Labelling cell types with single cell spatial data

Labelling cell types with spatial single cell data

Jan Super (i6205862), Philip Mühlenfeld (i6155813), Gianmarco Parise (i6326141), Arthur Goffinet (i6215214), Angeliki Vogiatzoglou (i6203869)

Supervisor: Dr. Rachel Cavill
Coordinator: Dr. Linda Rieswijk

Abstract. Colon cancer is a very deadly form of cancer and its treatment requires further research and aid. There is high importance in inspecting tissue samples from colons with cancer to understand the evolution of the disease and find a cure. To help identify the different cell types in a tissue sample this research proposes a couple of applications that use unsupervised clustering to label the cells. It is found that the clustering methods provide a good way of labeling large groups of cells and can help speed up individual labeling and assessing the state of a patient's colon.

Maastricht University
Faculty of Science and Engineering
Department of Advanced Computing Sciences
Maastricht, The Netherlands

Contents

1 Introduction

Colon cancer is one of the most deadly forms of cancer that exist in our modern society. It occurs when the cells in the colon are growing in an uncontrollable manner. Despite progress being made in recent years int the treatment of this disease such as chemotherapy and development of new surgical techniques. The scientific community still look to understand the underlying of this disease. In order to find ways to improve the treatments.

In order to gain a deeper understanding of the underlying mechanisms of colon cancer, biologists are utilizing the computational power available today to extract important information from biological images through the use of automatic segmentation techniques. By using sophisticated algorithms to label cell types and subtypes through specific markers and spatial information, biologists can gain insights into new biological mechanisms and predict therapeutic outcomes that would otherwise be difficult to detect. Ideally, algorithms that can label these types of cells with absolute precision would make it faster for biologists to study them.

This research builds on the work of Dr. Rachel Cavill and her Ph.D studies, with a focus on identifying the most effective approach for accurately labeling cells and subtypes through the use of specific markers. The aim is to uncover new relationships between the data and propose new techniques that may exploit the spatial information.

The main problem for this research to solve is that there is no gold standard in automated single cell labeling, fully solved and labeled cell images are also not available. This means that a simple supervised learning approach is not viable, as there is no solution to assess the performance of the model while learning. There are no fully labeled data sets, as creating solved solutions to cell images would take a lot of time as thousands of cells would have to be individually labeled. Another issue is the reliability of the markers used, earlier research was not able to derive the labels for all cells from the given data. It is speculated that the reason for this is markers leaking into other cells and markers being unreliable. This previous research was conducted without the use of spatial information about the cells, due to the suspected leaking of markers is it hypothesised that spatial data about the relative positions of the cells should help alleviate some of the noise caused by leaking problems.

One way to accomplish finding relationships in unclassified data is by using a SOM map. SOM is a form of unsupervised learning that uses techniques like neurons and lowering the dimensionality of data to create clusters. These clusters can than be assessed individually to label them. Due to the grand dimensionality of the data and the fact that there is no solution yet to single cell labeling, SOM map have potential to be a powerful tool in labeling the cells. To assess the applicability of SOM maps on this problem the following research questions need to be answered:

1. What type of normalisation should be used for data used by the SOM map?

2. What is the best amount of clusters to be used by the SOM map for this data?

3. Can the clusters resulting from the SOM map be labeled as individual cells?

4. Does spatial information influence the clusters resulting from the SOM map?

In order to benchmark the results received by the SOM, this paper also discusses another unsupervised learning approach, which utilizes a deep autoencoder for dimensionality reduction followed by a K-Means clustering algorithm. In the remainder of this paper, this method will be referred to as AE-KM. Given that the goal is to compare AE-KM with the SOM, the following research question will be discussed:

5. How does the SOM's performance compare to the AE-KM algorithm?

In addition, it was decided to explore another type of clustering technique based on the BIRCH algorithm. This was an opportunity to explore the presence of different cell types in each tissue type (Inflamed, Uninflamed and dysplastic) for seven different patients.

6. Is it better to do the clustering for all the data or do it individually for each group?

2 Related work

The study by Li and Feng 2022 introduced a method called Neural Network-based Cell Annotation (NeuCA) for the supervised labeling of single cell RNA-sequencing data. This method uses already labeled data as the training set to train a classifier to label new cell data. NeuCA first calculates the correlation between cell types, and depending on whether the correlation is high or low, it follows a different procedure for labeling. In cases of highly correlated cell types, the method uses a tree structure created by hierarchical clustering for training the neural networks. In cases of low correlated cell types, NeuCA trains a feed-forward neural network for predicting labels. This approach improves the accuracy of labelling and eliminates the high rate of unlabelled data.

scConsensus developed an algorithm that combines both supervised and unsupervised learning to cluster and label different cells. The algorithm is originally created and tested on RNA sequencing data. It works by taking the clustering results from both supervised and unsupervised learning and creating a contingency table. The data from this table is used to form sub-clusters and choose data points that will improve principal component analysis clustering. The algorithm is provided as an R package that is ready to use for free for academic purposes.

The Self-Organization Map (SOM) by Vununu, Lee, and Kwon 2020 is an unsupervised machine learning technique that creates a low-dimensional representation of higher-dimensional data by grouping it into different clusters. SOM is a type of artificial neural network that uses a competitive approach to train the network instead of an error-based approach. This approach allows for unsupervised learning on unlabelled data.

3 Dataset

The dataset provided for this research contains 394132 rows and 68 columns with labels, where each row represents as cell. The majority of this data consists of the amount of certain marker found in a specific cell. The cells themselves are labelled after the image they originate from, which patient that images originates from and contain information about the stage of cancer the patient was in while this cell was measured. The data was obtained by taking a tissue sample from patients at a certain stage of their colon cancer. Markers were added to this tissue, which would then bind to specific proteins. These markers are then measured and recorded.

4 Methods

4.1 SOM map

A self organizing map (SOM) Du 2010; Vununu, Lee, and Kwon 2020 is an unsupervised machine learning method which takes high dimensional data and reduces it into low-dimensional data of order two, by grouping them into similar clusters. Unlike traditional ANN approaches that are error based, SOM maps are using a competitive approach that allow unsupervised learning. This type of learning is needed in order to label unknown data without having access to already correctly classified data. The SOM algorithm works by creating a network of neurons where each represent a different cluster. The SOM algorithm will calculate the distance between each input of the high dimensional data and each neurons to update the weight according to the neighbors equation.

$$W(t+1) = W(t) + \theta(u, v, t)[d(x) - W(t)] \tag{1}$$

One of the key advantages of the SOM map is that it can reveal hidden patterns or structures, in addition there are several that represent the output graphically thanks to the dimensional reduction SOM map. SOM maps have already proven to be effective on biologic data from patients with cancer Hautaniemi et al. 2003.

To label the clusters formed by the map the mean amount of a certain marker is calculated, if it differs more than 0.75 times the standard deviation from the mean of that marker over the entire dataset it can be concluded that the cluster is at least partially based on that cluster. The types of markers that are that prevalent in a cluster are recorded for each cluster. Using the classification tree found in the Appendix the clusters are labeled. It is possible for a cluster to contain multiple kinds of cells.

4.1.1 Normalisation

To find the optimal type of normalisation for the given data three different normalisation methods will be tested. As mentioned before, the data used in this research has gone through a proprietary pipeline. Meaning it has in some way been cleaned or structured, because of this it is possible to use this data without any normalisation. This will be the basis behind the first two experiments, where one will be without spatial data and one with. The spatial data gets this treatment because of its importance to the paper overall. By excluding and including it in different experiments the importance of the spatial information can be derived from results.

As a second option, min-max normalisation will be used. Min-max normalisation is a simple yet effective type of normalisation that scales all the data between 0 and 1 while preserving the original relationships between the original data. This is done with the following formula.

$$X_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

A downside of this is that the standard deviation also becomes smaller due to this, which might make outliers less influential. As before the data will be normalised both with and without spatial data, however for the experiments with normalisation there will be an extra factor. More experiments will be carried out to observe the effect of normalising only a single patient's data or normalising all the data at once. This is done as here could be differences between for example the amount of marker used for each patient or different equipment, which could give different results. By normalising all the data at once there could also be too high of a loss in standard deviation and therefore noise and outliers, which is especially important when it comes to finding markers in cells.

The third option tested will be the standard scaler found in the *sklearn* library. This method works by removing the mean from the data and scaling to the unit variance. This is done with the following formula.

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

This is one to all columns individually to make them more normally distributed. The same amount of different tests will be performed as with the prior normalisation method for the same reasons.

For the comparison of these three types of normalisation a single image from patient 1 will be used. This image is used for the sake of consistency and because it contains a high amount of cells to be labeled. A simple $4x4$ SOM map will be trained on all the data from only patient 1, again for consistency's sake.

4.1.2 Amount of clusters

The experiments that are going to be conducted, to determine the best amount of clusters to use, will be using data normalised using the previously found most optimal method. To find the right amount of clusters outcomes from the different experiments will be recorded. The amount of unlabeled, multi-labeled and epithelium will be recorded. Unlabeled and multi-labeled are used as it can show how specific and sensitive the method is to single cells in groups with a lot of other cells. Epithelium is used as it has proven to be the most prevalent and consistently correctly classified. A predefined range of clusters between $XandY$ with step size Z will be tested. For the sake of consistency again the SOM map will be trained on exclusively data from patient 1. The recorded performance of the cluster sizes will be portrayed in a plot for comparison.

4.1.3 Individual cells

To asses the performance of the SOM map in individually labelling cells an image from all stages from patient 1 will be used for consistency. The amount of clusters will be assessed, together with the amount of unlabeled and multi-labeled cells. This will be done for multiple sizes of the SOM map as it is not expected for a single amount of clusters to perform a lot better than the rest. From these images there will also be a visual assessment, in which the evolution of the clusters is discussed. As it is impossible for a specific cluster of cells to be assessed for all stages the evolution of clusters is assessed over multiple different images. This assessment will also explain the usefulness of the clustering and labeling for biological research purposes.

4.2 AE-KM

Neural networks, such as deep autoencoders, have long been a prominent technique for dimensionality reduction (Hinton and Salakhutdinov 2006). Their advantage over, for instance, principal component analysis (PCA), is

that they can create a non-linear feature transformation. A deep autoencoder achieves this by being trained to reconstruct its input. It consists of two main parts: an encoder and a decoder. The encoder maps the input data to a lower-dimensional representation, called the bottleneck or latent representation. The decoder then maps the bottleneck representation back to the original input space. The training process involves minimizing the difference between the original input and the reconstructed output using backpropagation. In biology, autoencoders are often used in conjunction with clustering algorithms, to account for the high dimensional nature that is usually present in biological datasets (Hadipour et al. 2022; Karim et al. 2020). In this paper, we use the K-Means clustering algorithm on top of the autoencoder. K-Means is used over other, more powerful, clustering algorithms due to its computational efficiency and the option to manually specify the number of clusters.

4.3 Genetic Algorithm Optimization

An initial approach was the optimization of the existing genetic algorithm in Matlab. However due to the fact that this is already an optimization algorithm this was not possible. Thus, the approach was redirected and a clustering technique was implemented instead based on the BIRCH algorithm.

4.4 BIRCH clustering

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchy) is a data clustering algorithm that is designed to handle large datasets with high dimensionality. BIRCH performs well especially for datasets with a high amount of data points but significantly low amount of clusters.

The BIRCH algorithm creates a tree structure called a CF (Clustering Feature) tree to represent the data. The tree is built iteratively, with each iteration adding a new data point to the tree and rebalancing the tree by merging or splitting clusters as needed. A CF tree is a hierarchical tree structure, and each tree node represents a set of data points.

Large datasets are processed by the BIRCH algorithm using two key concepts:

1. CF-Tree: This is a hierarchical data structure that represents data sets. It stores s summary information about each cluster and the data points that belong to that cluster.
2. Cluster features: This is a vector that summarizes cluster properties It includes information such as the number of data points in a cluster, the cluster mean and variance etc.

Workflow of the BIRCH algorithm:

1. Initialization of the CF-tree with a single leaf node
2. Reading the data points individually and inserting them into the CF-tree
3. Splitting a mode into two smaller ones if it becomes too large
4. Merging two separate nodes into one bigger f these are similar
5. Repeating steps 2-4 up to the point where all data have been processed

Once the CF tree is established, the algorithm can use the tree to efficiently perform operations such as inserting new data points, retrieving the nearest neighbors of a data point, and finding clusters in the data.

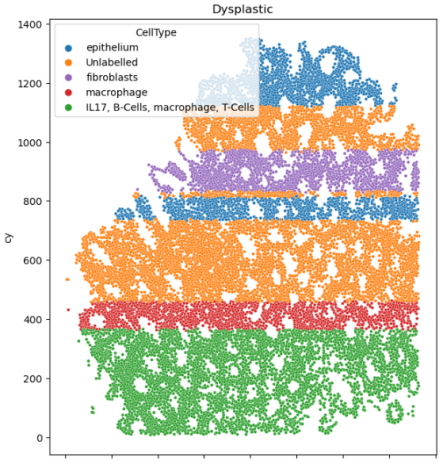
BIRCH has several advantages over other clustering algorithms, such as the ability to handle large data sets, the ability to handle high dimensionality, the ability to adapt to change of data distributions, and the ability t produce high-quality clusters. Based on these advantages, specifically the ability to handle large data sets and high dimensionality, as well as the fact that the algorithm follows a tree - like structure, as the one used as a reference for the different cell types in this project, BIRCH algorithm was chosen for the clustering implementation Zhang, Ramakrishnan, and Livny 1997.

5 Experiments and Results

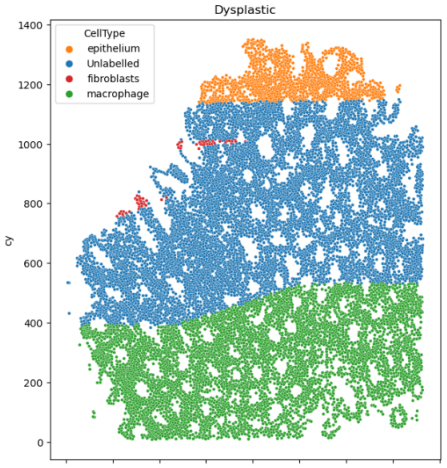
5.1 SOM map

5.1.1 Normalisation and Spatial information

After applying no normalisation on the data, using all data from patient 1 excluding spatial data to train a 4x4 SOM map and then labeling a random dysplastic image the following clusters were obtained



(a) No spatial info.

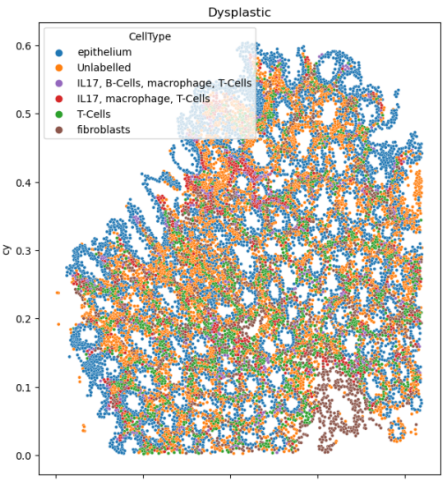


(b) With spatial info.

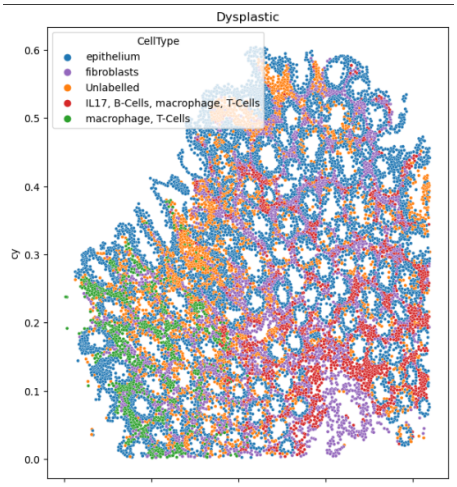
Figure 1: No Normalisation

As can be seen in the image there appear to be layers with different kinds of cells. It is very clear that this is not correct and that there are values with disproportionately more influence on the clusters than they should have. Something similar occurs when adding spatial data and running the same experiment.

Secondly normalisation between 0 and 1 is applied by dividing the data minus the minimum value by the maximum value of the data with the minimum. When normalising all the data at the same time and running the same test the following clusters were obtained.



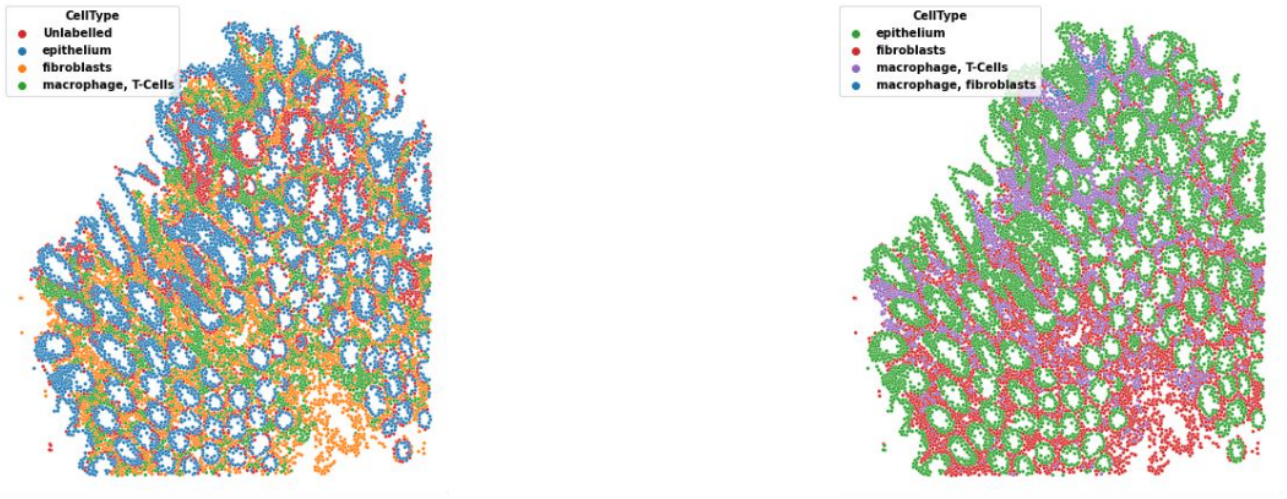
(a) No spatial info.



(b) With spatial info.

Figure 2: Min-Max Normalisation on all data

In this image a lot of cells remain unlabeled but the epithelium clusters seem to always form around the edges. Addiding spatial information seems to change the results. The results from normalising on a single patient's data seem to differ.

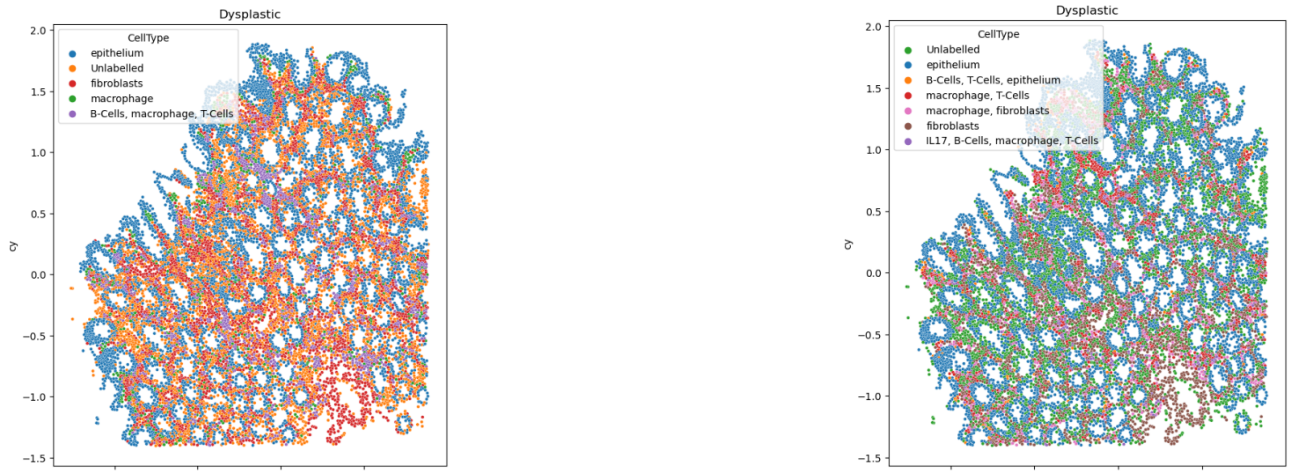


(a) No spatial info. (b) With spatial info.

Figure 3: Min-Max Normalisation on single patient data

It appears that the map generalises less aggressively without the spatial data but more aggressively with spatial data.

Lastly normalisation is applied by removing the mean and scaling to unit variance. When normalising on all the data at the same time and running the same test as before the following clusters were obtained.

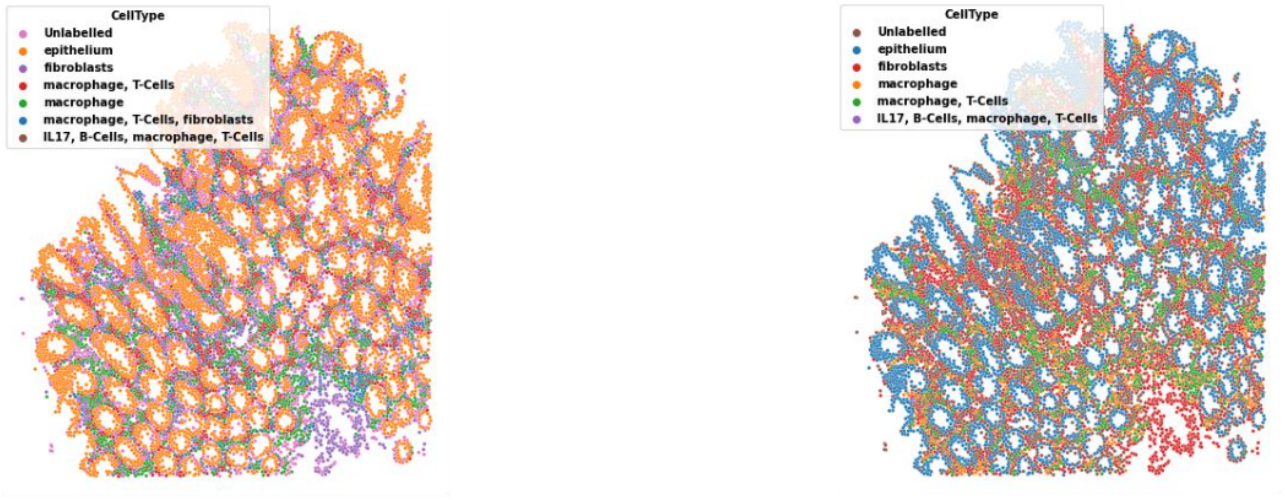


(a) No spatial info. (b) With spatial info.

Figure 4: Standard scaler on all data

It appears that even when scaling all data at the same time it generalises a lot less aggressively than its min-max normalisation counterpart.

This normalisation is also applied on only the data of patient 1. After performing the same experiment the following clusters were obtained.



(a) No spatial info.

(b) With spatial info.

Figure 5: Standard scaler on single patient data

From the results it appears that standard scaling with spatial information, standardised on single patient data provides results with a low amount of unlabeled cells while not seeming to overgeneralise. Therefore this configuration will be used to further assess the SOM map, however further conclusion about the SOM map will include the results of the other experiments.

5.1.2 Amount of clusters

The graph below depicts the recorded percentages of the given cell types found in the experiment testing a broad range of cluster sizes. As said before, the configuration used for this experiment is: standard scaling on the data of onlyn patient 1 including spatial information.

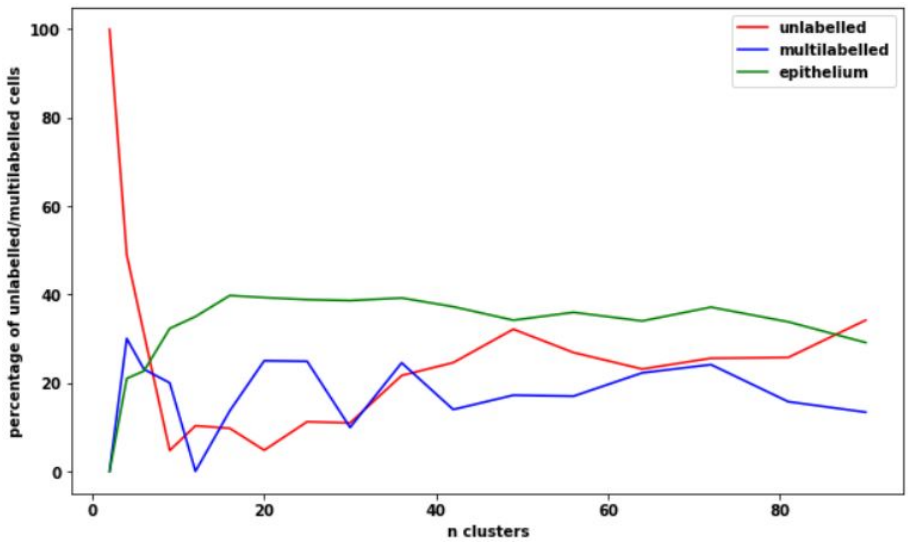


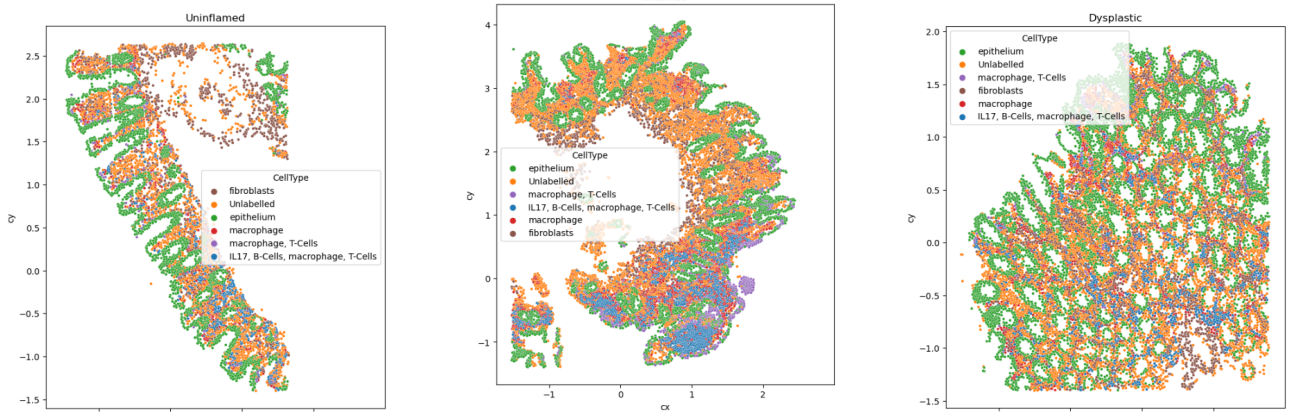
Figure 6: Graph of percentage of unlabeled, multi-labeled and epithelium cells

Generally both the percentage of epithelium and multi-labeled cells seems to stay consistent, where the amount of unlabeled cells dips down first and then stabilises. This can be attributed to the fact that the maps

with a low amount of clusters generalise more aggressively, resulting in a lower amount of unlabeled cells. With this in mind the 6x6 map size seems the most optimal size as after that point the percentages of unlabeled cells stay somewhat consistent.

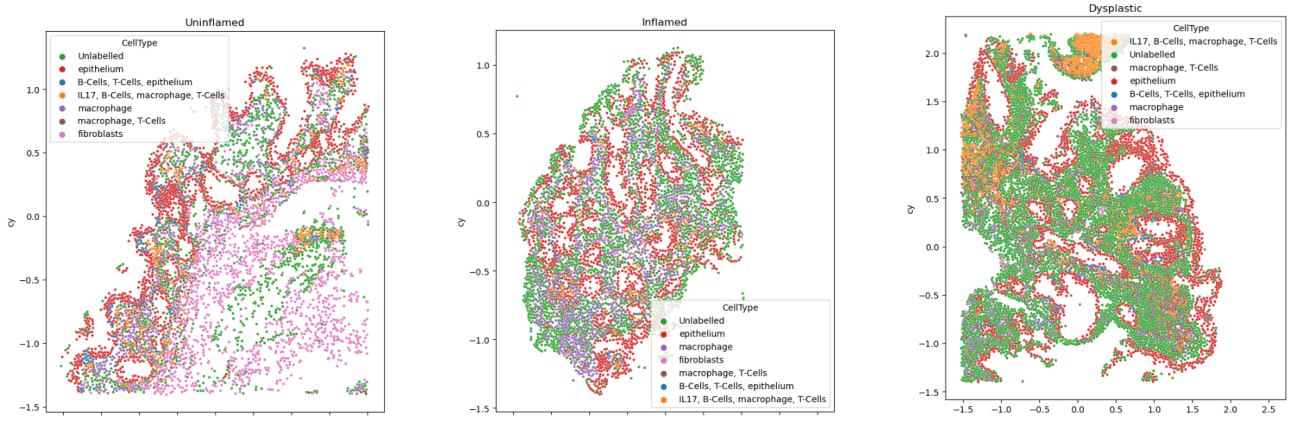
5.1.3 Cluster labeling

As details previously each cluster are assign a cell type based on the distribution of specific proteins marker compare to the rest of the data set. by consequence the result of the labelling will depend of on which subset of the data we will used.



(a) Uninflamed (b) Uninflamed (c) Dysplastic

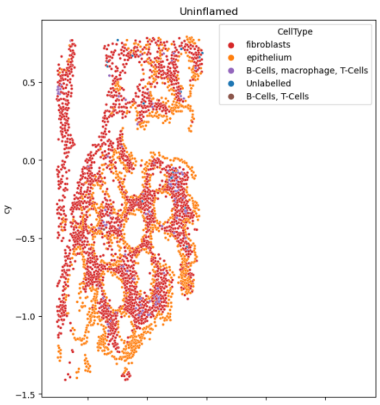
Figure 7: Patient 1



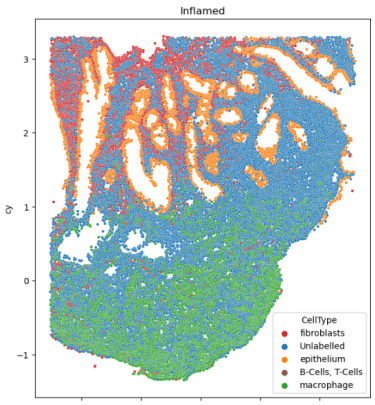
(a) Uninflamed (b) Uninflamed (c) Dysplastic

Figure 8: Patient 3

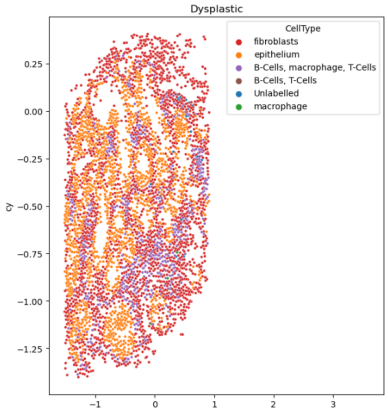
The results will be discussed in the conclusion.



(a) Uninflamed

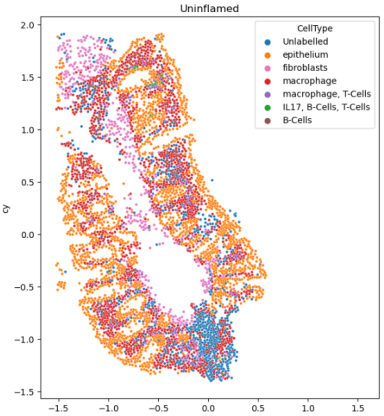


(b) Uninflamed

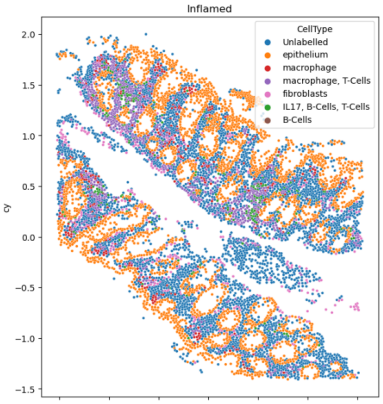


(c) Dysplastic

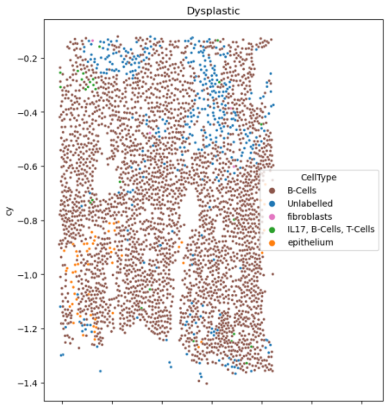
Figure 9: Patient 4



(a) Uninflamed



(b) Uninflamed



(c) Dysplastic

Figure 10: Patient 5

5.2 AE-KM

In order to compare the SOM with the AE-KM method, the experiments on the AE-KM were performed with the same processes that were already applied to the SOM. Specifically, in both cases, the standard scalar normalization was applied over the entire dataset, the spatial information was included and the number of clusters that were found by SOM (i.e. 16) was manually matched by the AE-KM method. Additionally, no concrete hyperparameter tuning was performed on the deep autoencoder, with respect to its initial weights and number of neurons. Instead, inspired by Maaten (2009), the network dimensions were set to 54 - 150 - 150 - 500 - 5, where 54 is the dimension of the dataset and 5 is the latent space representation.

(a) SOM

(b) AE-KM

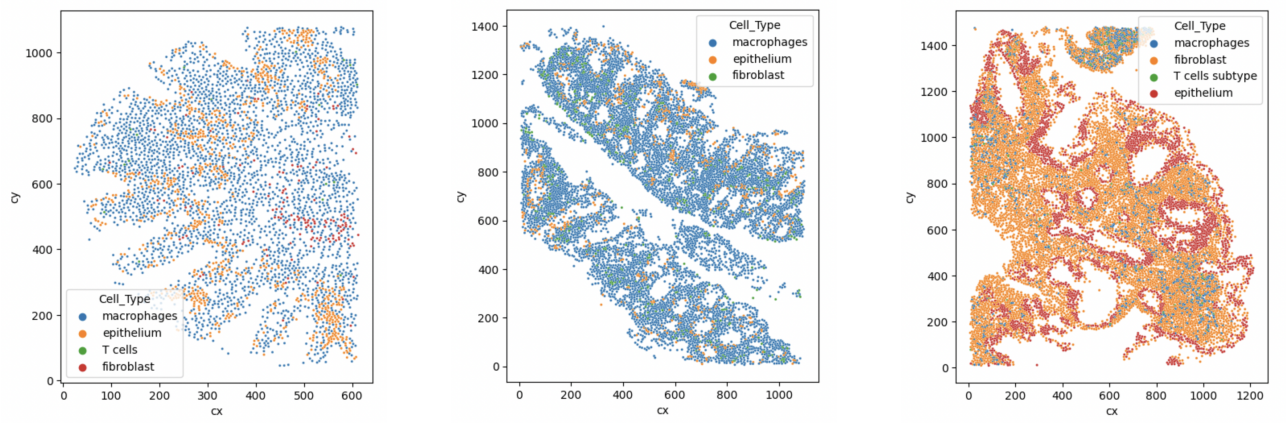
Figure 11: Comparison of SOM and AE-KM methods on dysplastic tissue sample

When comparing the results between the SOM and the AE-KM method, the main similarities lie in the amount and approximate location of the epithelium cells, while the differences are that the AE-KM did not classify any cells as multiple cell types and overall also has a much smaller percentage of unlabelled cells. Instead, the unlabelled cells found with the SOM approach are mostly classified as macrophage cells in the AE-KM method.

5.3 BIRCH Clustering

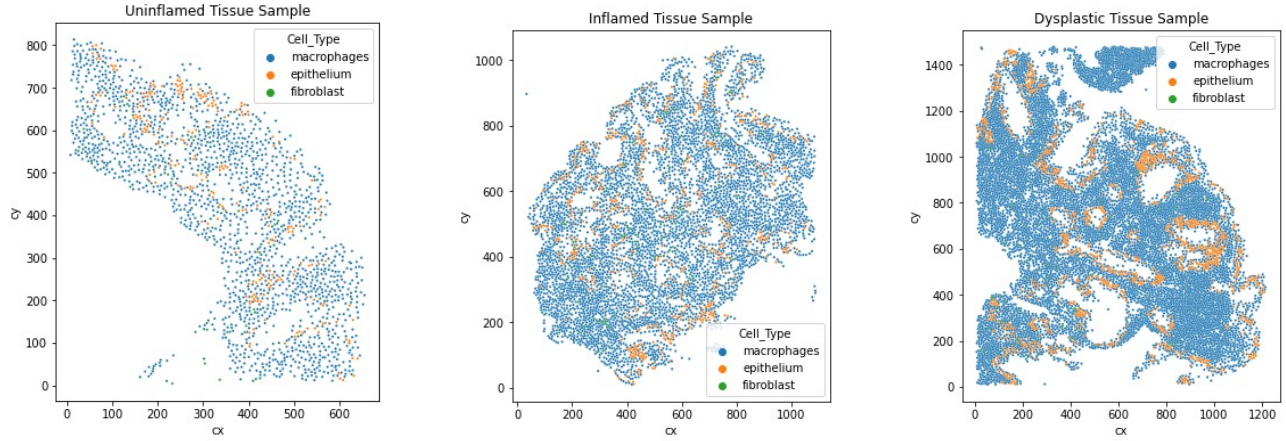
As mentioned before, the BIRCH clustering algorithm was implemented for the clustering of the cell types. The overall dataset (394,132 cases) was split into three subsets: dysplastic (98,303 cases), inflamed (115,405 cases), and uninflamed (180,424 cases) tissue. The analysis was conducted separately for each tissue type. More precisely, four different approaches were implemented; for the whole dataset including all patients (inflamed, uninflamed, dysplastic), for the uninflamed patients , for the inflamed patients and for the patients with dysplasia. The variables taken into account were the cells' area value and the markers described in the "tree markers" pdf file. Before fitting the model, the dataset was normalized by the normalizer() function provided by scikit-learn package. This function rescales the input data for each sample to have a unit norm, independently of the distribution of the samples. The L2 norm (Euclidean norm) was used. Scaling inputs to unit norms is a common operation for clustering. After fitting the model and predicting the cluster assignment two metrics were employed to assess the clustering goodness: the Davies Bouldin and silhouette scores. With the Davies Bouldin score, zero is the lowest possible score, so score values closer to zero indicate a better partition. With the silhouette score, the best value is 1 and the worst is -1, while values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar. An attempt was made to determine the cell types. The method relies on the iterative maximum mean values search of the markers in the individual cluster, according to the indications given in the abovementioned "tree markers" pdf file. Example figures of the three tissue types were generated by applying the models to show the cell types distribution and they are presented below.

When performed the clustering for the whole dataset, the results acquired for the three distinct tissue types (inflamed, uninflamed and dysplastic) are the following:



(a) Uninflamed (b) Inflamed (c) Dysplastic

Figure 12: Clustering performed based based on the tissue type



(a) Uninflamed (b) Inflamed (c) Dysplastic

Figure 13: Clustering performed based based on the whole dataset

Dataset	Davis Bouldin	Silhouette
Overall	1.59	0.320
Dysplastic	1.61	0.166
Inflamed	1.77	0.312
Uninflamed	1.47	0.349

6 Discussion and Conclusion

6.1 Conclusions from results

As was shown in the experiments section. the most optimal configuration for the SOM map is: the data of a single patient standardised with spatial data used to train a 6x6 SOM map. As this provides a very diverse group of clusters with a lot of different cells that does not generalise too much or too little, as is shown by its generally low amount of unlabeled cells. In general, due to the limitations discussed later it should be noted that it is not possible to draw a conclusion on which of the given options is the absolute best, as there is no way to benchmark the results. For the sake of simplicity and keeping the amount of experiments sustainable only a single configuration that seemed the most reliable from testing was used to compare the clusters. So to answer the first two research questions, standard scaling seems to provide the best kind of normalisation for the SOM

map as the resulting SOM maps produces the best results. The answer to the second question is that a 6x6 SOM map is the best size to use, as this was the least complex configuration that showed the same results in terms of unlabeled, multi-labeled and epithelium cells.

When analysing the clustering results from section 5.1.3 a couple of things can be seen, the epithelium cluster consistently forms around the edges of the structures in the images. This signifies how strongly these cells differ from others, which helps it serve as a benchmark for model performance. Generally speaking it seems that the images from the inflamed stage have the most unlabeled cells. Overall this is a consistent issue, images from the uninflamed and dysplastic stages contain less unlabeled cells and have stronger epithelium clusters. When comparing the first and last stage of the cancer it does show that the clusters containing B-cells seem to become more prevalent. This is consistent with the findings of Shimabukuro-Vornhagen et al. 2014 where it was concluded that "B-cells constitute a significant proportion of the immune infiltrate in colorectal cancer". This finding supports the notions that the SOM maps clusters are biologically sound and can be used for further research. The same goes for T-cells in some images, where research has also shown that T-cells play a role and become more prevalent in the human body fighting cancer (Koch et al. 2006). However, this trend is not noticed overall, this could be due to unbalanced data. To further analyse this occurrence graphs with comparisons between the percentage of cells made by using the most optimal configuration (6x6 SOM) will be provided in the appendix for some patients. It should be noted however that some clusters contain multiple cells, the so called multi-labeled clusters. As seen in the experiments the percentage of multi-labeled clusters stays consistent, meaning that it does not seem realistic for the SOM map to label all cells individually. So to answer the third research question, the SOM map can be used to individually label a large amount of cells, however there will remain clusters of cells that will need to be manually analysed. This amount is of course a lot less than the total amount of cells.

When analysing the results of 5.1.1, it is obvious that there are differences in the images that do or do not contain spatial information. It appears that the SOM map gives some value to the spatial information and actively uses it to determine the clusters. It should be noted however that the issue with normalising distance is that there is potential information loss where the absolute distance does have influence over the type of cell, relative distance might make this less prevalent. This can also explain the differences in images with and without spatial data, it is more likely for this issue to persist in clusters created by using min-max normalised data. As all data is bounded between 0 and 1 instead of being re-scaled like in standard scaling. Seeing as the configuration used to assess the performance of SOM maps overall is does include spatial information and this configuration performs rather well. The notion that spatial data is not useful cannot be proven with the data found. So to answer the research question, spatial information does influence the clusters resulting from the SOM map and seems to have a positive influence.

When comparing the SOM to the AE-KM method, we can mainly conclude that both methods classify a large number of cells as epithelium cell types. Moreover, the classified epithelium cell types can be found in similar locations for both algorithms. In order to get further insight into the quality of either method, an expert in the field of biology should evaluate the resulting clusters from both approaches.

Overall it can be concluded that SOM maps provide a fast and effective way of labeling and clustering large amount of cells from tissue images. It should not be expected to perfectly label every individual cell and replace human researchers for the time being. It should be used as a tool to help those labeling every cell individually to know what to expect in a certain region. It can also be used to look at general trends in the change of cells in tissue through the stages of cancer, this can help in the prognosis and further treatment of the cancer.

In addition, looking at BIRCH clustering observing the scores and in particular the silhouette score the performance of the model considering the whole database is comparable to the results of the models obtained with only inflamed stage or uninflamed stage cases. In contrast, for the model obtained with dysplastic stage cases the performance would seem to be lower. Comparing the figures of the dysplastic sample with the overall model, one can see that the same group of cells was labeled differently-this is most likely a problem. If we compare the inflamed and uninflamed model with the model that takes into account the whole dataset it will turn out that the labeling is almost the same.

In conclusion regarding the BIRCH algorithm we get similar results with two of the three types of stages available. Therefore, it is not possible to determine whether or not it is better to use the whole dataset or its subsets to fit the cluster.

6.2 Limitations

There are a couple of striking limitations with this research, making it so that the classification is not as specific as it could have been. First of all there is no biological background within the research team, meaning that there

is no possibility to draw academically grounded conclusions about the correctness of the clustering. In addition to this there was no further documentation provided other than the original data, the original matlab code, results from matlab code and the tree marker file. This means that viable metadata useful for understanding the data was missing. Because this information is missing there is potential for a loss in information in the normalisation or cleaning process as only the numeric data was used from the data originally provided at the start of the project. The aforementioned tree marker file used for classification is not easy to understand. It it not specific in its proportions but uses terms like 'a lot' and 'a little' without elaborating further. This results in the method used for labeling only relying on how much the mean of a marker within a cluster differs from the general mean over all the data, which is not directly grounded in academic truth. The original data was also provided as an R-workspace, meaning that the data had to be extracted and it cannot be verified that no data was lost in this process due to the size of the dataset.

With regard to the AE-KM method, the following limitations need to be discussed. First, no extensive hyperparameter tuning for the deep autoencoder has been performed. This is an important step, given that hyperparameters, such as the number of layers and weight initialization, can significantly influence the resulting latent space representation. Furthermore, while K-Means was used in this paper for its efficiency, the clustering algorithm also has multiple drawbacks, such as being sensitive to the initial partitioning, as well as only being able to cluster convex shapes.

6.3 Future Work

Future Work on this project can be broken down into three main categories. First, this project could benefit from the acquisition of additional resources. These could come in the form of additional documentation regarding the dataset or biological expert knowledge. Another area where future work could be performed is the improvement of the existing algorithms, particularly by focusing on the points that are mentioned in the Limitations section 6.2. Lastly, new algorithms can be implemented. One specific family of algorithms that could deliver strong results are deep clustering methods (i.e. deep learning-based clustering). The advantage of such algorithms over other methods is that the neural network and clustering algorithm are trained at the same time, rather than sequentially. In their study, Karim et al. (2020) benchmark different state-of-the-art deep clustering methods on several bioinformatics use cases. Thus, their paper could serve as guidance in finding the appropriate deep clustering method for the problem at hand.

References

[1] K.-L. Du. "Clustering: A neural network approach". In: *Neural Networks* 23.1 (2010), pp. 89–107. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2009.08.007. URL: https://www.sciencedirect.com/science/article/pii/S089360800900207X.

[2] Hamid Hadipour et al. "Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means". In: *BMC Bioinformatics* 23 (Apr. 2022). DOI: 10.1186/s12859-022-04667-1.

[3] Sampsa Hautaniemi et al. "Analysis and visualization of gene expression microarray data in human cancer using self-organizing maps". In: *Machine learning* 52.1 (2003), pp. 45–66.

[4] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786 (2006), pp. 504–507. DOI: 10.1126/science.1127647. eprint: https://www.science.org/doi/pdf/10.1126/science.1127647. URL: https://www.science.org/doi/abs/10.1126/science.1127647.

[5] Md Rezaul Karim et al. "Deep learning-based clustering approaches for bioinformatics". In: *Briefings in Bioinformatics* 22.1 (Feb. 2020), pp. 393–415. ISSN: 1477-4054. DOI: 10.1093/bib/bbz170. eprint: https://academic.oup.com/bib/article-pdf/22/1/393/35934885/bbz170.pdf. URL: https://doi.org/10.1093/bib/bbz170.

[6] Moritz Koch et al. "Tumor infiltrating T lymphocytes in colorectal cancer: Tumor-selective activation and cytotoxic activity in situ". en. In: *Ann. Surg.* 244.6 (Dec. 2006), 986–92, discussion 992–3.

[7] Ziyi Li and Hao Feng. "A neural network-based method for exhaustive cell label assignment using single cell RNA-seq data". In: *Scientific reports* 12.1 (2022), pp. 1–12.

[8] Laurens van der Maaten. "Learning a Parametric Embedding by Preserving Local Structure". In: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*. Ed. by David van Dyk and Max Welling. Vol. 5. Proceedings of Machine Learning Research. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR, 16–18 Apr 2009, pp. 384–391. URL: https://proceedings.mlr.press/v5/maaten09a.html.

[9] Alexander Shimabukuro-Vornhagen et al. "Characterization of tumor-associated B-cell subsets in patients with colorectal cancer". en. In: *Oncotarget* 5.13 (July 2014), pp. 4651–4664.

[10] Caleb Vununu, Suk-Hwan Lee, and Ki-Ryong Kwon. "A strictly unsupervised deep learning method for HEp-2 cell image classification". In: *Sensors* 20.9 (2020), p. 2717.

[11] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. "BIRCH: A new data clustering algorithm and its applications". In: *Data mining and knowledge discovery* 1.2 (1997), pp. 141–182.

Appendices

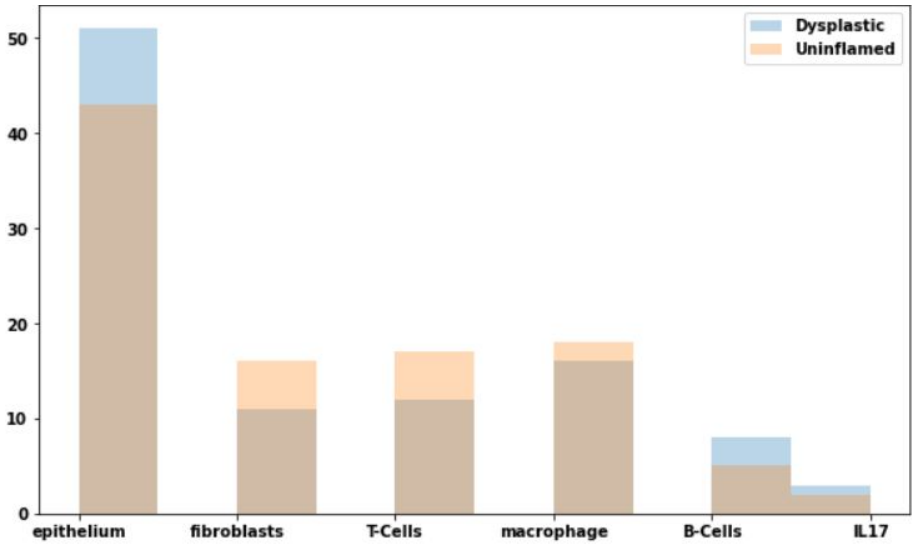


Figure 14: Evolution of cell types patient 1

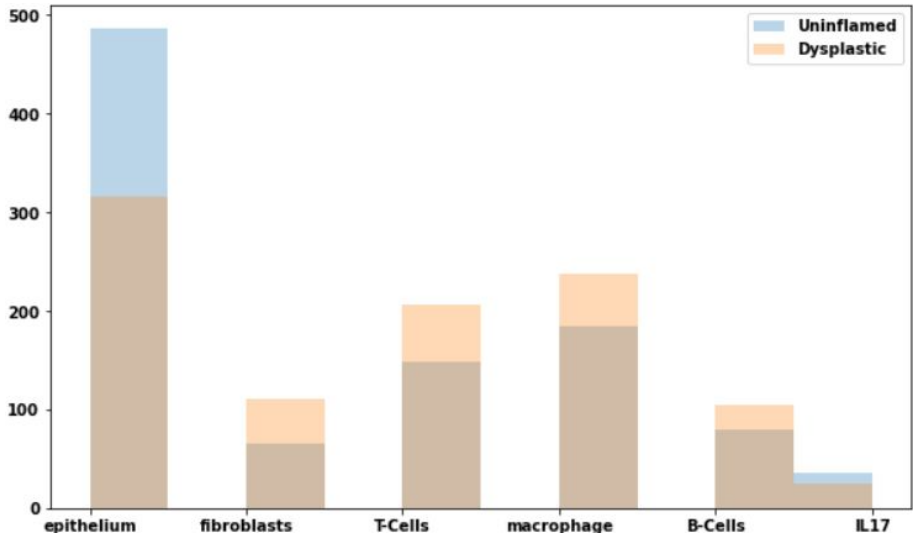


Figure 15: Evolution of cell types patient 3

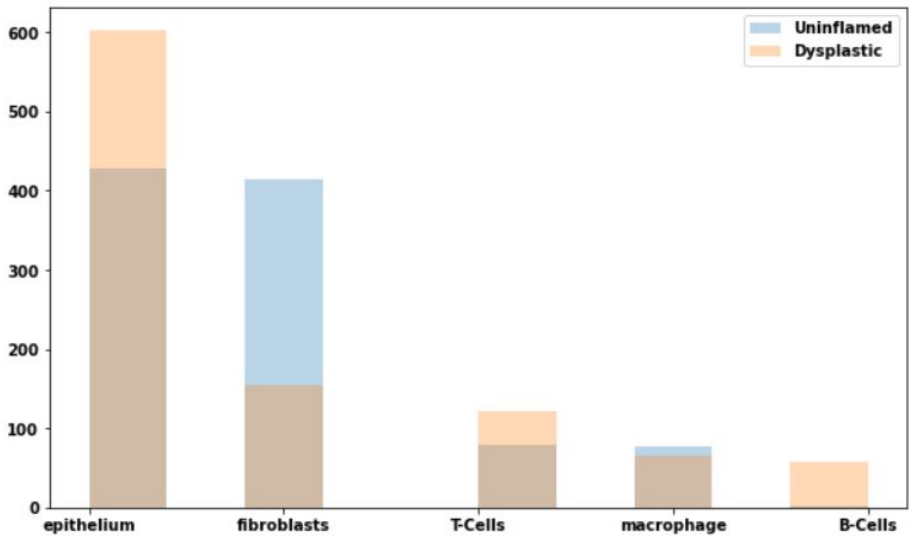


Figure 16: Evolution of cell types patient 4

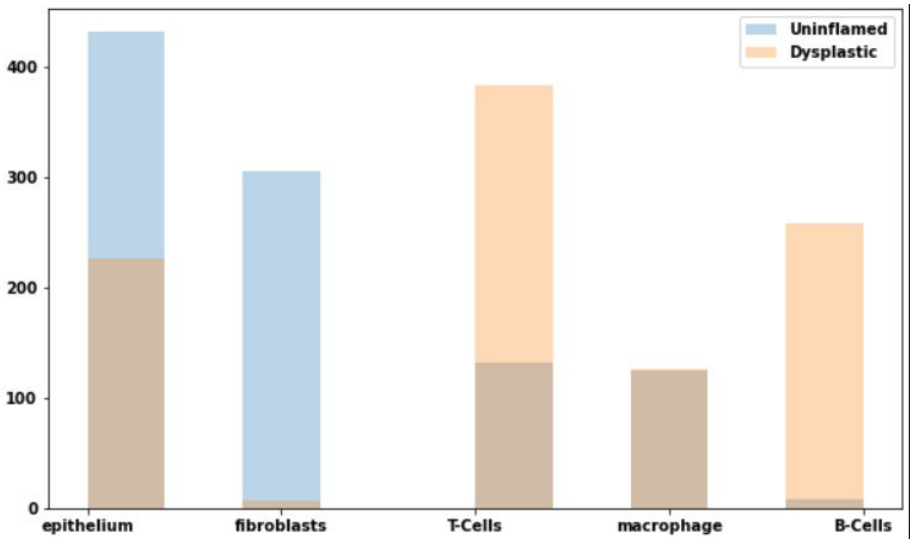
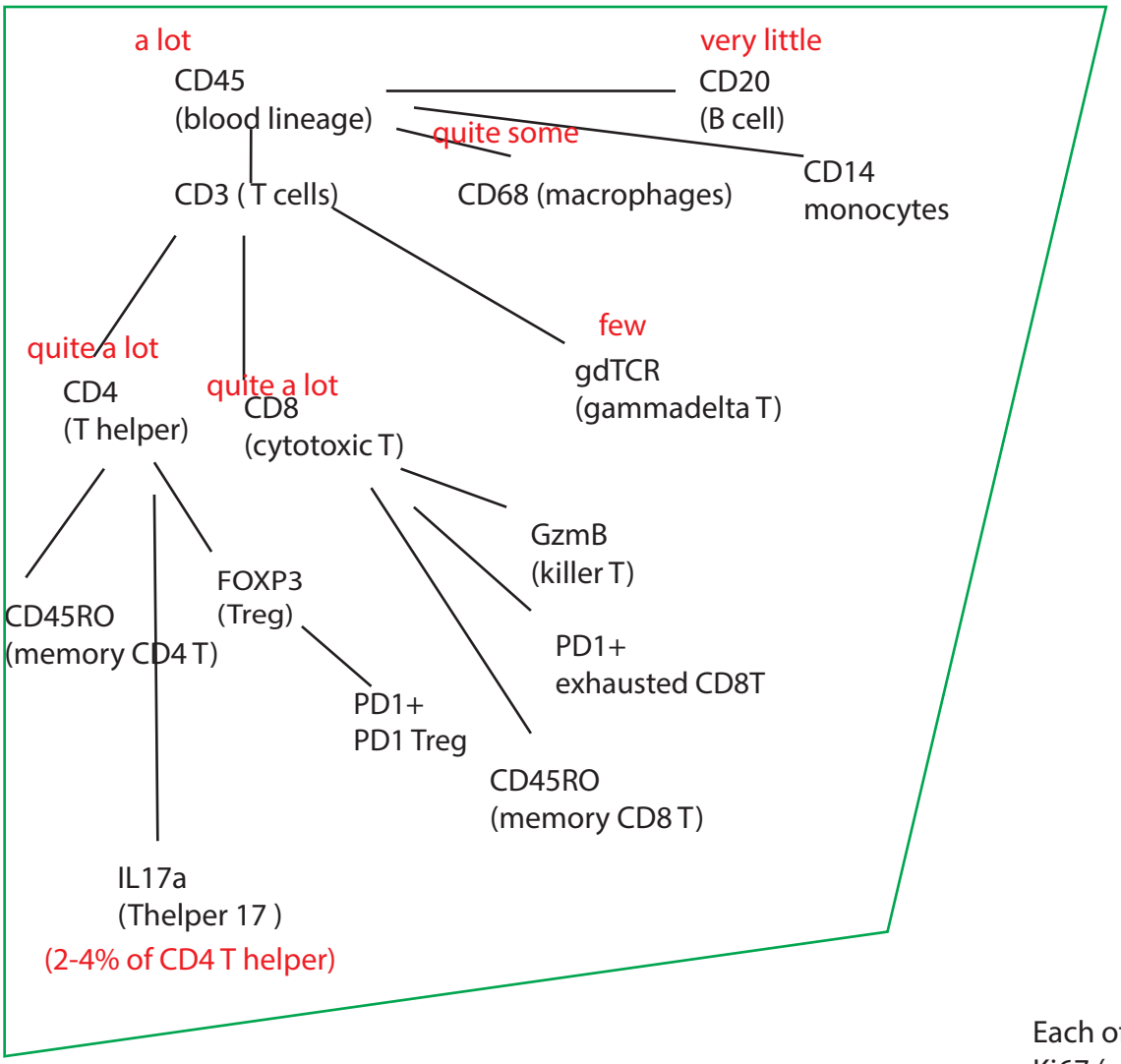


Figure 17: Evolution of cell types patient 5

immune cells



Each of these cells can express
Ki67 (cell division)
pan-AKT (present in all cells)
P-ERK (active Ras signaling)
P-S6 (active AKT signaling)

Ecadherin+ pan-keratin
(epithelium)

many, very close to fibroblasts, some T cells (gd/ CD8, maybe CD4+CD8+ in between epithelium)

aSMA
(fibroblasts)

quite some, very close to epithelium (surrounding the crypt)