Dimensionality reduction (PCA, umap...)

June 2, 2021

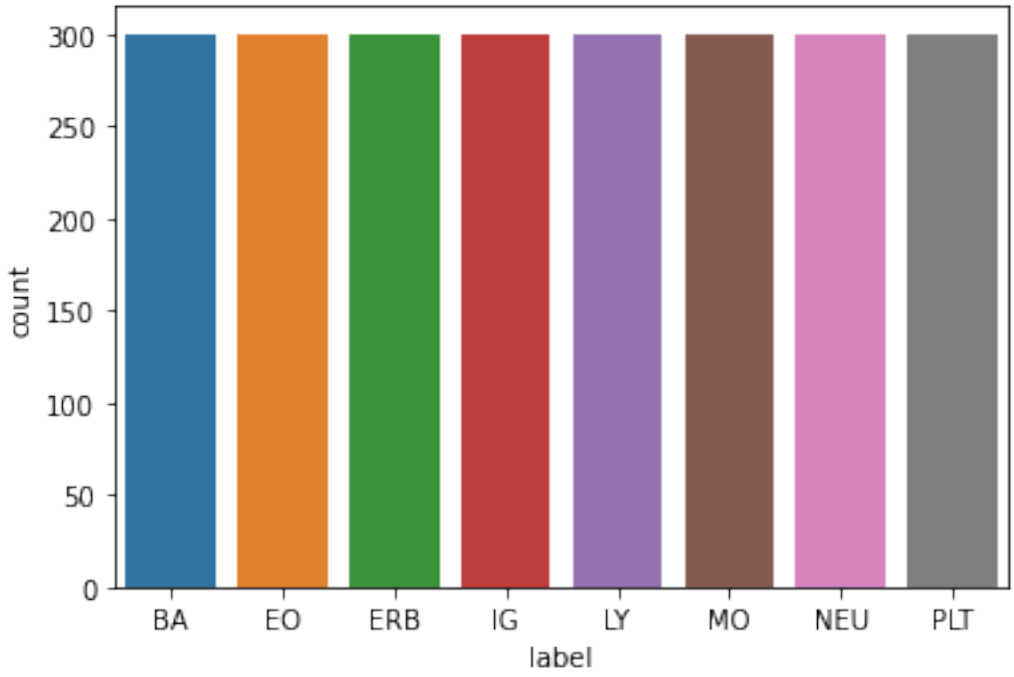
```
                                          img_paths      id label cell_type  \
0  ../../data/PBC_dataset_normal_DIB/monocyte/MO_…  225079    MO  monocyte
1  ../../data/PBC_dataset_normal_DIB/monocyte/MO_…  582430    MO  monocyte
2  ../../data/PBC_dataset_normal_DIB/monocyte/MO_…  436409    MO  monocyte
3  ../../data/PBC_dataset_normal_DIB/monocyte/MO_…  648815    MO  monocyte
4  ../../data/PBC_dataset_normal_DIB/monocyte/MO_…  668574    MO  monocyte

   height  width  mean_brightness  mean_luminance
0     363    360       196.225564        0.756902
1     363    360       196.672727        0.757366
2     363    360       204.348235        0.797640
3     363    360       199.038259        0.770929
4     363    360       191.020018        0.734784
```

```
array(['BA', 'EO', 'ERB', 'IG', 'LY', 'MO', 'NEU', 'PLT'], dtype=object)
```

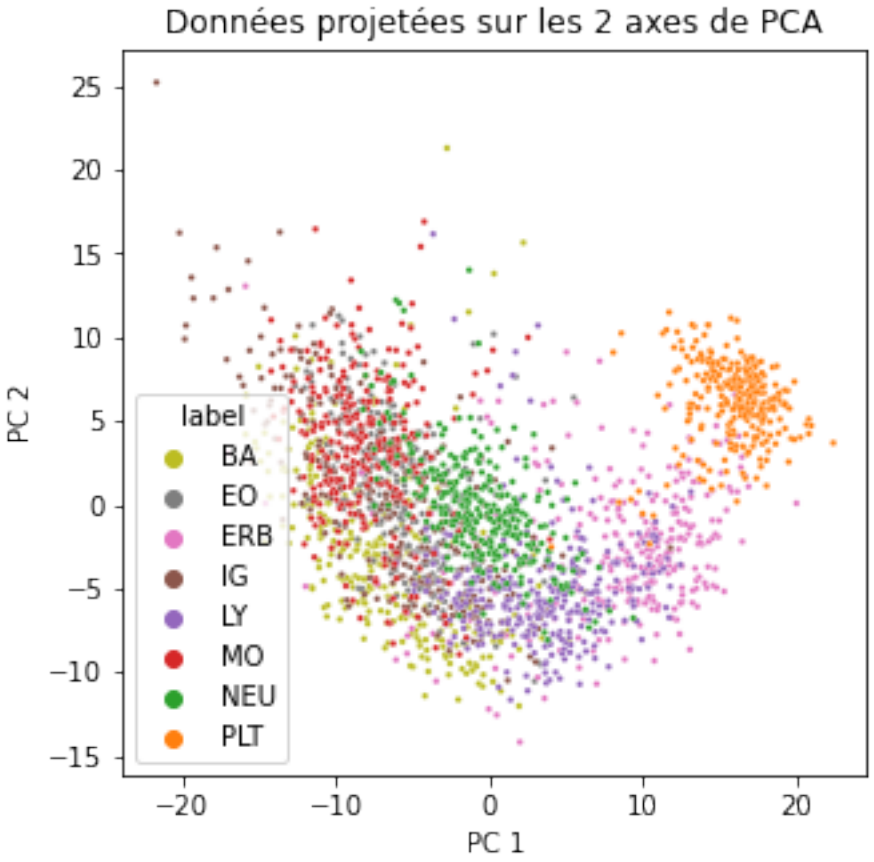
1 Preparing the data for PCA

1.0.1 Sampling from dataframe

```
<AxesSubplot:xlabel='label', ylabel='count'>
```

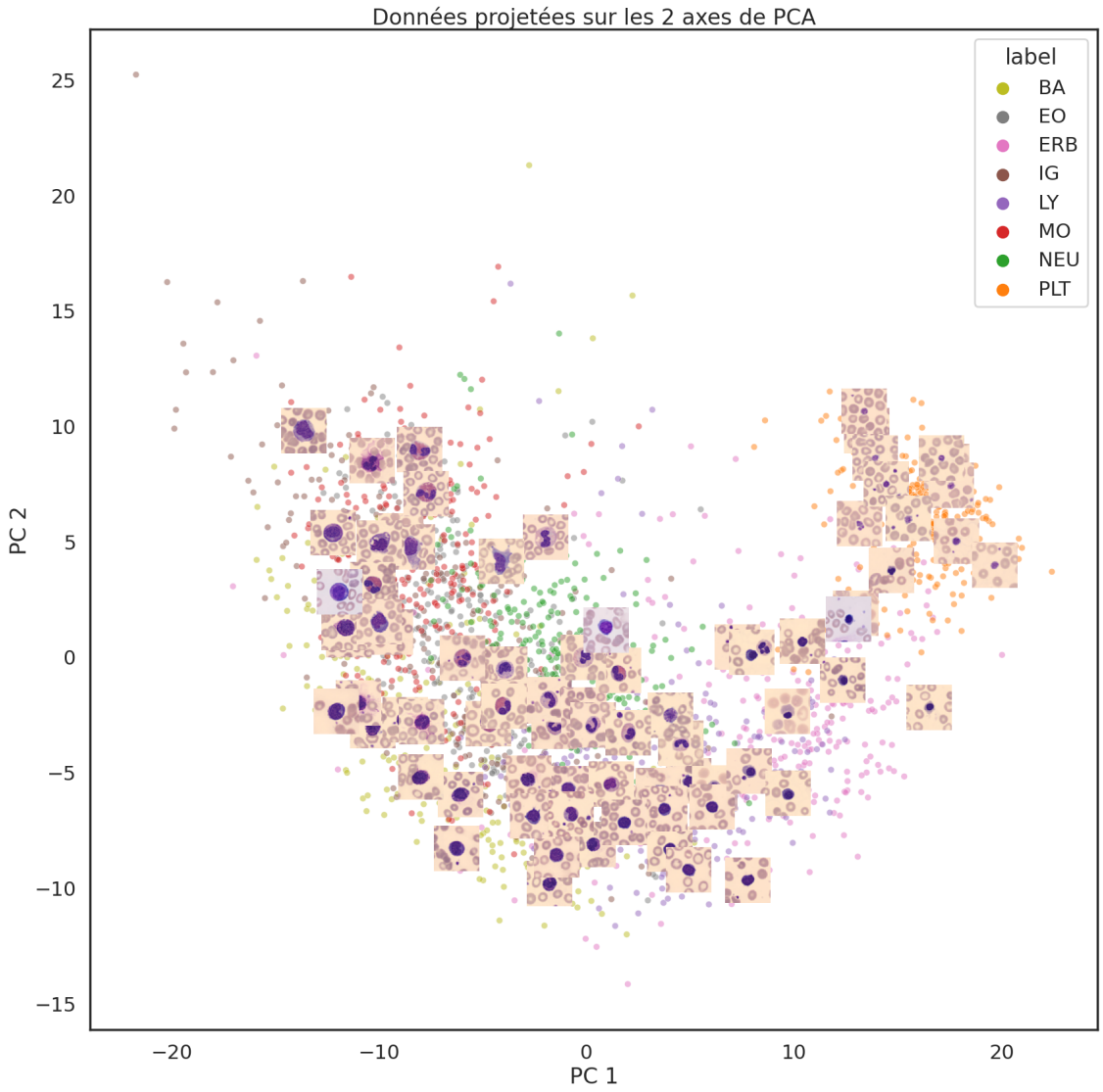

300
250
200
150
100
50
0
count
BA
EO
ERB
IG
LY
MO
NEU
PLT
label

2 PCA embedding



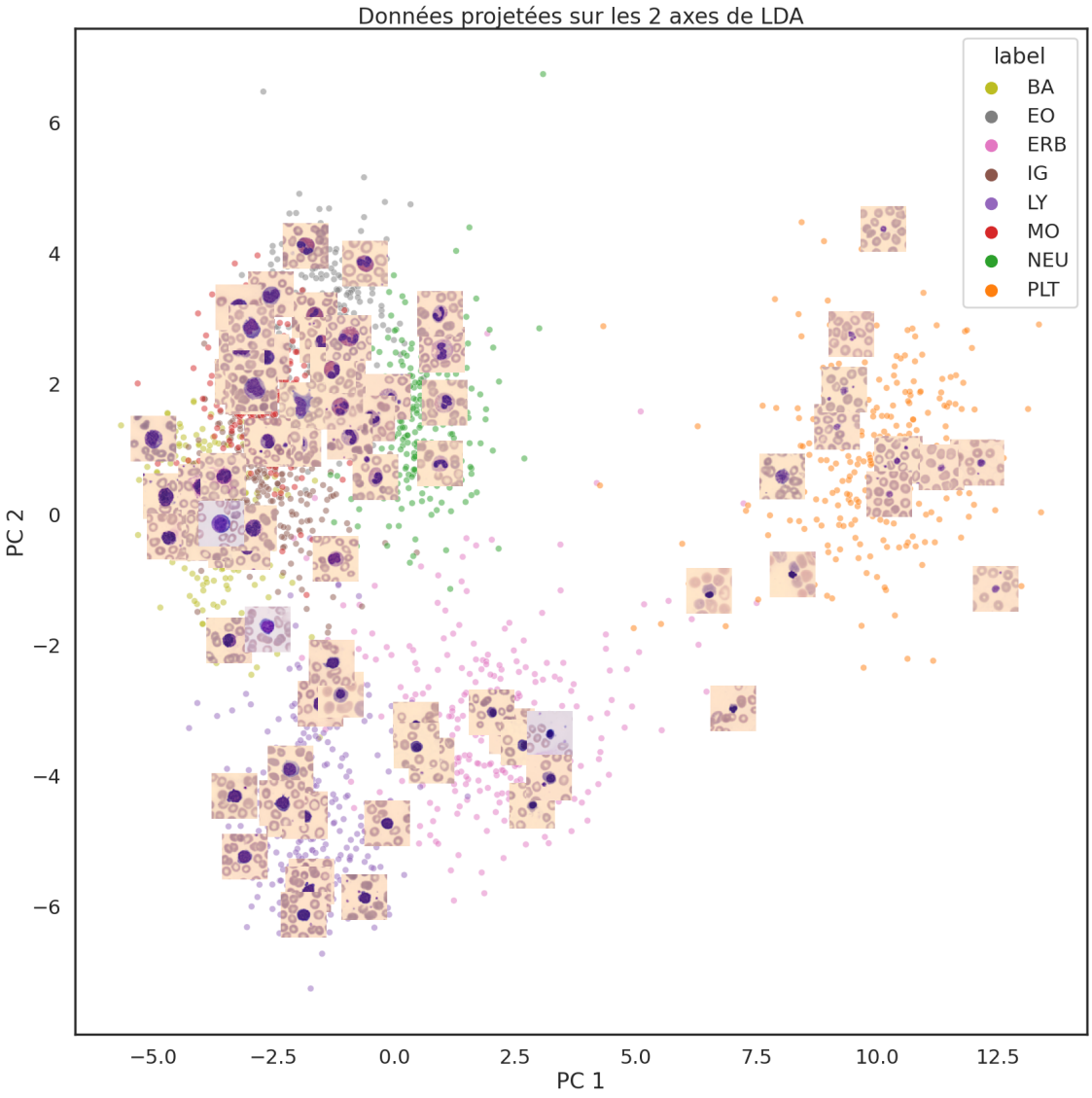
It looks like our PCA model is able to distinguish between some categories. In fact, as we will see later on, it is very efficient in clustering the platelets (orange points).

Plotting the previous scatter plot with annotated images for our data



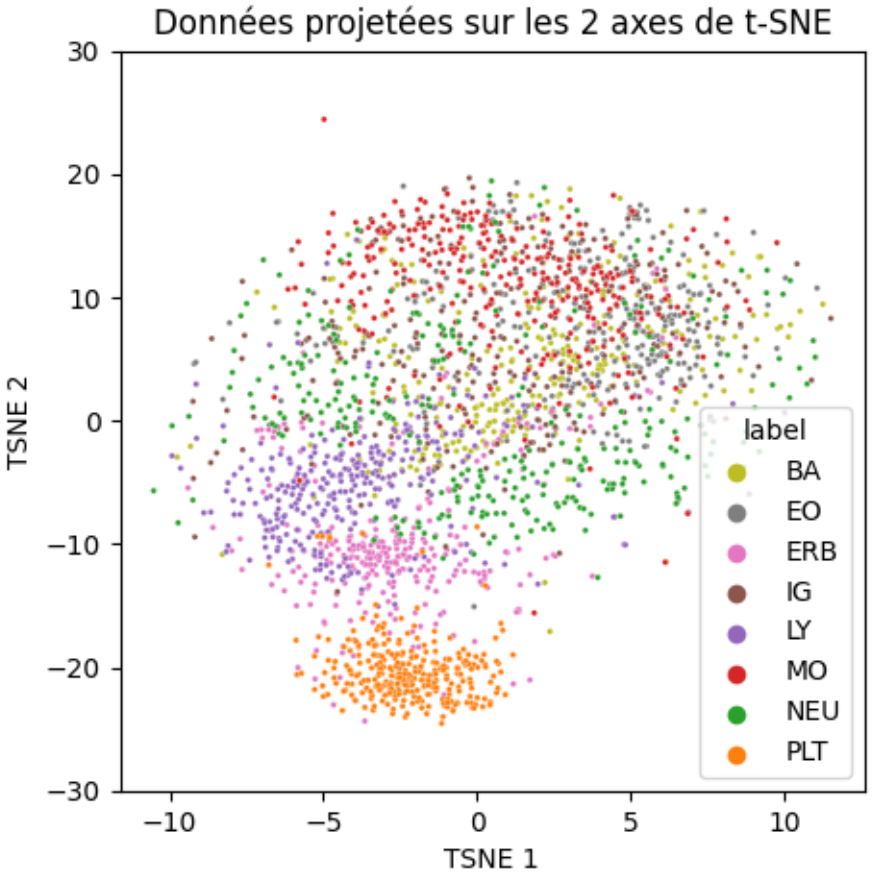
Just as we said previously, the PCA has been able to **distinguish between categories** according to the size of the bloodcells in the images, which makes it really useful for detecting platelet.

3 LDA embedding



Amazing ! Once again, we have been able to visually cluster some categories according to the bloodcell size.

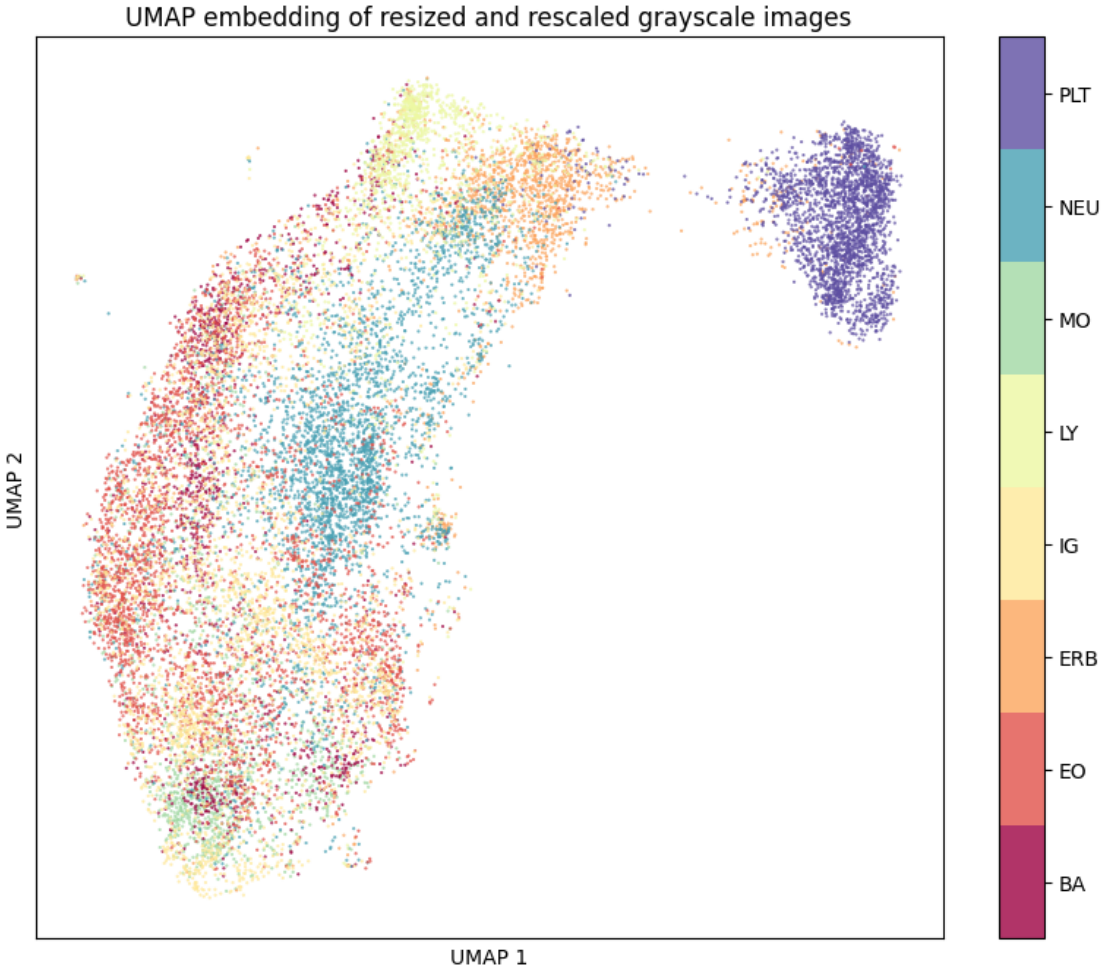
4 t-SNE embedding

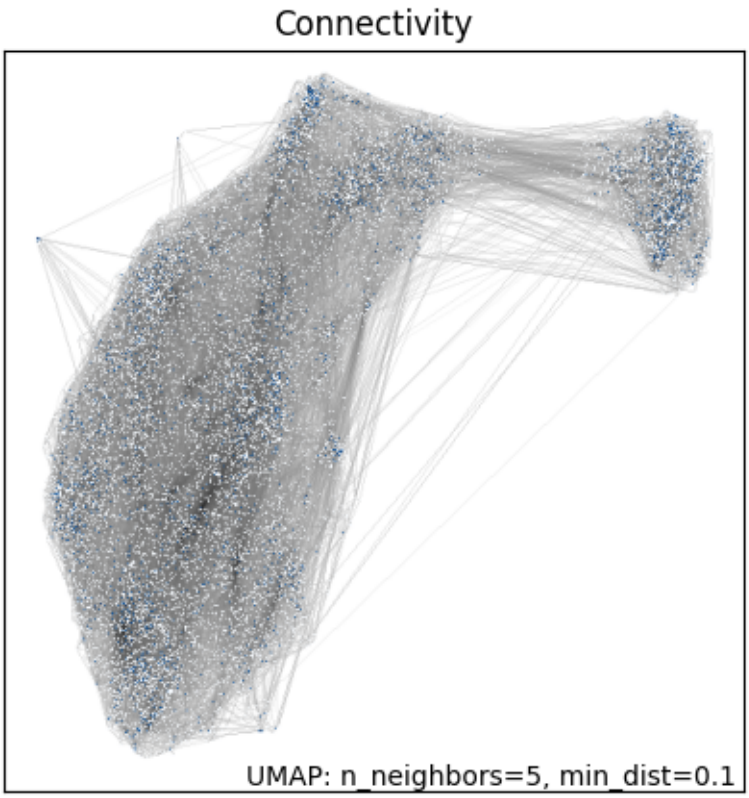


5 UMAP emdedding

While t-SNE is really good at conserving local distance, PCA is good at conversing global structure. The UMAP algorithm is good at maintaining both. Let's have a look.

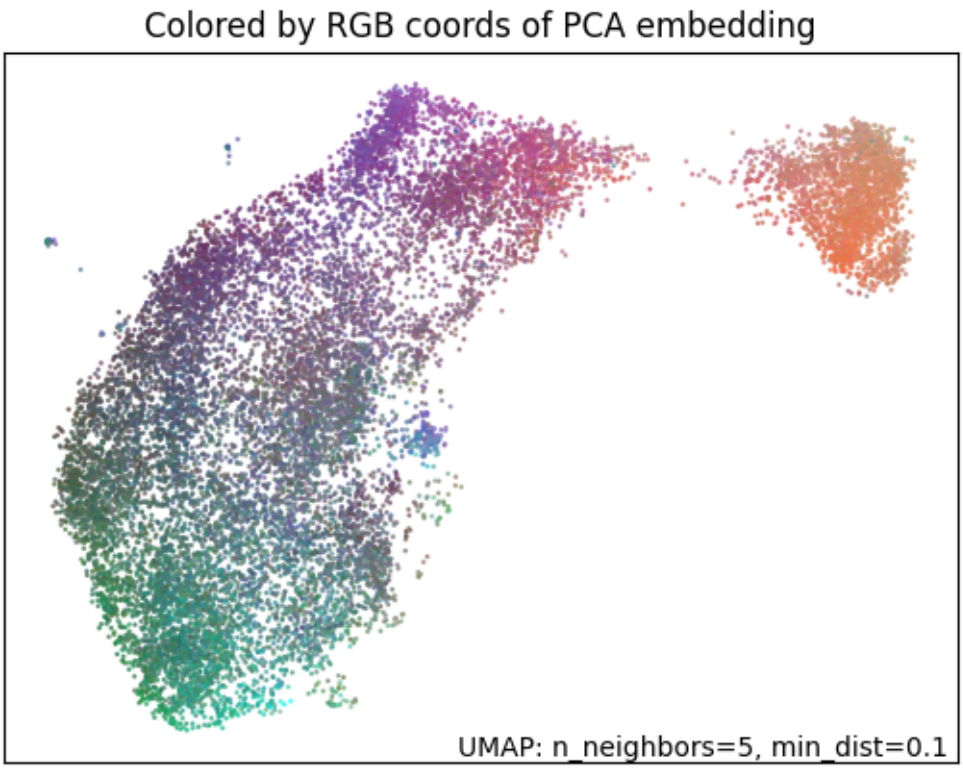
5.0.1 Applying our UMAP model





As with the other methods of dimension reduction, the platelets form a single cluster. We can also see that some of the cells appears to be in sub-cluster within the main cluster.

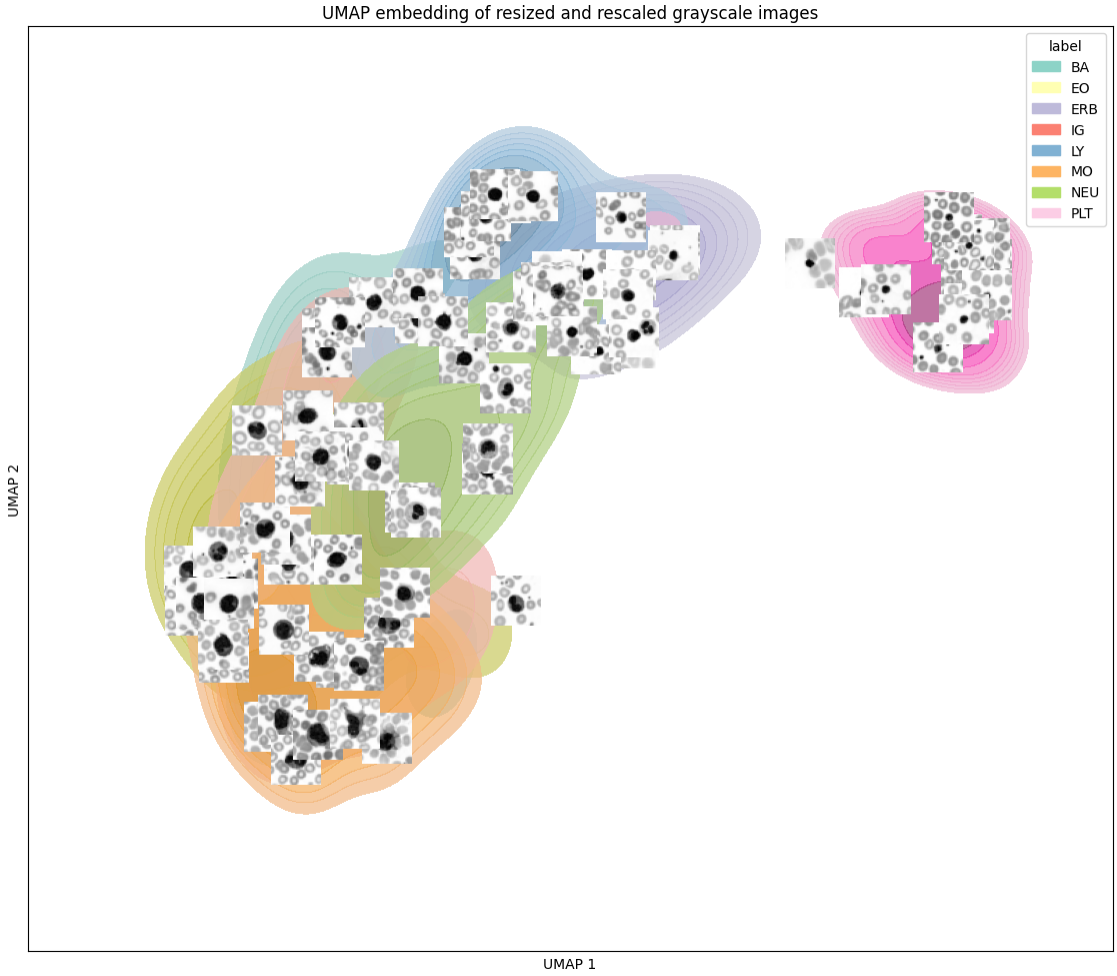
We can check whethet the UMAP is performing good on the global distances using the PCA diagnosis tool.



The plot that the UMAP was able to integrate **global variance** within its dimension reduction.

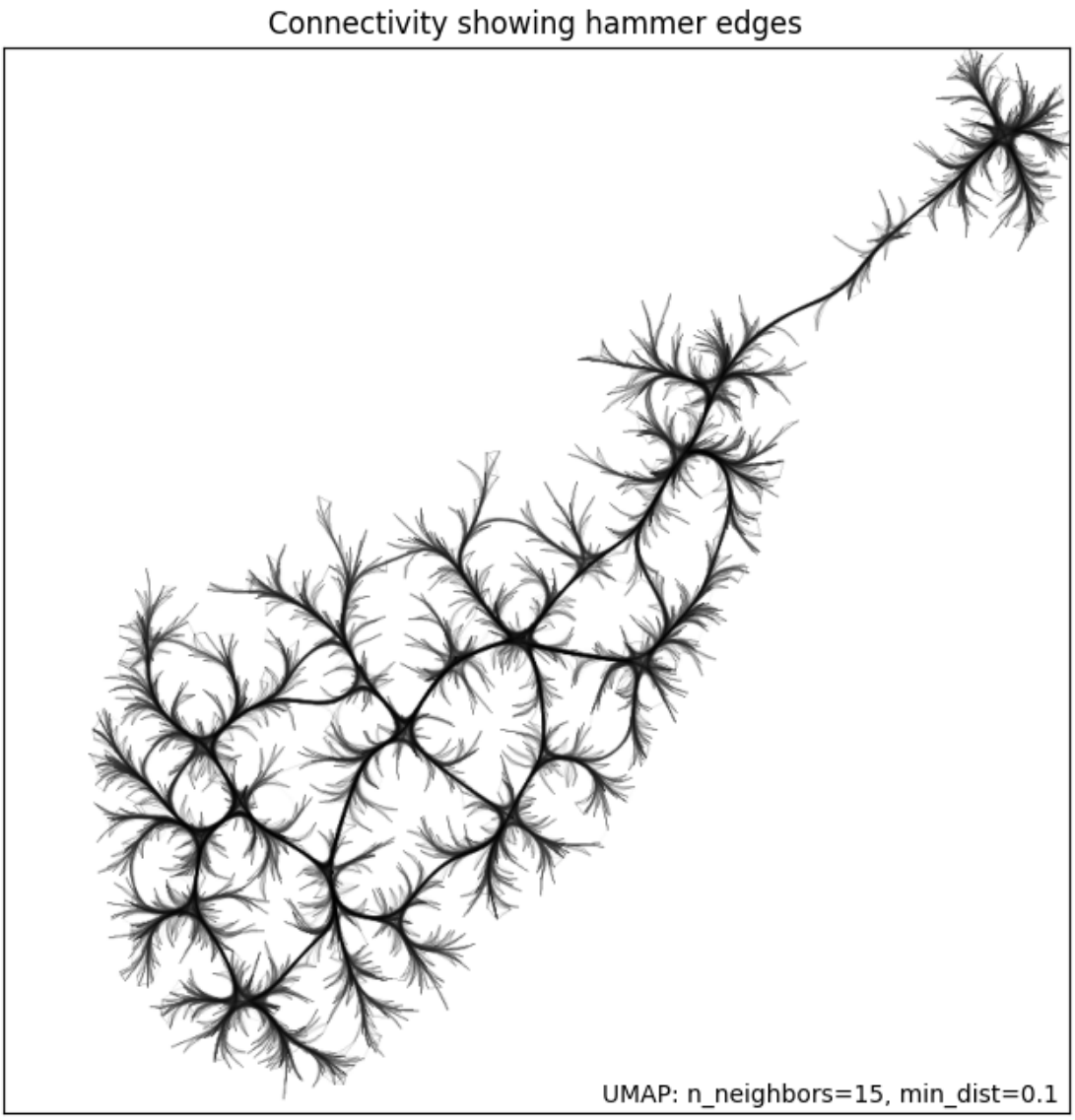
5.0.2 UMAP on a subsample

Let's plot the image on a **subsample fitted with UMAP**



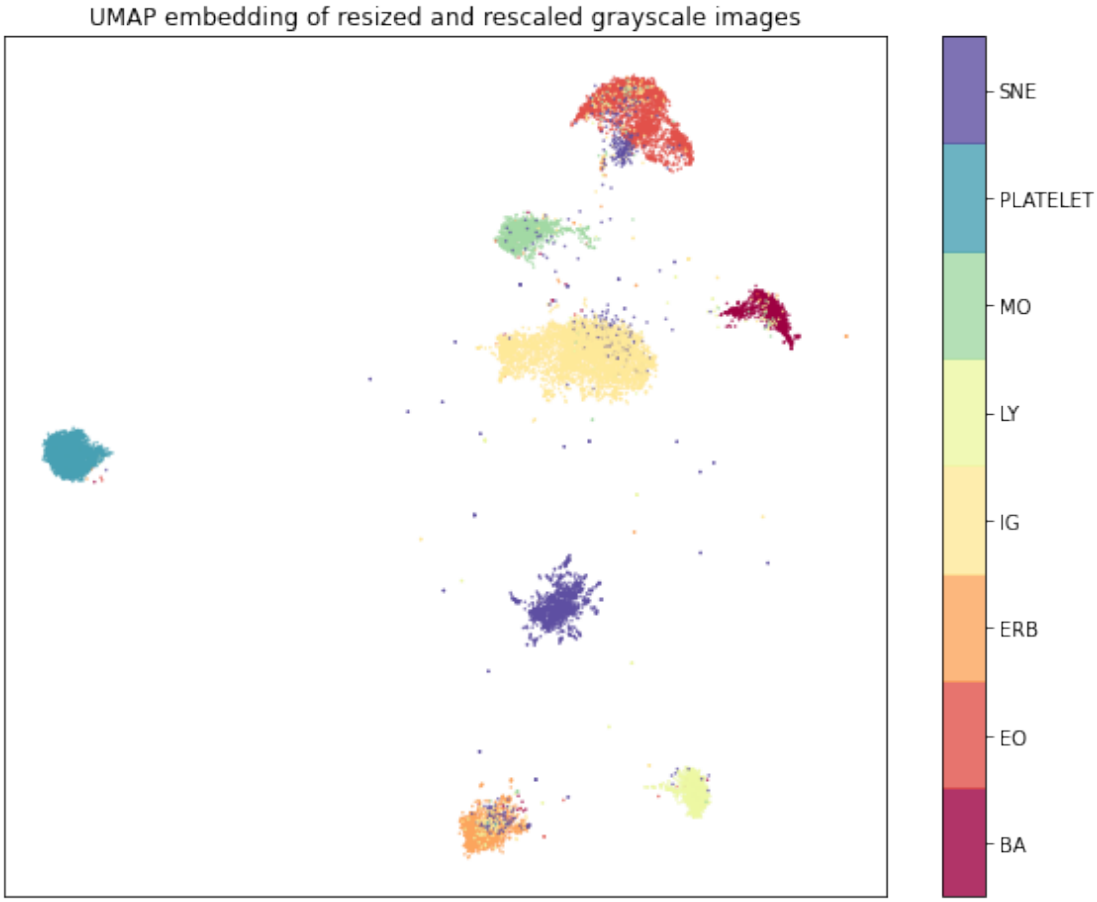
We can clearly see here that the platelets stand out, moreover the UMAP is able to group other cells into subclusters and UMAP component 2 appears to differentiate cells by size and brightness.

```
Text(0.5, 1.0, 'Connectivity showing hammer edges')
```



5.0.3 Supervised UMAP

We can also try to **supervised UMAP** with our labels



6 Conclusion

Using dimension reduction, we found consistent separation of platelets from the rest of the cell. Platelets thus appears as an easy cell to classify. Those results are in agreements with the data from the luminance and deconvolution exploration. The UMAP revealed that the cells tend to cluster together within an important cluster. On the PCA and UMAP, we could also observed that the size of cells and the brightness contribute to the global variance.

The supervised UMAP is an interesting model for dimension reduction as it manage to truly separate each cell groups. We will use it for our base model.