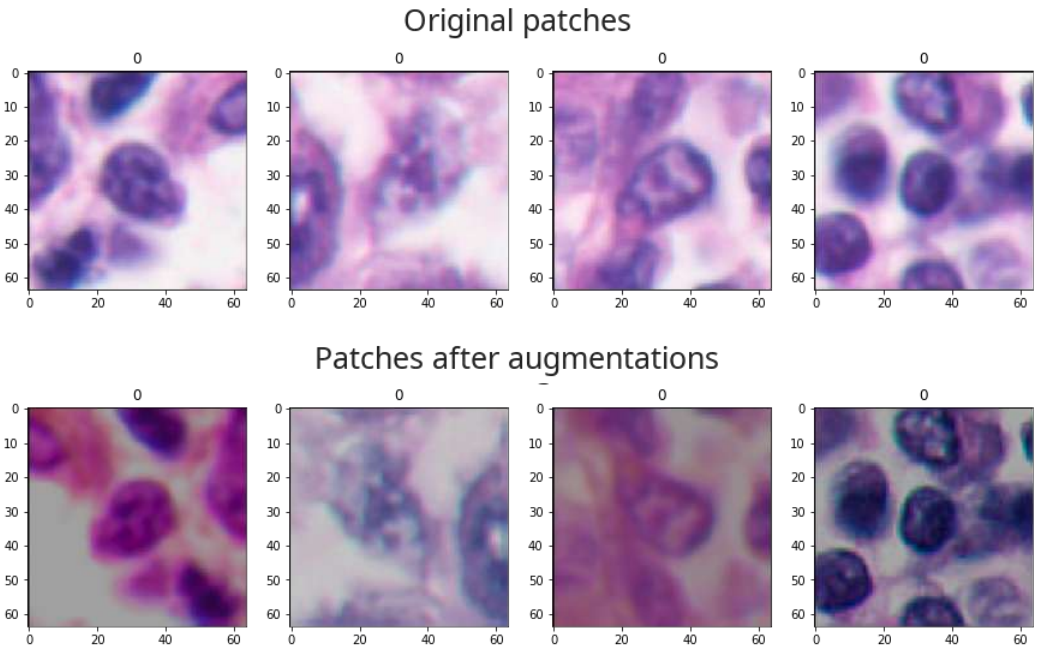

Figure 2.12: Four patches extracted from an H&E WSI and these same patches after augmentations.

Binary Mask Channel

One of the problems we faced because of the nature of our input data was the difficulty for our models to ensure that they identify the correct cell, as there may be many cells close together in an image. Therefore, to help the model focus on the area of interest, a fourth channel was created. This channel is a binary mask of the nucleus, thus identifying the correct cell.

Adding a fourth channel at the input is a simple modification of the network. It relies on the idea of giving information to the model without imposing it. The disadvantage is that it relies on the binary mask's accuracy. The binary mask we decide to use comes from the HoVer-Net model. The HoVer-Net model has proven its good performance over the years, but no model is infallible. Thus, a filtering is performed on the size of the mask because we have seen that even if it happens rarely, the binary mask can still be excessively

large and covers several cells. Then, a modification of the mask is performed to allow more flexibility. The mask may not always be correct; a slight shift to the ground truth is inevitable. To compensate for the uncertainty, we add a Gaussian filter to the binary mask.

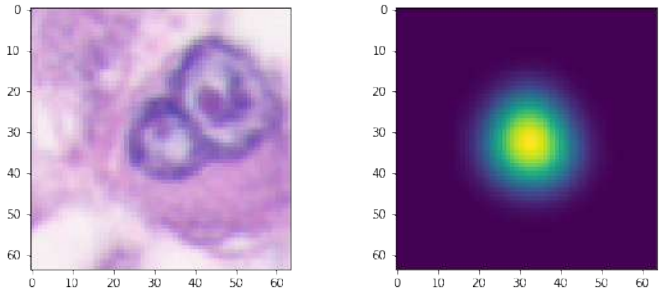

Figure 2.13: Patch extracted from an H&E WSI and the corresponding binary mask with Gaussian filter.

Datasets

Three datasets were used for this project. One was created with our registration and the other two are external datasets used to validate our models on external tissue. Of these two external datasets, one is a well-known dataset named CoNSeP, and the other is an unpublished dataset from the CHUV called DeepMEL.

3.1 Melanoma Tissues

As part of this work, we introduce a new dataset consisting of 178 H&E stained image tiles, each of size 6,000×6,000 pixels at 40X objective magnification. Images were extracted from an FFPE block of a melanoma and scanned with a Vectra Polaris Imaging System from Akoya within the ILL laboratory at the CHUV in Lausanne, Switzerland. Pathologists have not created the dataset. The positions of cells have been obtained with the HoVer-Net model and inForm®. The labeling of cells has been achieved with the inForm® analysis software in the mIF and the registration described in the section method. The dataset is composed of six labels: melanoma tumor cell, T cell, B cell, natural killer, macrophage, and dendritic cell.

The distribution in fig. 3.1 shows a wide disparity between the amount of each protein. The majority of the dataset is composed of melanoma tumor cells, with about 67%, lymphocytes represent 27%, and myeloid cells only 5%. We particularly note a very low number of nuclei identified with CD68, CD11C, and CD56.

Our really unbalanced dataset is a problem for training our models. To counter this problem, different methods exist and were used in this project. First, by adding class weights, which gives different weights to the majority and minority classes. Second, by oversampling the three underrepresented classes.

In addition to these six labels, another one was created, called "other", which consists of 100,000 cells identified by HoVer-Net but not by inForm®. This addition helps our models

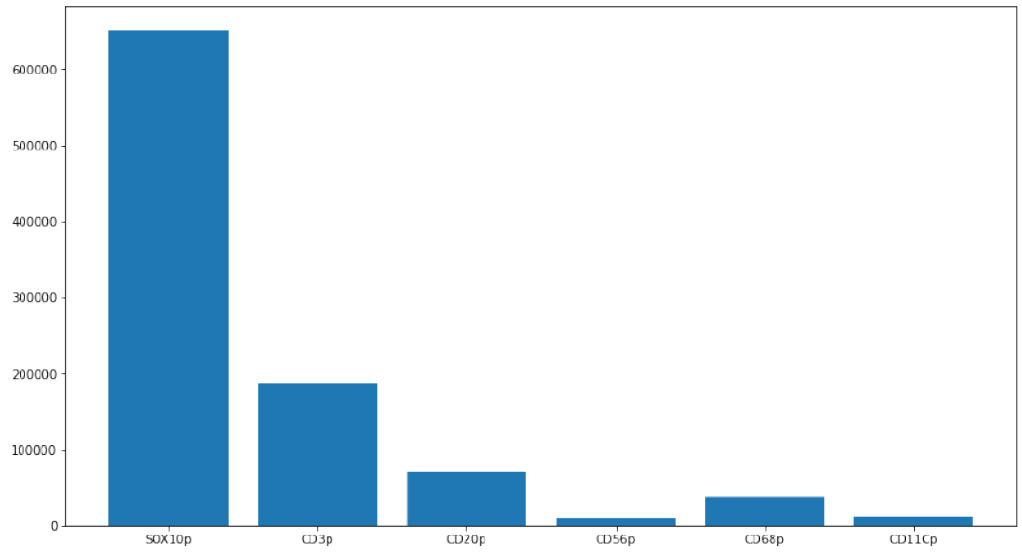

Figure 3.1: Distribution of each protein in the dataset created with the registration on the three melanoma tissues.

to cope with the diversity of cells that our models are expected to encounter in the tissues.

3.2 Dataset to Test the Generalizability of the Models

CoNSeP

The CoNSeP dataset [3] is a publicly available dataset comprising both the segmentation masks and class labels. It consists of 41 H&E stained image tiles, each of size 1,000×1,000 pixels at 40X magnification. CoNSeP contains image regions extracted from WSIs from University Hospitals Coventry and Warwickshire (UHCW) and are extracted from 16 WSI colorectal adenocarcinomas (CRA), each belonging to an individual patient, and scanned with an Omnyx VL120. The dataset contains 24,319 exhaustively annotated nuclei with associated class labels from pathologists. Each nucleus is labeled by one of the seven categories: inflammatory, healthy epithelial, dysplastic/malignant epithelial, fibroblast, muscle, endothelial and other.

Unfortunately, only one label is interesting for the project. Indeed, only the inflammatory label is useful because it can be associated with our lymphocyte label. This is a recurring problem, as current datasets do not have a label for myeloid cells or melanoma tumor cells or the specific type of lymphocyte (e.g., T cells).

DeepMEL

DeepMEL is a project of the CHUV. The main objective of this project is to use machine learning techniques to train a deep learning model to identify lymphocytes in melanoma tissues and then apply feature extraction and survival analysis on the patient based on the clustering of these cells.

The project consists of 127 H&E stained image tiles, each of size 1,000×1,000 pixels at 40X magnification. Due to the current lack of labels for these tiles, we had to use the help of a pathologist to estimate the accuracy of our models on two selected tiles.

Experimental Design

4.1 Registration

The pipeline for registration is tested on our three available H&E WSIs as specified in the datasets section. Each WSI is divided into ROIs of size 6,000 by 6,000 pixels (1,500 by 1,500 μm). The pipeline records the registration status for each major process step. Three major stages are identified: rigid-affine registration on ROIs, non-rigid registration on ROIs, and mapping to H&E segmentation. The registered states are visualizations to evaluate the current registration. Quantitative evaluations are not used because it is difficult to obtain good measurements due to the different nature of the two images being compared (H&E and mIF).

The rigid-affine registration on the ROIs is a structure-level registration. The visualization that has been chosen to check the correct behavior is, therefore, a checkerboard. A BWR visualization is more appropriate for the non-rigid registration on ROIs because it allows better analysis of the registration at the nucleus level. Finally, the mapping to the H&E segmentation is verified by manually checking many H&E patches.

From the registration output, three datasets are created, a training dataset consisting of two WSIs, a validation dataset consisting of half of the third WSI, and a test dataset consisting of the other half of the third WSI. The datasets created with the registration are error-prone. Even if the registration is visually verified, the inForm® labeling can still be error-prone. For this reason, the three datasets were partially verified by comparing the inForm® label and a visual estimate of the mIF label. The dataset was also filtered in various ways to mitigate errors from inForm® and various other factors. The primary filter is the confidence value. As discussed in the methods section, the confidence value is the probability of the winning phenotype. Therefore, the higher the value, the more likely we can say that our dataset is correct. In our work, a confidence value of 65% was decided because it eliminates most of the mislabeled cells without eliminating too many correct

cells. Indeed, the higher the confidence, the lower the number of cells in the dataset. Another filter was the size of the nuclei. Some identified cells are only partially visible, making it impossible to identify their type. These cells are mainly the result of tissue cutting. Thus, the size of the nucleus obtained with HoVer-Net was used as a value for the second filter. This filtered dataset is checked manually, examining 100 cells of each cell type. For each cell, a visual estimation is performed using each channel of the mIF image at the nucleus location and the H&E at different magnifications.

4.2 Deep Learning Models

The models used for cell identification are DenseNet models. The DenseNet parameters are as follows:

- Growth rate = 32 - How many filters to add each layer.
- Block config = (2,2,2,2) - How many layers in each pooling block.
- Num init features = 80 - The number of filters to learn in the first convolution layer.
- Bn size = 4 - Multiplicative factor for number of bottle neck layers.
- Drop rate = 0 - Dropout rate after each dense layer.
- Num classes = 2 - Number of classification classes.

Training is performed using a batch size of 64 with a learning rate of 0.002. The dataset used for training is the same as described above, with one difference. The number of CD56p, CD68p, and CD11Cp cells is too low to train our models properly. Therefore, two modifications were made: First, by adding a class weight, which gives different weights to the majority and minority classes. Second, by oversampling the three underrepresented classes.

The evaluation of the models is performed on different tissues. First, on the test dataset, which is similar to the training dataset because the tissue is equivalent (same FFPE block, same scanner, similar staining). Second, on tissue from other projects (CoNSeP and DeepMEL).

To evaluate the model, different metrics are used: accuracy, recall, precision, and f1-score. In addition to these metrics, a pathologist evaluates the performance of our models on other tissues to confirm the metrics or to compensate for the absence of these metrics.

The CoNSeP dataset is the only one of our external datasets with ground truth. The only problem with this dataset is that the available labels are not the ones we are interested in. Indeed, as said in the dataset section, only the inflammatory label is useful because it can be associated with our lymphocyte label.

As specified in the methods section, the input feed to the models is composed of four channels. To evaluate the addition of the fourth channel, a comparison is made with two models with and without the fourth channel. The models used for this experiment is the melanoma tumor cell model and myeloid cell model.

Results

5.1 Image Registration

Rigid-Affine Registration on ROI

In order to evaluate the registration performance, a visual comparison between the H&E and the registration output at each step has been made. Quantitative evaluations are less used for the registration because it is challenging to have good metrics due to the different nature of the two images (H&E and mIF).

The first visualization is done after the rigid-affine registration on the ROI, as the first registration objective is to align the two images at the global structure level. A checkerboard testifies that the global structure coincides between the two images. Indeed, checkerboards are well suited when the squares used for the visualization are large enough.

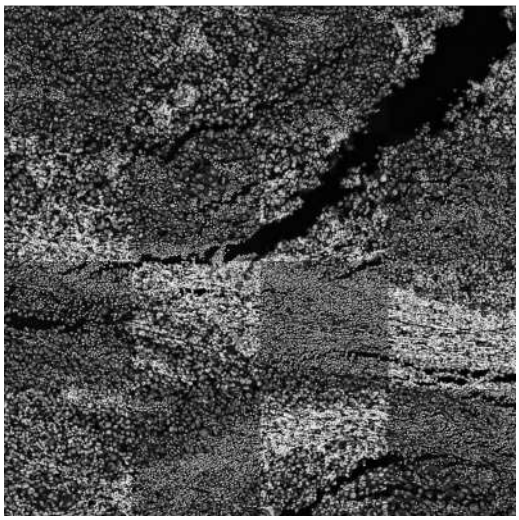

Figure 5.1: CheckerBoard between H&E ROI and the corresponding mIf ROI after the rigid-affine registration.

Of the 178 ROIs we have, the registration went well for 96% of them. Unfortunately, the registration can fail, primarily if the tissue is composed of dense regions or has a repetitive pattern. Indeed, those two scenarios create difficult local minimums for the similarity metric (C) used in the registration. The registration can therefore be stuck in those local minima and can not find a path to the global minimum.

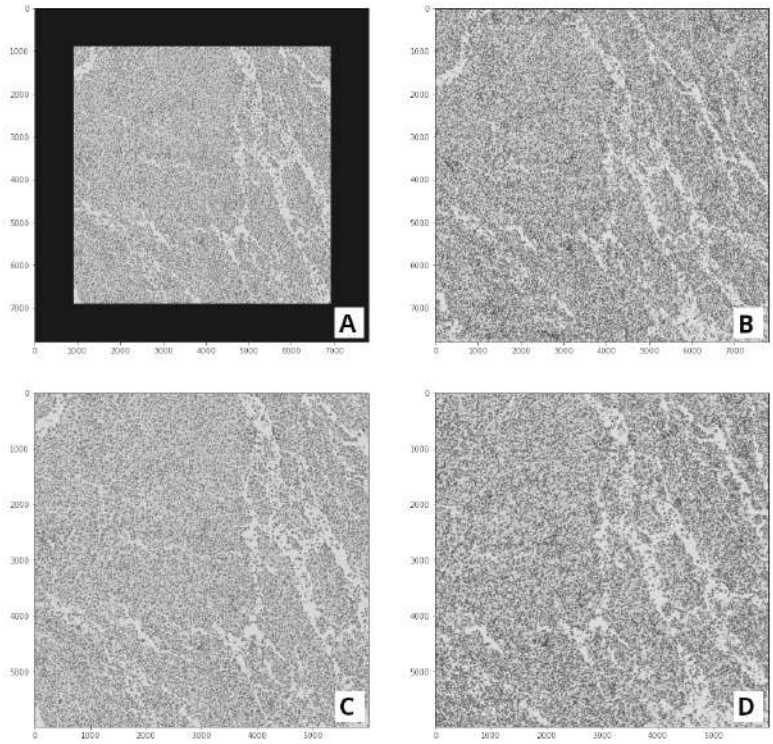

Figure 5.2: Failed rigid-affine registration on an ROI with a repeating pattern. (A) H&E ROI with the border. (B) mIF ROI before the registration. (C) H&E ROI without the border. (D) mIF ROI after the registration.

To solve these rare cases, it was decided to manually assist the recalibration by transforming the H&E and mIF ROIs by placing a visual marker on each image. The newly added marker is designed to significantly impact the similarity metric, thus escaping the previous local minimum. The marker is placed in approximately the same location on both images, with an easily recognizable location being preferred.

Non-rigid Registration on ROI

The non-rigid registration uses a B-spline transformation. Due to the H&E discoloration and the second staining, local deformation appears around the tissue. Those deformations cannot be corrected with the previous registration. Therefore, the second type of

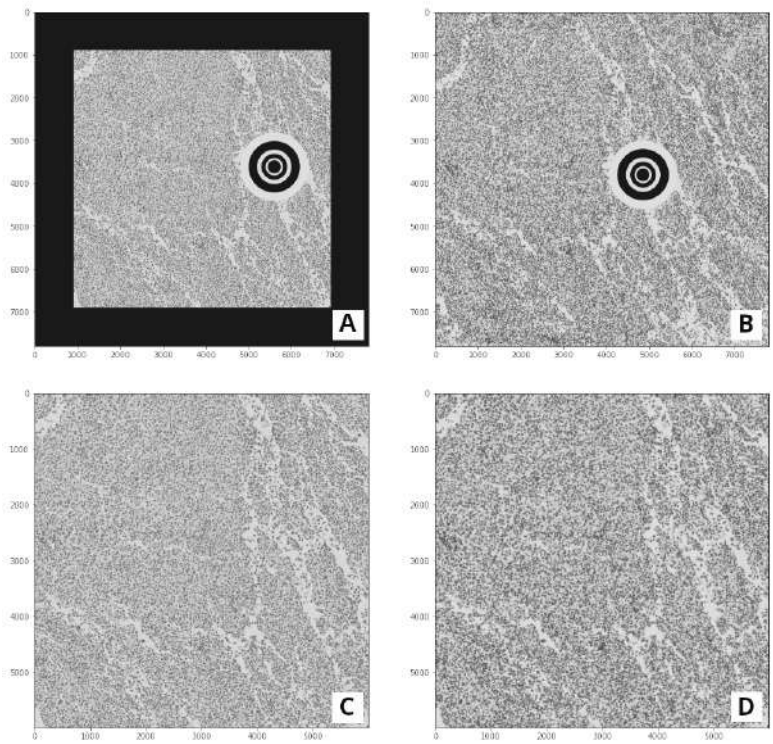

Figure 5.3: Correction of a rigid-affine registration failure on an ROI with a repeating pattern. (A) H&E ROI with the border and the marker. (B) mIF ROI with the marker before the registration. (C) H&E ROI without the border. (D) mIF ROI after the registration.

registration was necessary. This non-rigid registration (B-spline transformation) locally modifies the image, but as we want to keep a coherent tissue architecture, we decided to use a regularization term that penalizes transformations that involve too much variability. Our results show good performance during this registration.

BWR visualizations [fig. 5.4] examine the difference between the H&E and mIF ROIs after registration. The visualization is composed of three colors: blue, white, and red. In a perfect world, the visualization is entirely white when the two images are identical. Red and blue indicate a difference between the two images, which we want to minimize. The blue color means the presence of tissue for the H&E only and the red color for the mIF only.

The example in fig. 5.4 has a lot of red and blue at first sight. However, they are, for the most part, the contour of a white globular area. It means that cells in H&E and mIF are both centered, but we have a difference in the size of the cells.

Indeed, with destaining and re-staining, the cells are distorted and sometimes become

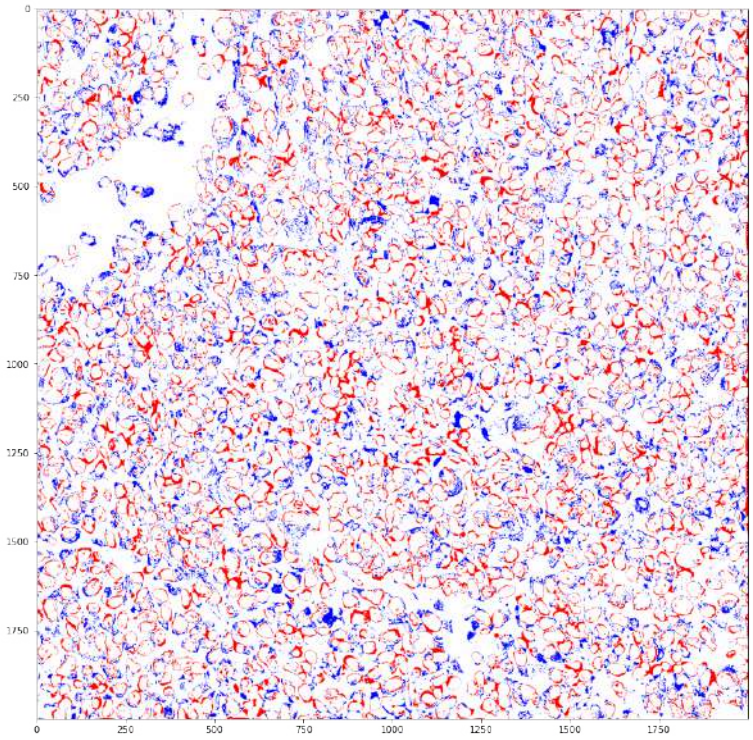

Figure 5.4: BWR visualization on a region of an ROI.

larger or smaller than when they were first stained. Therefore, even if the H&E cell and the mIF cell are aligned, we will have a difference in tissue at the edge of the cell.

Mapping with the H&E Segmentation

The mapping to the H&E segmentation enables to have H&E centered patches. Indeed, the dataset would be mainly composed of uncentered cells without it. For example, in the inherited project, the dataset had many patches of cells that were not centered fig. 5.5.

Figure 5.5: Examples of patches at the beginning of the project.

With mapping, no patches were found with this problem. Thousands of cells were

examined, and each cell was centered. Example of patches after the mapping in fig. 5.6.

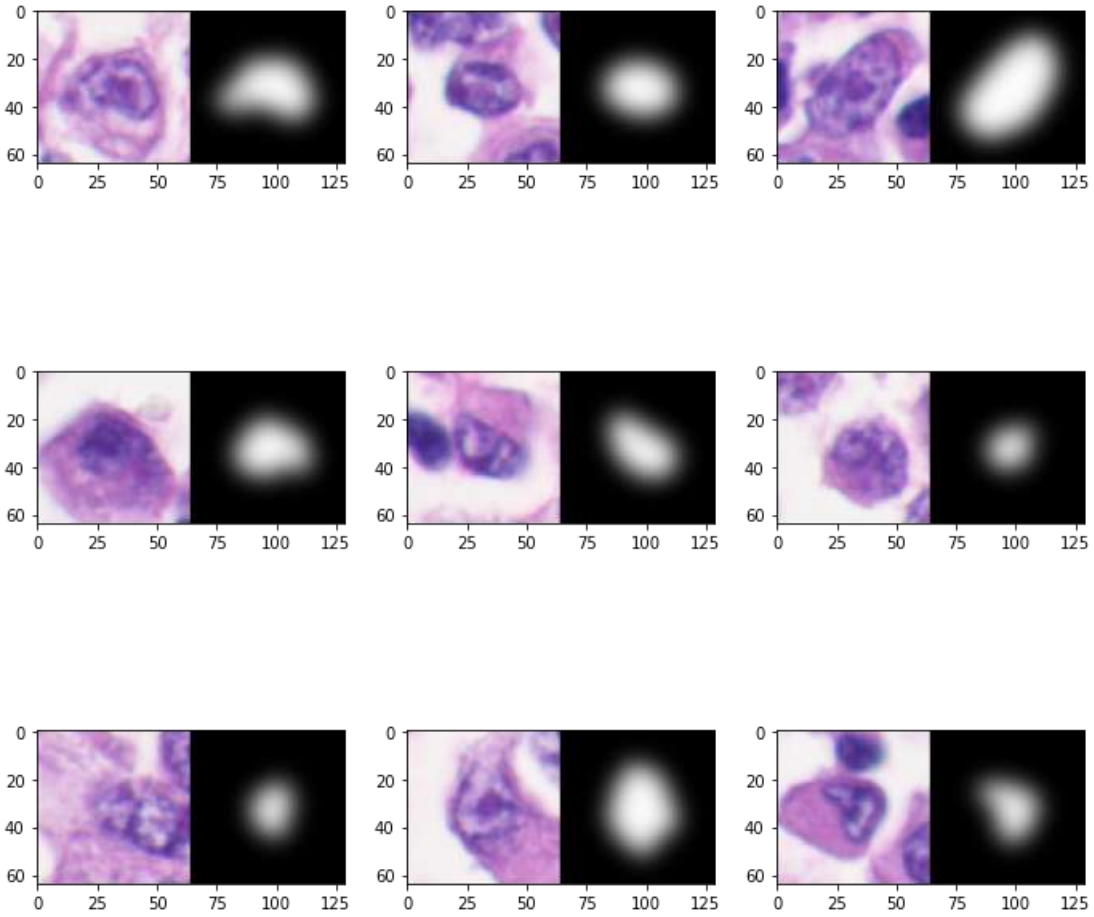

Figure 5.6: Examples of patches after mapping to the H&E segmentation. Each H&E patch is shown next to its nuclei mask.

The New Dataset Created with the Registration

In total, in our three datasets created with the registration, we identified 965,169 cells. The dataset is composed of the six different cell types. To which we add a group "others" to improve the performance of our models on external tissues. The composition of the cells obtained from the three tissues can be found in table 5.1. We can see that the dataset is generally well balanced between the three WSIs, although WSI 1 has more CD56p than the other two.

A visualization representing the cells composing our dataset in a particular section of a WSI can be found in fig. 5.9. One can notice the precision of the registration since the circles are always centered on a nucleus.

Table 5.1: Number of each cell type in the dataset for each WSI.

	SOX10p	CD3p	CD20p	CD56p	CD68p	CD11Cp
WSI 1	210461	55787	26868	4157	13920	2871
WSI 2	232632	62707	22030	2350	9877	3285
WSI 3	207445	68404	21591	2432	14468	3884

However, as discussed in the experimental design section, the labeling of the inForm® can still be subject to error. The comparison between the label of inForm® and a visual estimate of the label from the mIF is given below in table 5.2.

Table 5.2: Estimation of labeling error in a dataset. The dataset used is from a single WSI. The error estimation is performed visually on the mIF with 100 cells for each cell type.

True label	Myeloid cell	Lymphocyte	melanoma tumor cell	other
Labeled by inForm®:				
Myeloid cell	80%	16%	2%	2%
Lymphocyte	2%	80%	0%	18%
melanoma tumor cell	0%	3%	97%	0%

Visual estimation is performed using each channel of the mIF image at the location of the nuclei, as well as the H&E at different magnifications. An example can be seen in fig. 5.7, where a cell was identified as CD68p by inForm®, but by looking at the mIF, we can confirm that the labeling is wrong. The proper label should be CD3p.

One particular problem was identified during this analysis. InForm® labeled many cells as CD20p, but no trace is visible in the mIF of the protein. This error is significant as it was estimated to represent 17% of the lymphocytes identified by inForm®. An example can be seen in fig. 5.8. These mislabeled cells tend to cause our lymphocyte model to classify a wide variety of cells as lymphocytes, even if they are not.

5.2 Deep Learning Models

The large dataset created allows us to train models that could never have been trained with datasets created only with manual labeling. Our first goal was to train three binary models to identify lymphocytes, melanoma tumor cells, and myeloid cells.

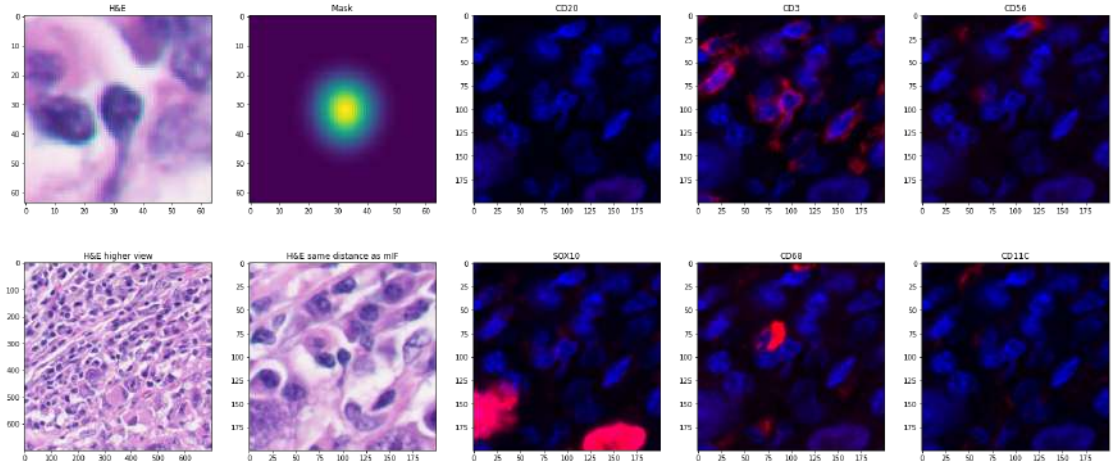

Figure 5.7: Example of mislabeling by inForm®. The cell is identified by inForm® as CD68p, but visually it can be seen that it should be CD3p. The cell that should have the CD68p label is the neighboring cell on the top left. The six images on the right are mIF images. The blue color represents the DAPI channel, and the red color represents the protein channel.

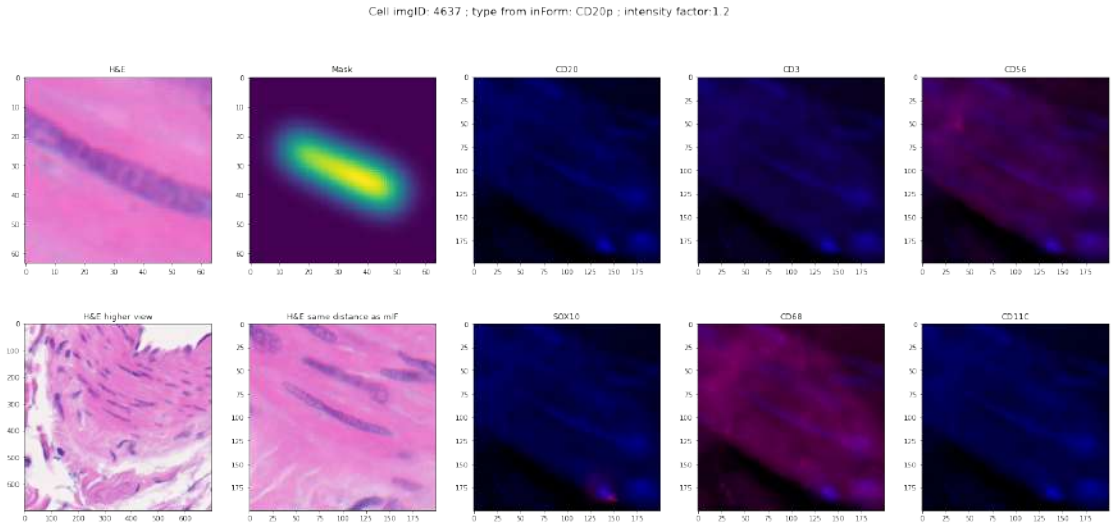

Figure 5.8: Example of mislabeling by inForm®. The cell is identified by inForm® as CD20p, but visually it can be seen that it should not be CD20p. The six images on the right are mIF images. The blue color represents the DAPI channel, and the red color represents the protein channel.

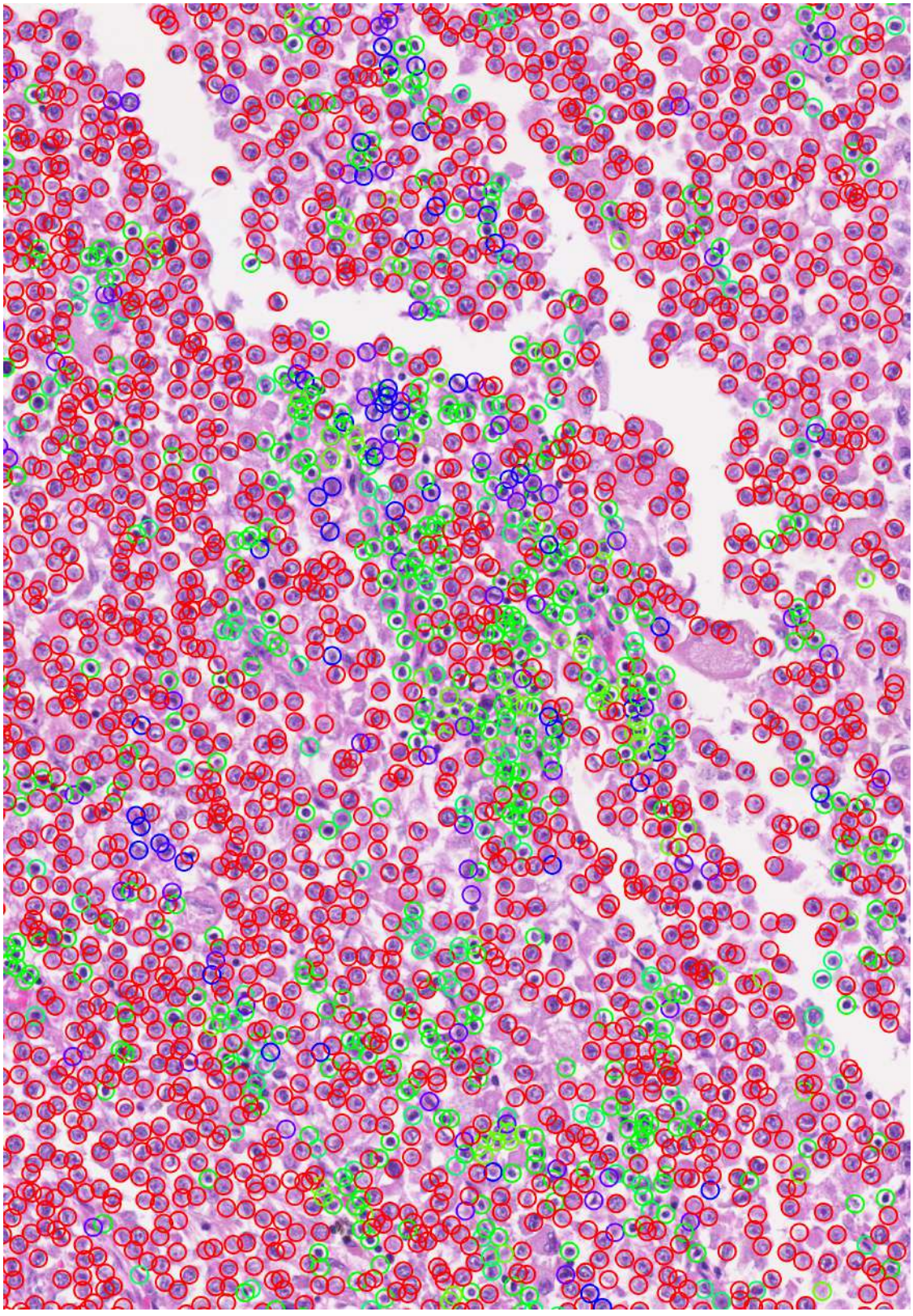

Figure 5.9: Position of cells after complete registration on a section of a H&E WSI. Red circle = SOX10p, green circle = CD3p, CD20p and CD56p, blue circle = CD68p and CD11Cp.

Lymphocyte Model

The lymphocyte model performs very well based on the metrics in the test dataset (see table 5.3). The accuracy is about 93% on the test dataset, which is a significant improvement over the 78% obtained at the beginning of the project for a similar task. Furthermore, the training shows no visible problems such as overfitting or underfitting (see fig. 5.10).

Table 5.3: Metrics of the lymphocyte model on the test dataset.

	Precision	Recall	F1-score	Support
Other	0.91	0.98	0.94	163478
Lymphocyte	0.94	0.74	0.83	60000
Accuracy			0.93	223478

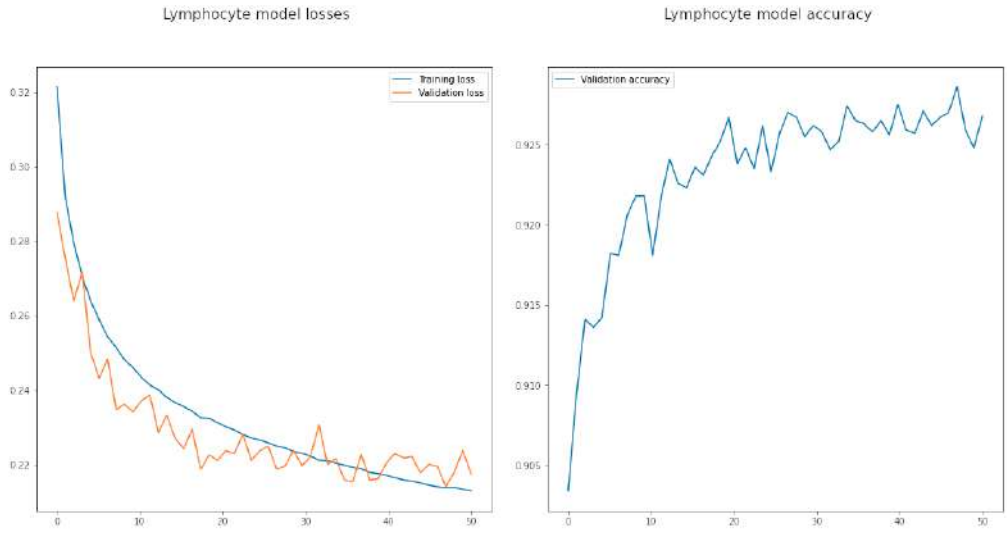

Figure 5.10: Training loss, validation loss, and validation accuracy of the lymphocyte model for 50 epoch training.

The qualitative results on the test dataset are also convincing and reflect the metrics (fig. 5.11). The model demonstrates good precision but misses some lymphocytes, which explains the lower recall. Looking at the problem at a more structural level, the model works well because the false positives are spread over the whole tissue, so we are not missing a specific region.

It is in the second part of the evaluation that we have the most difficulty obtaining good results because our models must be able to generalizable to different tissues, scanners, and staining protocols.

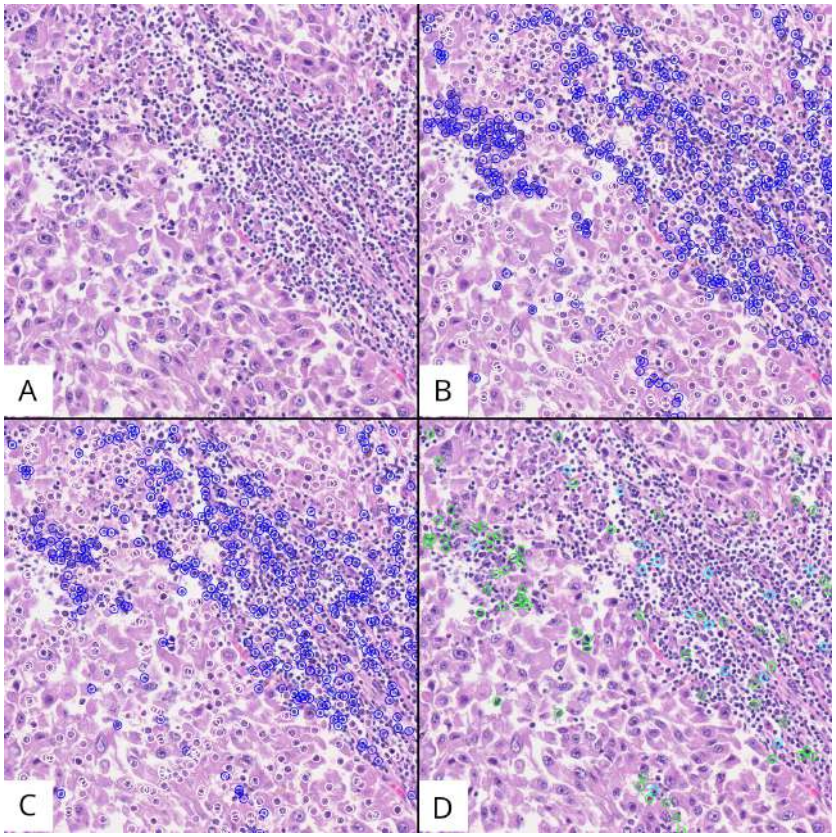

Figure 5.11: Example of comparison of the lymphocyte model with the test data set. Image A: The H&E area of interest. Image B: The ground truth obtained with the registration and the inForm® output. Blue circles are cells identified as lymphocytes. White circles are cells identified as something else. Image C: The prediction of the lymphocyte model on the tissue. Image D: The difference between truth and prediction. The blue circles are false positives, and the green circles are false negatives.

Two datasets are available: CoNSeP and DeepMEL. The CoNSeP dataset is the only one where we have ground truth for lymphocytes. Therefore, it is the only one where we can have metrics.

Table 5.4: Metrics of the lymphocyte model on the CoNSeP dataset.

	Precision	Recall	F1-score	Support
Other	0.93	0.84	0.89	6194
Lymphocyte	0.52	0.75	0.62	1422
Accuracy			0.82	7616

The qualitative results on a CoNSeP tissue in fig. 5.13 demonstrate a good ability to detect lymphocytes. However, we face a problem with the number of false positives, as seen

in the metrics. This problem is mainly due to the dataset we currently use for training. Indeed, the analysis of the dataset showed some errors. The inForm® output is incorrect for some lymphocytes, especially those labeled as CD20p, many of which are not proper lymphocytes (about 18%). These mislabeled cells tend to cause our lymphocyte model to classify a wide variety of cells as lymphocytes.

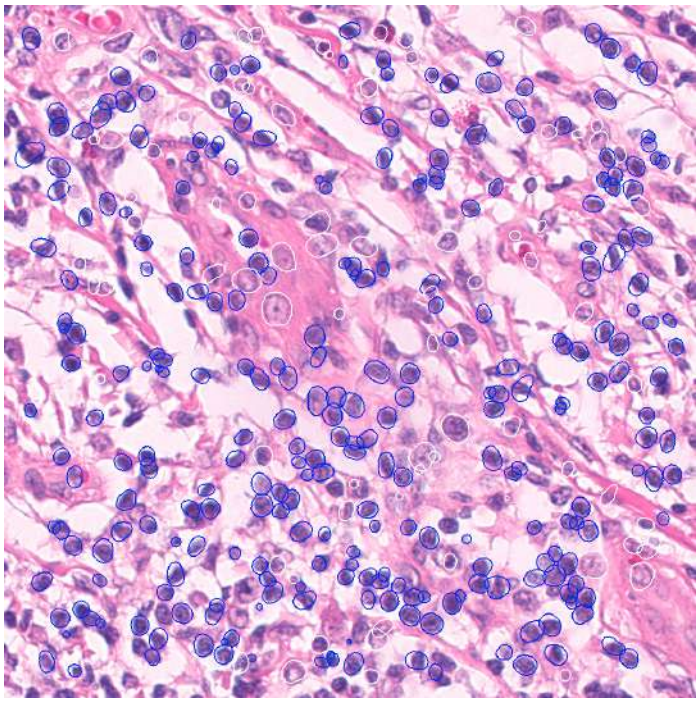

Figure 5.12: Example of the lymphocyte model on tissue from the DeepMEL project. The blue circles are cells identified as lymphocytes. The white circles are cells identified as something else.

Concerning the qualitative result of the model on the DeepMEL dataset: The lymphocyte detection seems to have the right sensitivity estimate between 80-90%. This result and the other two on the DeepMEL dataset need to be taken with precaution. As mentioned by our pathologist Dr. Pierre Moulin, a proper evaluation of the accuracy should take more examples and give more context, i.e. larger images. However, due to time constraints, only two images could be analyzed.

Nevertheless, this result helps us to have a good estimate of the performance of our models on external tissues. It gives us a precious hint on the behavior of our models. For example, it has been noticed that our lymphocyte model tends to it classifies some tumour cells as lymphocytes.

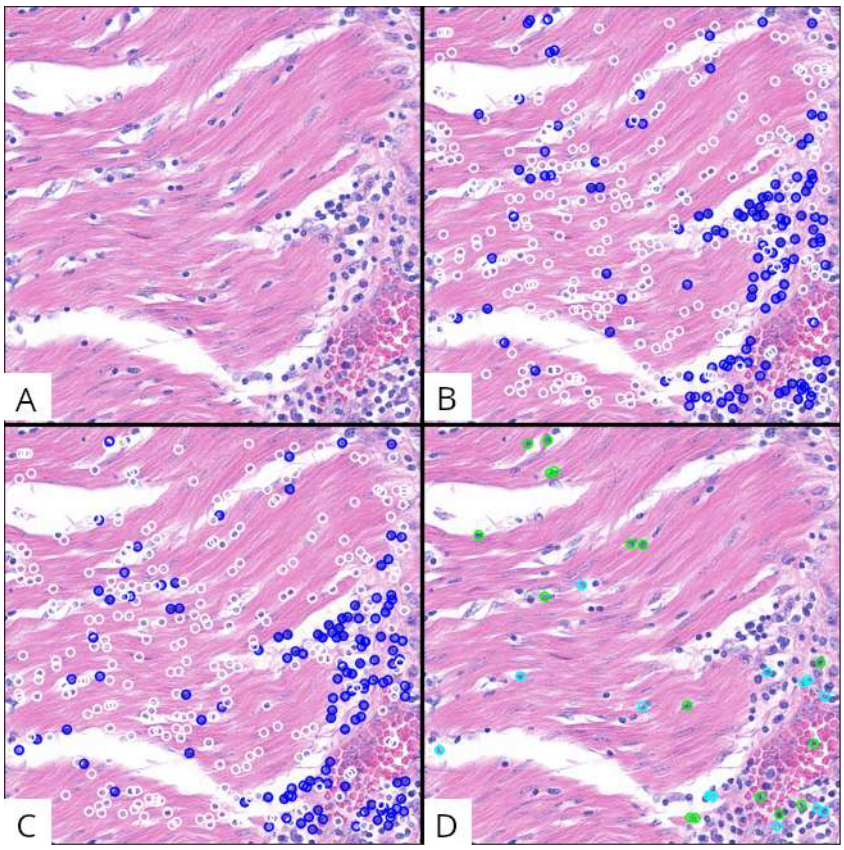

Figure 5.13: Example of the lymphocyte model compared to the CoNSeP dataset. Image A: The H&E area of interest. Image B: The ground truth. Blue circles are cells identified as lymphocytes. White circles are cells identified as something else. Image C: The prediction of the lymphocyte model on the tissue. Image D: The difference between truth and prediction. The blue circles are false positives, and the green circles are false negatives.

Melanoma Tumor Cell Model

The melanoma tumor cell model was our best model based on the metrics in the test dataset (table 5.5). It achieved 95% accuracy, with precision and recall of 97% and 96%, respectively. This result is related to the dataset's excellent quality and the task's simplicity. The dataset is near perfect for melanoma tumor cells (SOX10p) because there is low ambiguity on the mIF marker (fig. .1), and only a few errors were found (table 5.2). The task is also more straightforward because melanoma tumor cells are less likely to resemble myeloid cells or lymphocytes.

The qualitative results on the test dataset are also good and reflect the metrics (fig. 5.15). We have few false positives and false negatives, which corroborates the excellent precision and recall obtained in the metrics.

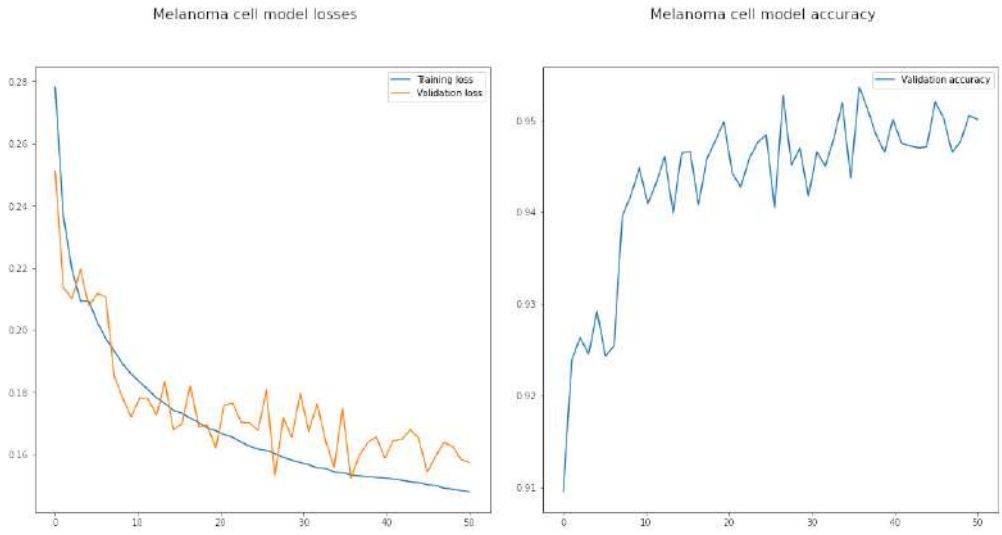

Figure 5.14: Training loss, validation loss, and validation accuracy of the melanoma tumor cell model for 50 epoch training.

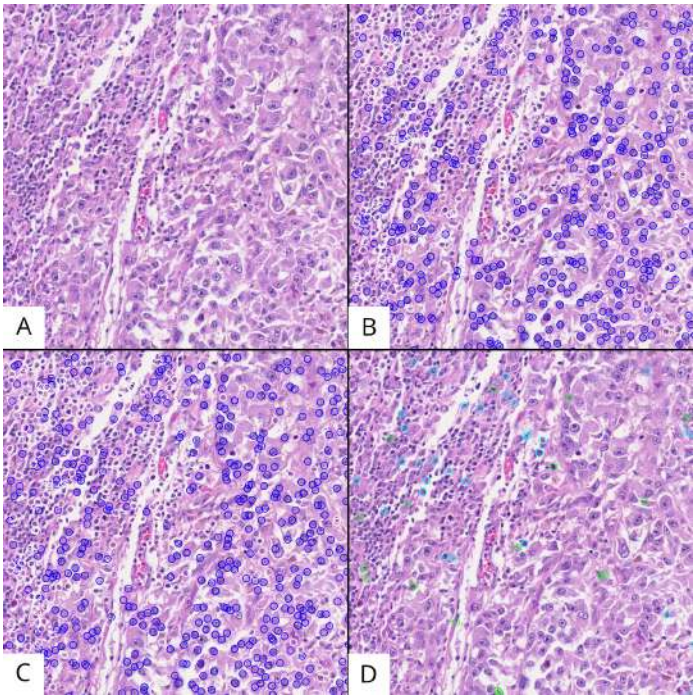

Figure 5.15: Example of comparison of the melanoma tumor cell model with the test dataset. Image A: The H&E area of interest. Image B: The ground truth obtained with the registration and the inForm® output. Blue circles are cells identified as melanoma tumor cells. White circles are cells identified as something else. Image C: The prediction of the melanoma tumor cell model on the tissue. Image D: The difference between truth and prediction. The blue circles are false positives, and the green circles are false negatives.

Table 5.5: Metrics of the melanoma tumor cell model on the test dataset.

	Precision	Recall	F1-score	Support
Other	0.91	0.94	0.93	72834
melanoma tumor cell	0.97	0.96	0.96	150644
Accuracy			0.95	223478

The second part of the evaluation is done on the DeepMEL dataset. The absence of ground truth does not allow us to have metrics. However, the accuracy was estimated by a pathologist on two images.

The pathologist believes that the melanoma cell model is rather sensitive. However, it could still be improved. Many false positives have been found, for example: large cells with large nuclei belonging to arteries are mislabeled. This is a tricky problem because there are many different cell types in the images. To solve this problem, our dataset should include more cells of different types labeled as "other" so that our models are more robust to the different cells that can be encountered.

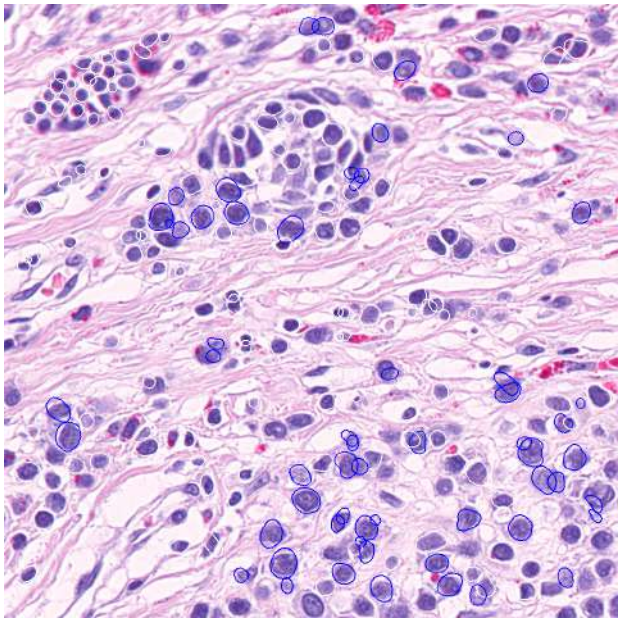

Figure 5.16: Example of the melanoma tumor cell model on tissue from the DeepMEL project. The blue circles are cells identified as melanoma tumor cells. The white circles are cells identified as something else.

In addition to these analyses, another was performed to evaluate the improvement of adding the fourth channel designed to help the model recognize the correct cell to label.

Surprisingly, the melanoma tumor cell model with only three channels gave almost the same result [table 5.6].

Table 5.6: Metrics of the myeloid model with only 3 channels on the test dataset.

	Precision	Recall	F1-score	Support
Other	0.92	0.94	0.93	72834
Melanoma tumor cell	0.97	0.95	0.96	150644
Accuracy			0.93	223478

The presumed reasons for this result is the size of the dataset and the size of the nuclei of the melanoma tumor cells. Because melanoma tumor cells are larger in size than the other two categories, a model wishing to identify melanoma tumor cells would have to mainly learn to differentiate a large cell from two small cells that are close together. With over 440,000 melanoma tumor cells in the training dataset, the model should easily understand when there is ambiguity about which cell the model should focus on.

This is not the case for myeloid cells because the number of myeloid cells is smaller in our dataset, and the size is only not as crucial for the model. Lymphocytes being of similar dimensions, the shape and structure of the cells should be more critical for the model in the first place.

Myeloid Model

The myeloid model is our worst model based on the metrics on the test dataset in table 5.7. The accuracy is high (93%), but this is due to the low proportion of myeloid cells in the dataset. Indeed, the myeloid cells represent only 5% of the dataset, so it is difficult to correctly evaluate the model just with this metric. Precision and recall help us to better understand the accuracy of the model. Precision and recall are low compared to the previous two models. Two reasons for this: the number of myeloid cells is smaller and the labeling by inForm® is prone to errors (see table 5.2).

Table 5.7: Metrics of the myeloid model on the test dataset.

	Precision	Recall	F1-score	Support
Other	0.97	0.95	0.96	210644
Myeloid cell	0.44	0.59	0.50	12834
Accuracy			0.93	223478

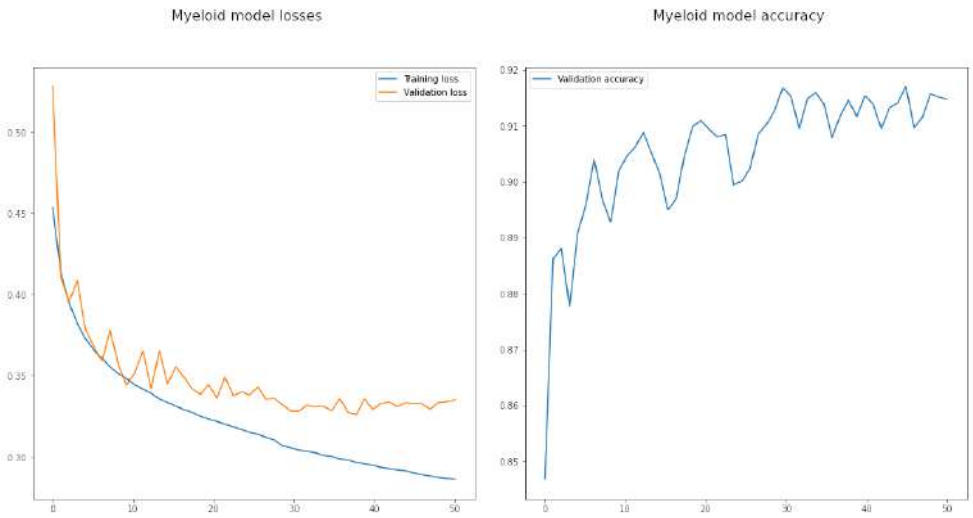

Figure 5.17: Training loss, validation loss, and validation accuracy of the myeloid model for a 50 epochs training.

The qualitative results on the test dataset are not as bad as one might think (fig. 5.18). Some false positives are found around true myeloid cells, which is a good sign. Myeloid cells are challenging to identify with mIF. Often we know that a myeloid cell is in a particular area, but it is difficult to know which cells in that area are the true myeloid cells. This scenario can explain some false positives. But it can't explain everything. As we said before, the main problem comes from our dataset.

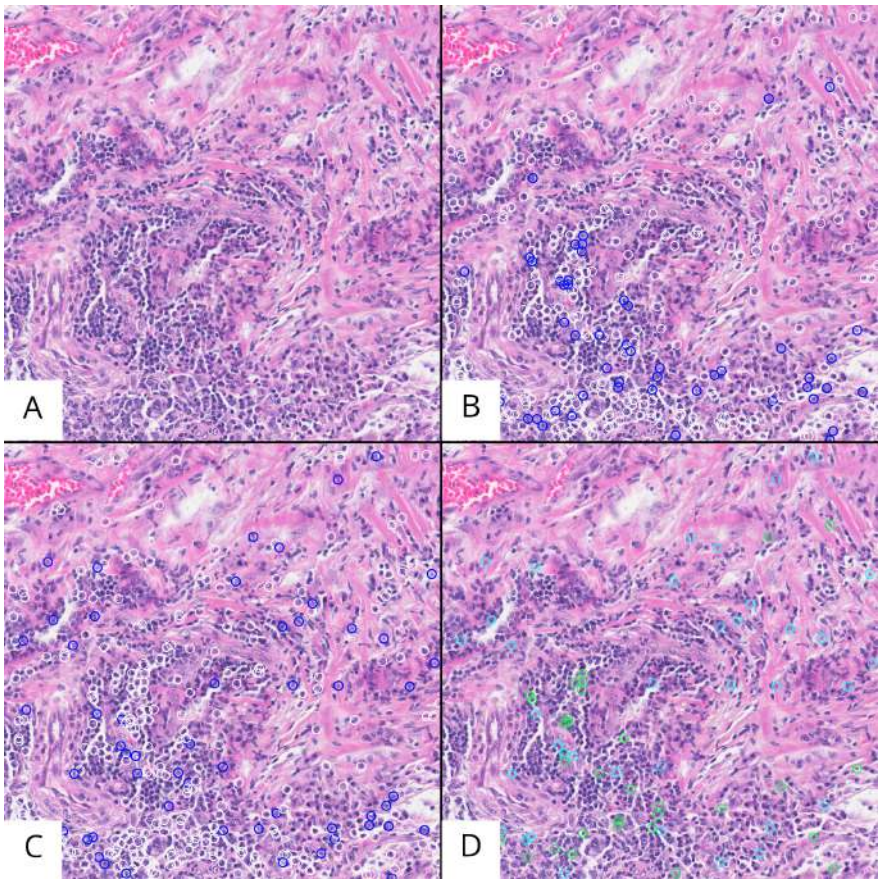

Figure 5.18: Example of comparison of the myeloid model with the test data set. Image A: The H&E area of interest. Image B: The ground truth obtained with the registration and the inForm® output. Blue circles are cells identified as myeloid cells. White circles are cells identified as something else. Image C: The prediction of the myeloid model on the tissue. Image D: The difference between truth and prediction. The blue circles are false positives, and the green circles are false negatives.

The second part of the evaluation is done on the DeepMEL dataset. The absence of ground truth does not allow us to have metrics. However, a pathologist gave his feedback on two images.

Regarding the myeloid cell model, the pathologist mentions the difficulty of establishing the presence of myeloid cells in both images. Detection of myeloid cells on slide is a very difficult task, even with immunohistochemistry. From what can be seen, some of the cells detected are consistent with the morphology of a macrophage, but at the same time, many of the objects detected seem quite random.

In addition to these analyses, another was performed to evaluate the improvement of adding the fourth channel designed to help the myeloid model recognize the correct cell

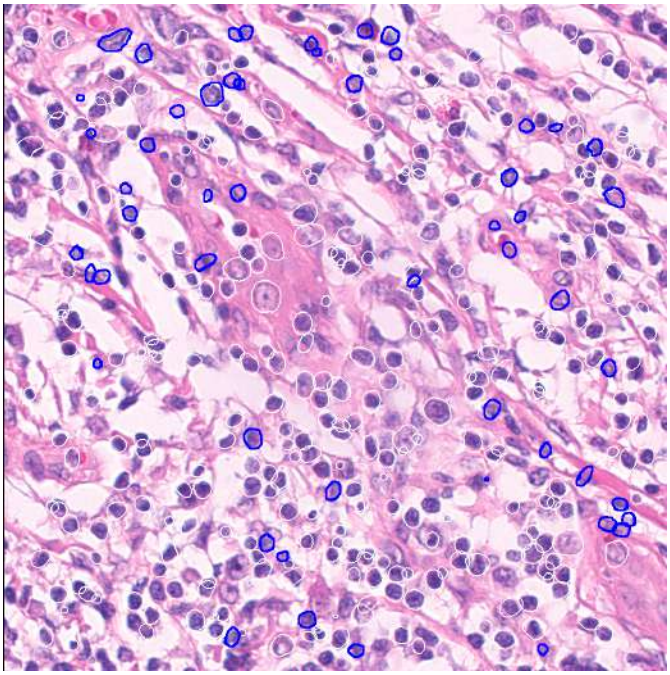

Figure 5.19: Example of the melanoma tumor cell model on tissue from the DeepMEL project. The blue circles are cells identified as melanoma tumor cells. The white circles are cells identified as something else.

to label.

As expected, without the fourth channel, the model is slightly worse [table 5.8]. The precision is the same as before, as myeloid cells make up 5% of the dataset, the accuracy barely moves. Precision and recall have gone down, however, by about 2-3%. This difference may seem significant, but since the number of myeloid cells is small, the value of precision and recall can change quickly.

Table 5.8: Metrics of the myeloid model with only three channels on the test dataset.

	Precision	Recall	F1-score	Support
Other	0.97	0.95	0.96	210644
Myeloid cell	0.42	0.56	0.49	12834
Accuracy			0.93	223478

The fourth channel seems to have some impact on the model as desired. However, it is not always essential. A difference between the model with and without the fourth channel was tested in the middle of the project when the dataset was 33% of its current size and showed more pronounced differences. Thus, the 4th channel seems useful when the dataset is

smaller. The larger the dataset, the less necessary it is. It could eventually harm the model because, with more data, the model could better learn the cell mask.

Sub-models

Sub-models were briefly tested, but the inaccuracy of our dataset made training these models nearly impossible.

Concerning lymphocytes, one task was known to be extremely difficult. A model attempting to distinguish T cells (CD3p) from B cells (CD20p) [1]. However, obtaining a large dataset as we did could hopefully help with this task. Nevertheless, since our dataset has many cells incorrectly identified as CD20p, the task is even more complex, and our model could not obtain a convincing result.

The myeloid sub-model that had to distinguish macrophages (CD68p) from dendritic cells (CD11Cp) was also tested. The result is very mixed, as the dataset is prone to many errors for myeloid cells, and the number of myeloid cells in our dataset is considered very low.

The trained myeloid sub-model shows a poor accuracy of 68% (table 5.9) on our test dataset. Indeed, if the model had only predicted that the cells were only macrophages, the precision would have been 81%. The recall for macrophages and dendritic cells is relatively good. The major problem lies in the precision for dendritic cells.

Table 5.9: Metrics of the myeloid sub-model on the test dataset.

	Precision	Recall	F1-score	Support
Dentritic cell	0.25	0.62	0.36	1906
Macrophage	0.91	0.69	0.78	10928
Accuracy			0.68	12834

The presumed reason for the behavior of this model is that the model correctly predicts true macrophages. However, everything else is considered dendritic cells, true dendritic cells, but also cells wrongly identified as myeloid cells (macrophages and dendritic cells).

Limitations and Future Work

The project has reached a point where the accuracy is above 93% for our three main models on the test dataset, at least 15% higher than the best accuracy obtained in the inherited project. However, these results can still be improved if we can clean up the dataset, especially errors due to inForm® mislabeling. The use of better software can be investigated, such as IFQuant or MCMICRO [10]. IFQuant is probably the software that will replace inForm® because we already have the result of IFQuant for one of our WSI and the CHUV wants to establish this software as the new software for this type of task in their new projects.

The second problem of the project is the generalizability of our models on external tissues. Improvements can be made by diversifying our dataset. Our dataset currently consists mainly of six cell types plus a few other cells that are grouped together as "others", but we do not know exactly which ones. Better control of these "other" cells by ensuring that the dataset is composed of a sufficient number of cells of each cell type that we may encounter in other tissues could greatly improve the model in real-world situations. Other tissues are also different from those we use for training in terms of scanning and staining. A more diverse tissue dataset could help the model be more robust on new tissues. A second possibility would be the addition of a low magnification image of the cell in the input as mentioned in the introduction. Finally, to verify the improvement of the models, a dataset with external tissues and truth should be built up because, for the moment, only lymphocytes can be verified with the CoNSeP dataset.

Discussion and Conclusion

In the course of this work, we developed a robust and accurate system. This system allows the registration of WSIs at the nucleus level with excellent accuracy. In this project, the success of the system allowed us to create large datasets consisting of H&E cells with labels obtained from mIF WSIs. With nearly one million cells in total, the datasets created were used to train different models to predict cell types. The three main models we trained achieved outstanding accuracy with a minimum of 93%. Two sub-models were trained without much success. The myeloid sub-model and the lymphocyte sub-models were trained with partially incorrect datasets, making them impossible to train correctly. Cleaning the database will surely allow a significant improvement in accuracy.

The addition of the input nucleus mask was not a triumphant success. It showed an improvement in the initial phase of the system when we had fewer data. However, with the further progress of the dataset, the need for this mask should potentially be questioned.

This work should have great potential to be useful for clinical research. Even if improvements can still be made, the project's current state can already be beneficial for researchers to obtain accurate labels without necessarily needing the help of a pathologist. The workflow is fast and accurate, which could lead to a new era where labeled datasets would no longer be a major problem in digital pathology.

Acknowledgements

I am deeply indebted to Dr. Andrew Janowczyk for his supervision during this project, for his invaluable support and feedback. I would also like to thank Dr. Pierre Moulin, the pathologist who helped me with this project by giving his evaluation of our models. The team of the CHUV provided a precious feedback at the end of each week. I would like to extended my sincere thanks to Céliane De luca for her editorial advice. Finally, I am grateful to Prof. Pascal Frossard for allowing me to do my master thesis in his laboratory.

Bibliography

[1] R Luz Elena Cano and H. Damaris E. Lopera. "Introduction to T and B lymphocytes". In: (July 2013). URL: https://www.ncbi.nlm.nih.gov/books/NBK459471/.

[2] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: (). URL: http://www.iro.umontreal..

[3] Simon Graham et al. "HoVer-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-Tissue Histology Images". In: *Medical Image Analysis* 58 (Dec. 2018). ISSN: 13618423. DOI: 10.48550/arxiv.1812.06499. URL: https://arxiv.org/abs/1812.06499v5.

[4] Gao Huang et al. "Densely Connected Convolutional Networks". In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-January (Aug. 2016), pp. 2261–2269. DOI: 10.48550/arxiv.1608.06993. URL: https://arxiv.org/abs/1608.06993v5.

[5] Anne S. Kramer et al. "InForm software: a semi-automated research tool to identify presumptive human hepatic progenitor cells, and other histological features of pathological significance". In: *Scientific Reports 2018 8:1* 8 (1 Feb. 2018), pp. 1–10. ISSN: 2045-2322. DOI: 10.1038/s41598-018-21757-4. URL: https://www.nature.com/articles/s41598-018-21757-4.

[6] Varun Kumar et al. "Data Augmentation using Pre-trained Transformer Models". In: (Mar. 2020). DOI: 10.48550/arxiv.2003.02245. URL: https://arxiv.org/abs/2003.02245v2.

[7] Giuseppe Lippolis et al. "Automatic registration of multi-modal microscopy images for integrative analysis of prostate tissue sections". In: *BMC Cancer* 13 (1 Sept. 2013), pp. 1–11. ISSN: 14712407. DOI: 10.1186/1471-2407-13-408/TABLES/4. URL: https://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-13-408.

[8] Xavier Moles Lopez et al. "Registration of whole immunohistochemical slide images: An efficient way to characterize biomarker colocalization". In: *Journal of the American Medical Informatics Association* 22 (1 Aug. 2014), pp. 86–99. ISSN: 1527974X. DOI: 10.1136/AMIAJNL-2014-002710/-/DC1. URL:

/pmc/articles/PMC4433366/%20/pmc/articles/PMC4433366/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4433366/.

[9] Kasper Marstal et al. "SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (Dec. 2016), pp. 574–582. ISSN: 21607516. DOI: 10.1109/CVPRW.2016.78.

[10] Denis Schapiro et al. "MCMICRO: A scalable, modular image-processing pipeline for multiplexed tissue imaging". In: *bioRxiv* (Mar. 2021), p. 2021.03.15.435473. DOI: 10.1101/2021.03.15.435473. URL: https://www.biorxiv.org/content/10.1101/2021.03.15.435473v1%20https://www.biorxiv.org/content/10.1101/2021.03.15.435473v1.abstract.

[11] Korsuk Sirinukunwattana et al. "Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images". In: *IEEE transactions on medical imaging* 35 (5 May 2016), pp. 1196–1206. ISSN: 1558-254X. DOI: 10.1109/TMI.2016.2525803. URL: https://pubmed.ncbi.nlm.nih.gov/26863654/.

[12] Andrew Su et al. "A Deep Learning Model for Molecular Label Transfer that Enables Cancer Cell Identification from Histopathology Images". In: *bioRxiv* (Mar. 2021), p. 2021.03.18.436004. DOI: 10.1101/2021.03.18.436004. URL: https://www.biorxiv.org/content/10.1101/2021.03.18.436004v1%20https://www.biorxiv.org/content/10.1101/2021.03.18.436004v1.abstract.

[13] Philippe Thévenaz and Michael Unser. "Optimization of mutual information for multiresolution image registration". In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 9 (12 Dec. 2000), pp. 2083–2099. ISSN: 1057-7149. DOI: 10.1109/83.887976. URL: https://pubmed.ncbi.nlm.nih.gov/18262946/.

Appendix

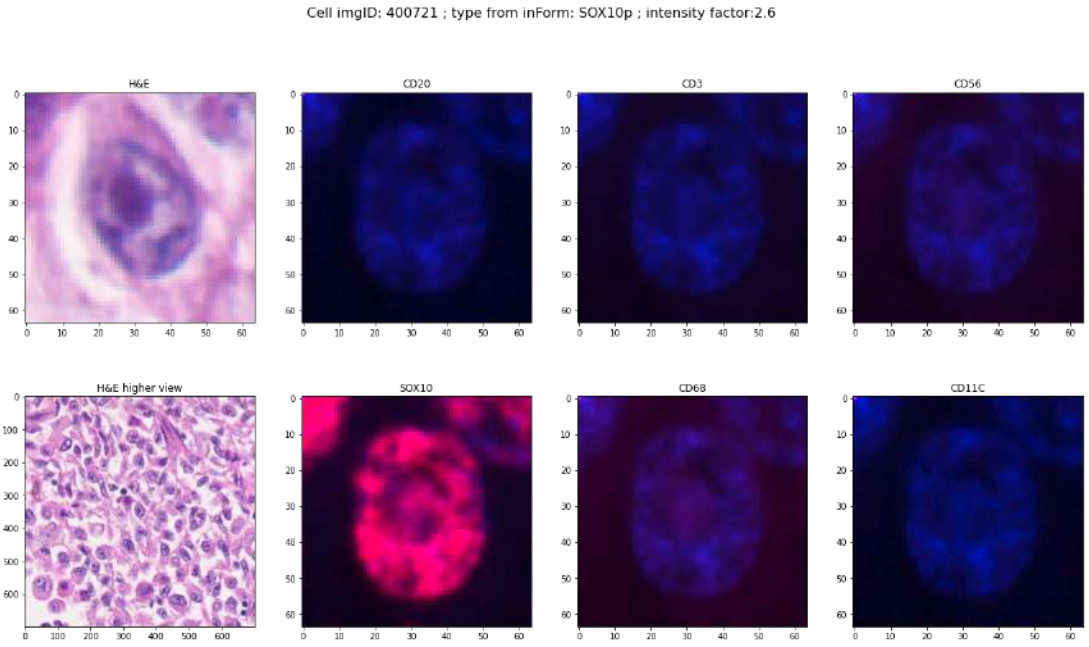

Figure .1: Example of a melanoma tumor cell in H&E and his corresponding mIF images. The six images on the right are mIF images. The blue color representes the DAPI channel and the red the channel of the protein.

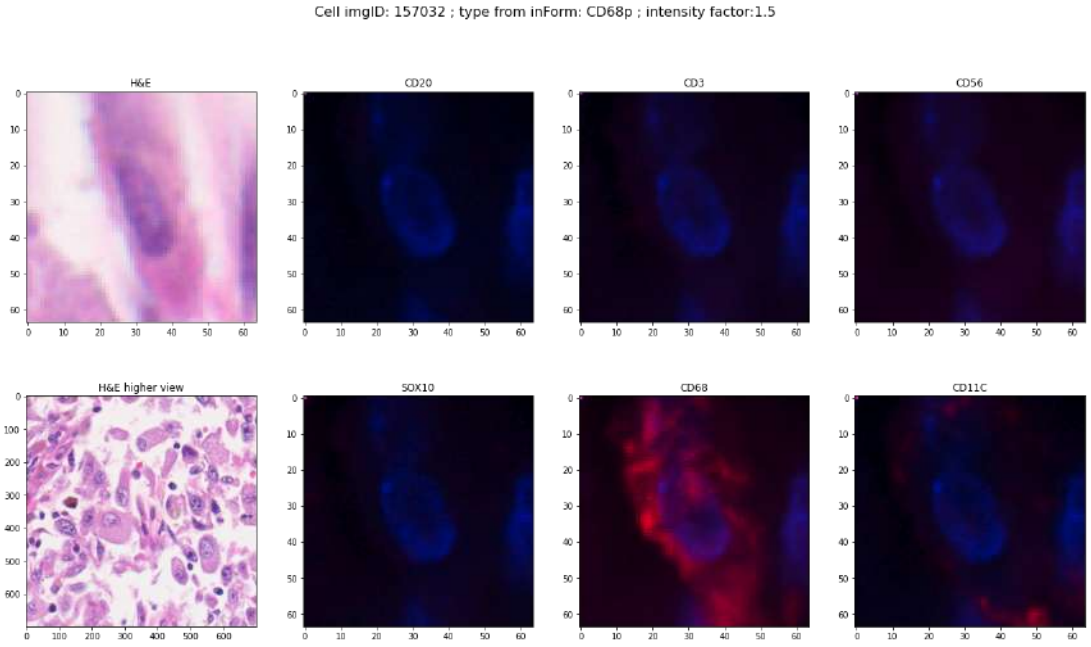

Figure .2: Example of a myeloid in H&E and his corresponding mIF images. The six images on the right are mIF images. The blue color representes the DAPI channel and the red the channel of the protein.