Exploration (stats, brightness radial profiles, deconvolution)

June 2, 2021

1 EDA - stats, brightness radial profiles, deconvolution

2 DataFrame et première exploration

(PBC_dataset : cellules saines)

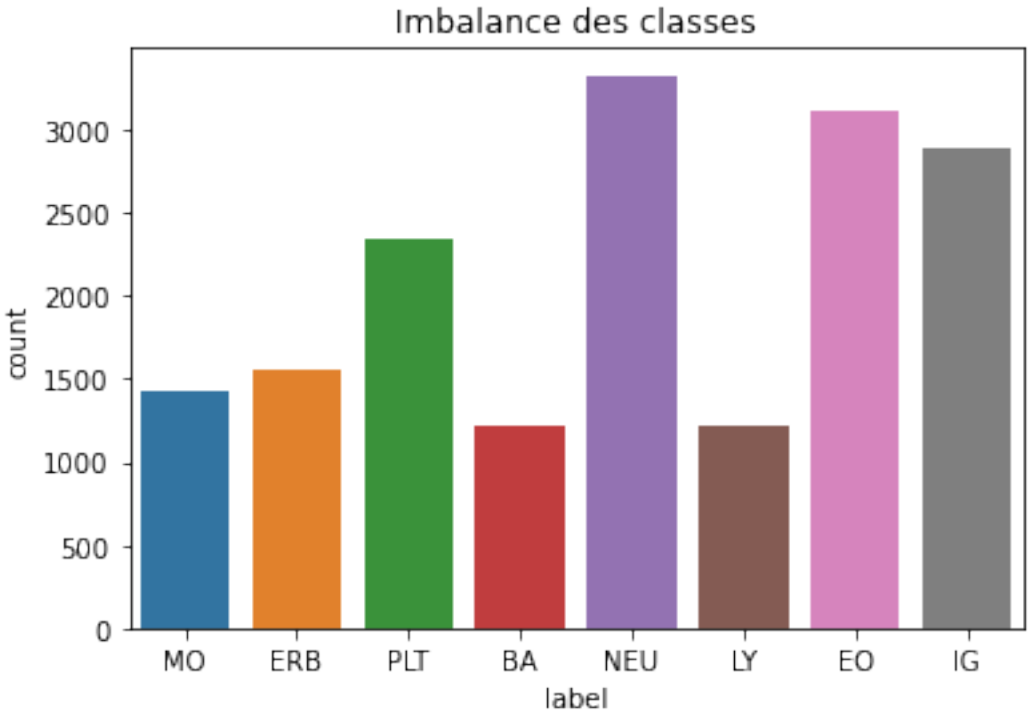
```
                                          img_paths      id label cell_type  \
0  ../../data/PBC_dataset_normal_DIB/monocyte/MO_…  225079    MO  monocyte
1  ../../data/PBC_dataset_normal_DIB/monocyte/MO_…  582430    MO  monocyte
2  ../../data/PBC_dataset_normal_DIB/monocyte/MO_…  436409    MO  monocyte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

Liste des variables du DF :

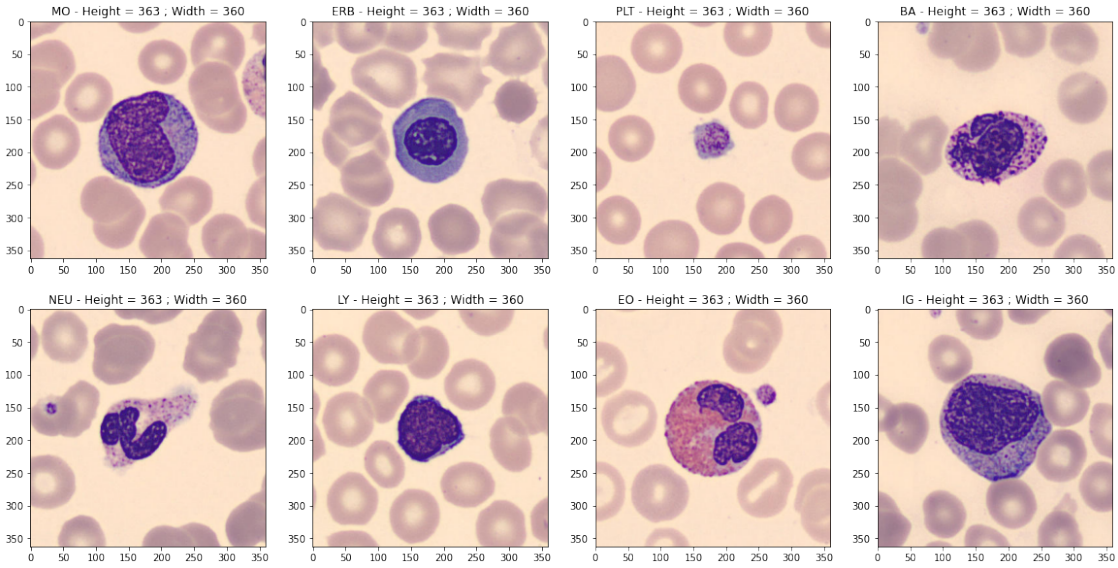
- img_paths : chemin d'accès de l'image
- id : numero de chaque image
- label : classe de la cellule
  - MO : monocytes
  - BA : basophiles
  - EO : éosinophiles
  - PLT : plaquettes
  - LY : lymphocytes
  - ERB : erythroblastes
  - IG : granulocytes immatures
  - NEU : neutrophiles
- cell_type : même information que label, en toutes lettres
- height : hauteur en nombre de pixels

- width : largeur en nombre de pixels
- mean_brightness : luminosité moyennée sur toute l'image (couleur)
- mean_luminance : luminosité moyennée sur toute l'image (niveaux de gris)

	label
NEU	3329
EO	3117
IG	2895
PLT	2348
ERB	1551
MO	1420
BA	1218
LY	1214



Les classes ne sont pas équilibrées, il faudra peut-être avoir recours à de l'augmentation de données.



On a un premier aperçu des différences entre chaque classe de cellules : - couleur - taille - forme - forme du noyau

A première vue, on peut remarquer des similitudes entre certaines types de cellules : par exemple MO et IG (avec BA, dans une moindre mesure), NEU et EO (forme du noyau, parfois coloration).

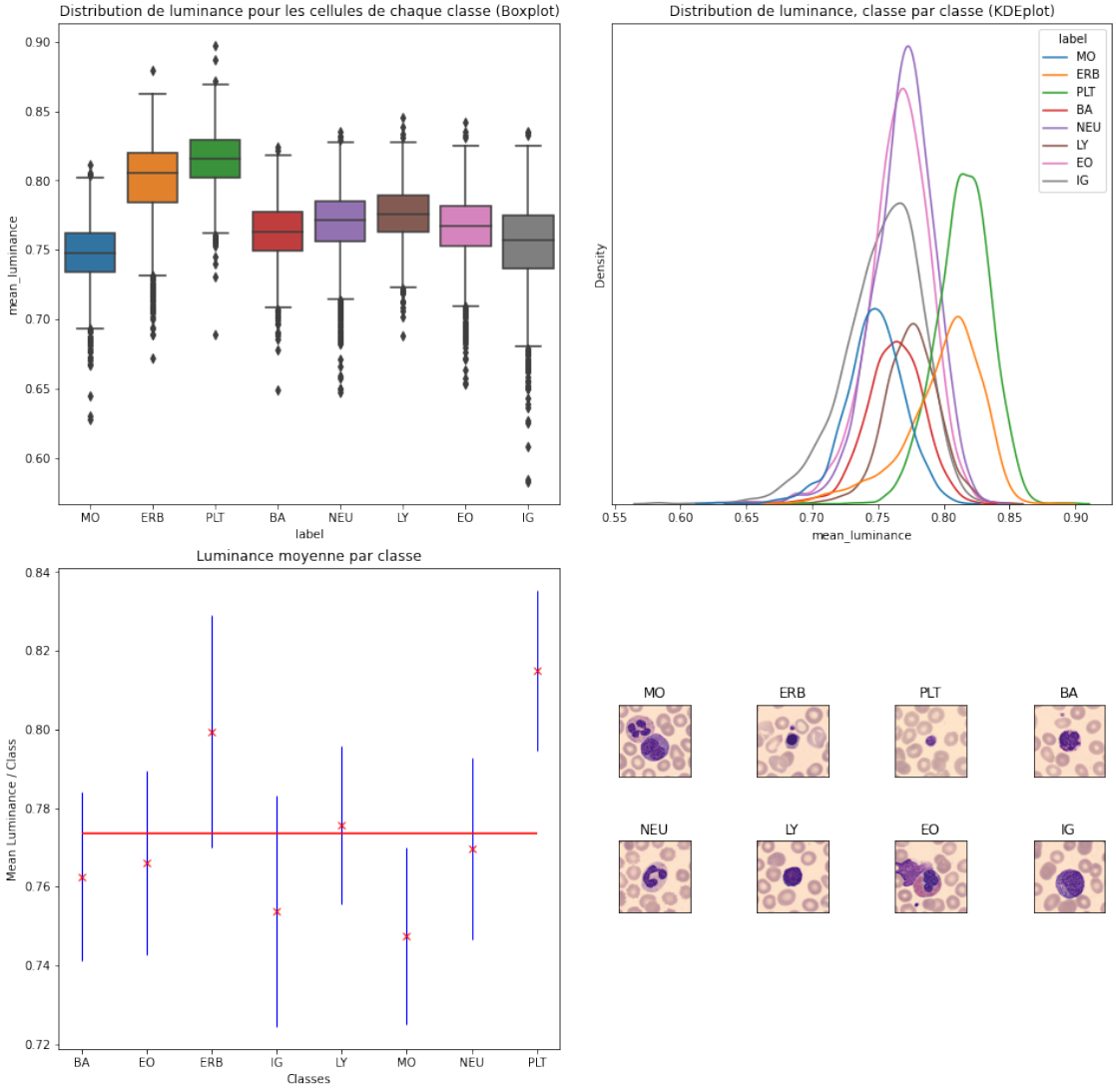
Des éléments peuvent venir parasiter l'information importante : - globules rouges en fond, dont le nombre et l'aspect peuvebt différer fortement d'une image à l'autre, - couleur de l'image : la majorité sont globalement roses, mais d'autres tirent vers le gris.

Les images sont au format RGB : M*N pixels + 3 canaux (rouge, vert, bleu) : on remarque que toutes n'ont pas la même taille.

3 Distribution de la taille des images

La majorité des images du dataset sont au format 363*360 : - pour la suite de l'exploration, les images seront redimensionnées en 360*360 - pour la modélisation, elles seront réduites au format 256*256

4 Distribution de luminosité moyenne, classe par classe



Ces deux graphiques mettent en évidence les points suivants : - deux classes sortent légèrement du lot : les plaquettes (PLT) et les érythroblastes (ERB), - toutes les classes présentent des outliers, essentiellement de faible luminosité.

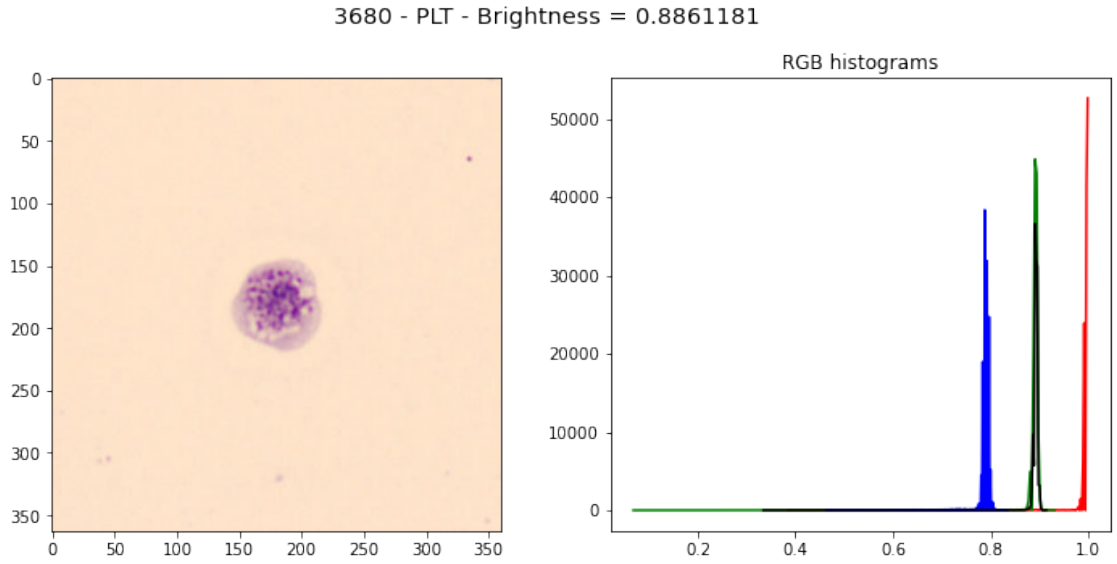
L'obscurité sur nos images est liée à trois facteurs : - la cellule d'intérêt (surtout son noyau lorsqu'il y en a un), - les globules rouges en fond, - dans une moindre mesure, la présence d'une ou plusieurs autres cellules d'intérêt, en périphérie de l'image.

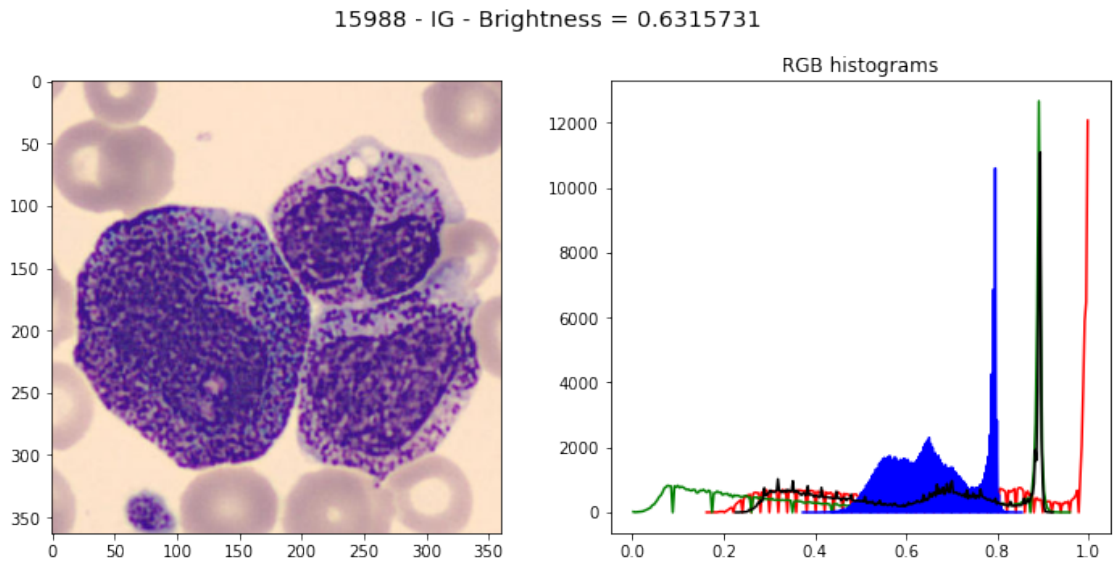
Comme les globules rouges et les cellules "parasites" sont distribués aléatoirement d'une image à l'autre, leur présence ne doit pas vraiment influencer la luminosité moyenne d'une classe. On peut donc expliquer qualitativement les différences de luminosité moyenne d'une classe à l'autre en regardant un exemple pour chaque type de cellule : - les plaquettes (PLT) sont les cellules les plus petites et sont en plus dépourvues de noyau : l'image dans sa globalité est donc moins sombre que

pour les autres cellules ; les érythroblastes sontn plus volumineux, mais leur noyau (la partie la plus sombre) est relativement petit par rapport au volume de la cellule et par rapport aux autres types de cellules : les images ERB sont donc plus lumineuses que pour les autres classes, sauf PLT ; - les images les plus sombres sont les granulocytes (IG) et les monocytes (MO), qui sont de grosses cellules avec des noyaux assez volumineux (et sombres) ; - le noyau d'un lymphocyte (LY) occupe la quasi-totalité du volume de la cellule, qui est donc plutôt sombre, mais cette cellule est plutôt petite, l'image est donc en moyenne plus lumineuse qu'une image d'éosinophile (EO) ou de neutrophile (NEU), cellules dont le noyau a un volume similaire à celui d'un lymphocyte mais qui sont plus grosses que lui.

Il n'y a donc **probablement pas de biais de luminosité** d'une classe de cellules à l'autre.

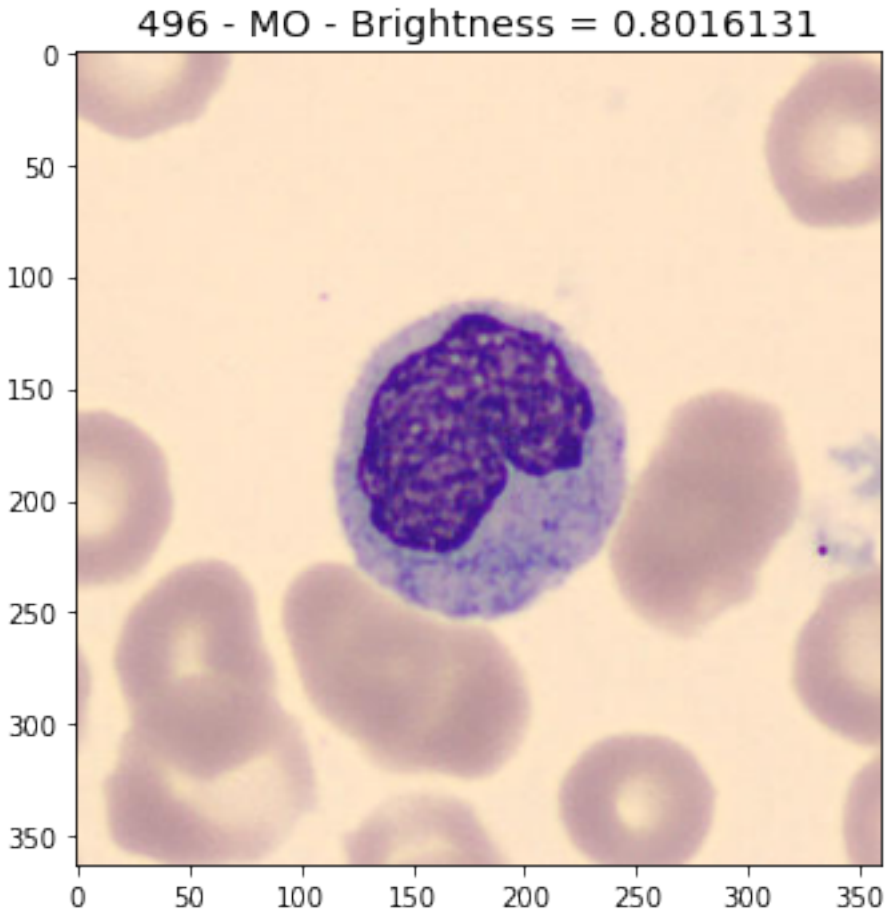
Affichons maintenant les outliers extrêmes : l'image la plus lumineuse, et l'image la plus sombre.

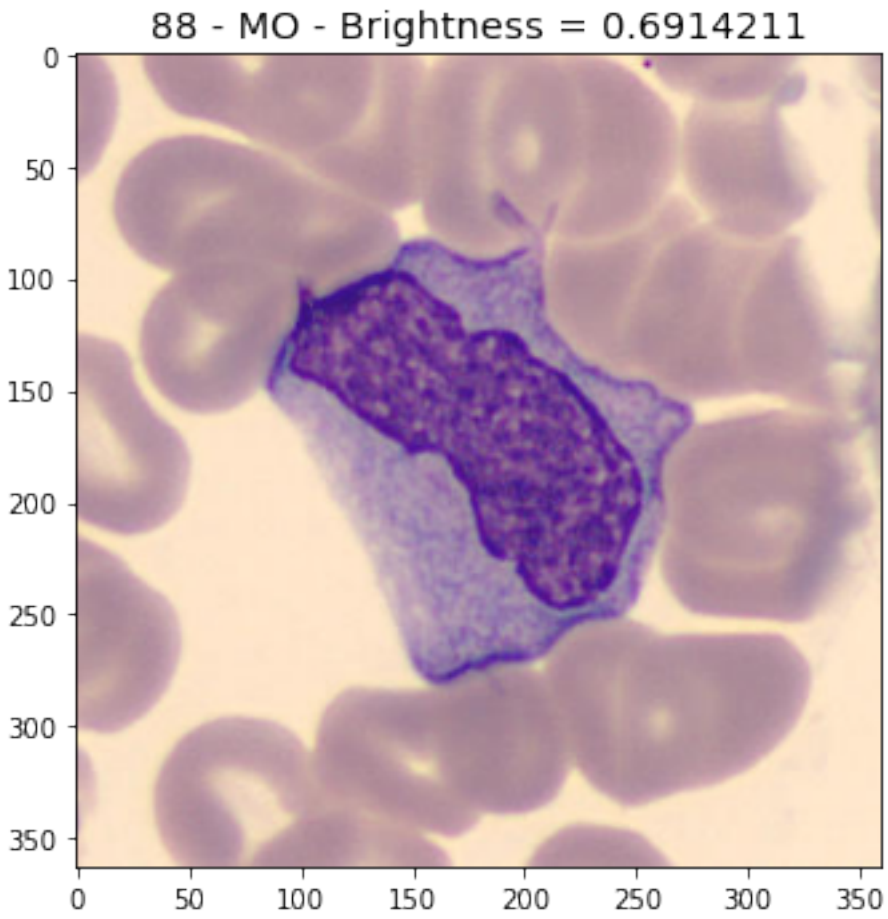


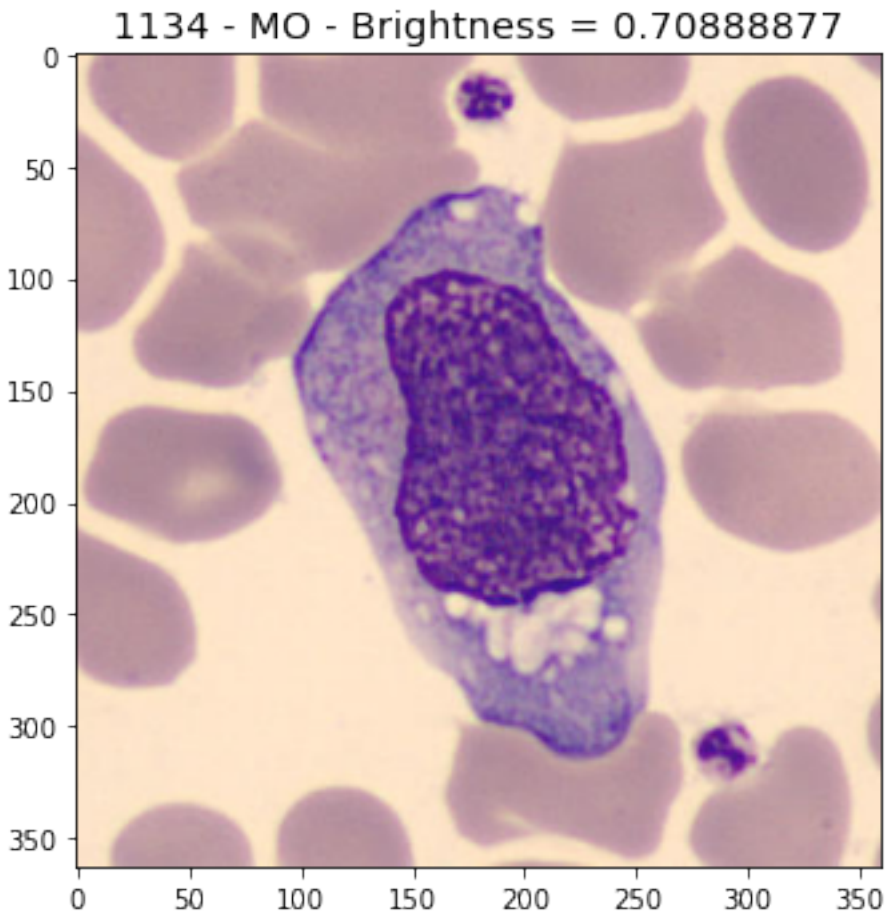


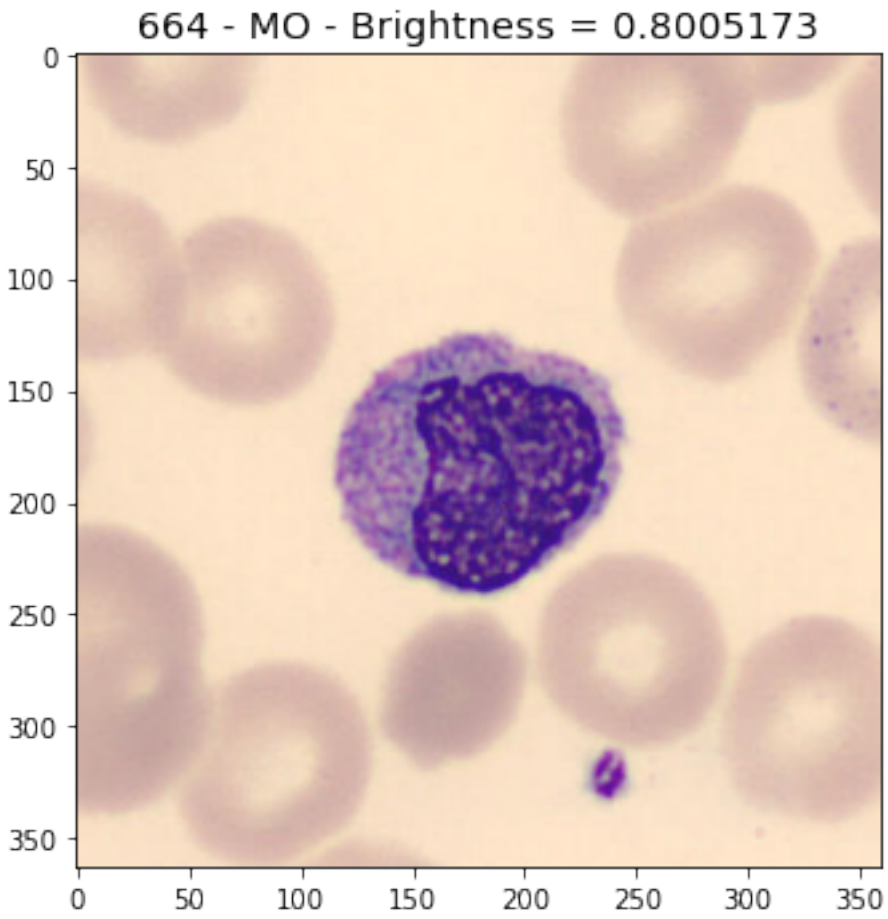
- L'image la plus lumineuse représente **une plaquette toute seule**, sans globules rouges en arrière plan.
- L'image la plus sombre représente une **agglomération de granulocytes immatures** qui occupe une grande partie de l'image. Le dataset comporte plusieurs images de ce type.

On peut maintenant afficher quelques outliers pris au hasard dans le DataFrame :




88 - MO - Brightness = 0.6914211
0
50
100
150
200
250
300
350
0
50
100
150
200
250
300
350


1134 - MO - Brightness = 0.70888877
0
50
100
150
200
250
300
350
0
50
100
150
200
250
300
350

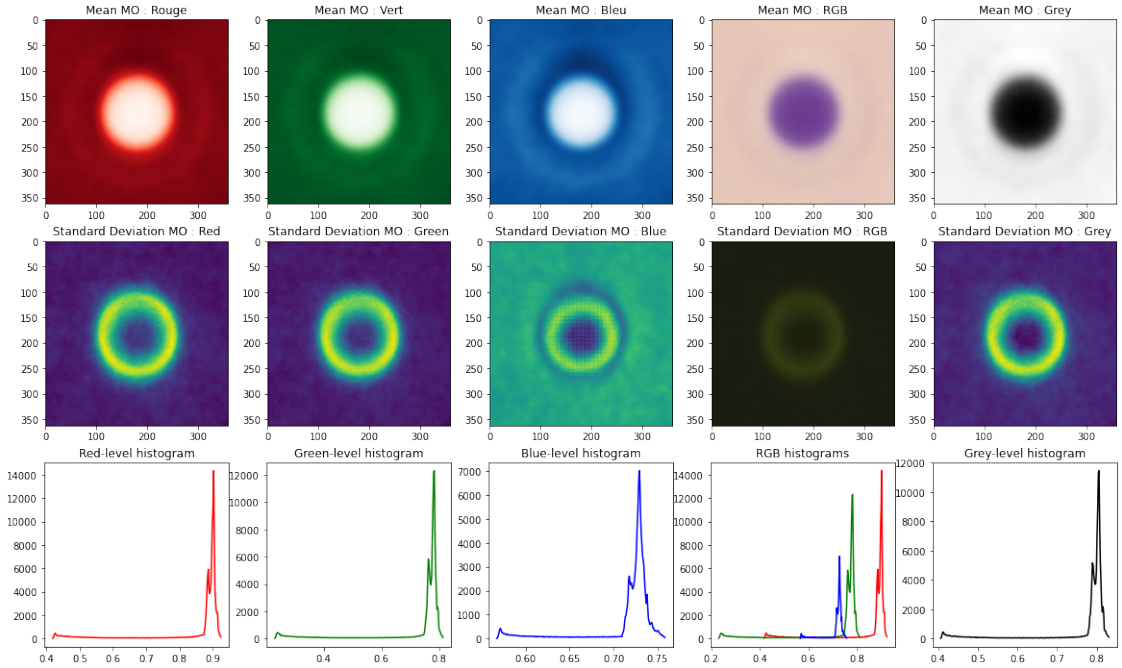


On remarque que beaucoup des outliers sombres sont caractérisés par la présence d'un nombre important de globules rouges autour de la cellule d'intérêt, ou par la présence de plusieurs cellules à noyau. Quant aux outliers lumineux, ce sont des images dans lesquelles il y a très peu (voire aucun) de globules rouges.

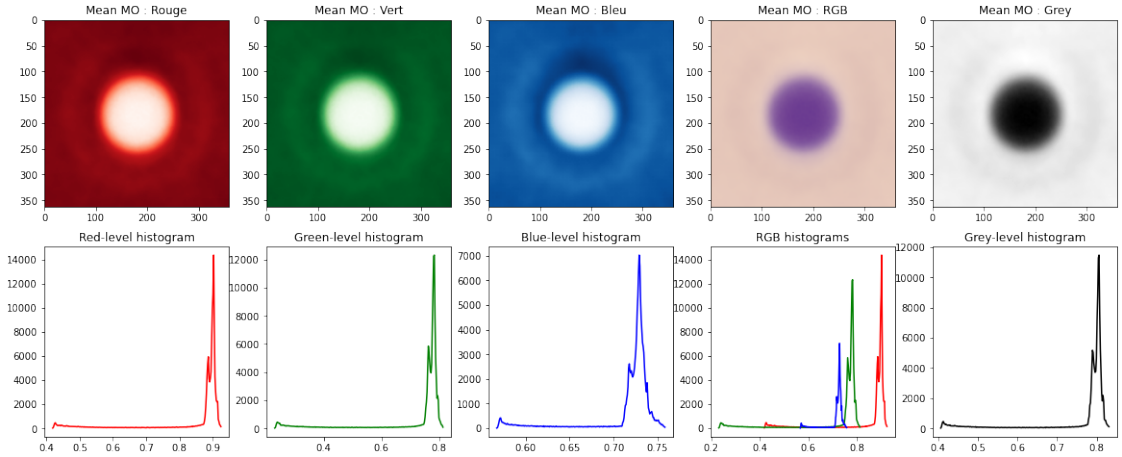
5 Cellules "moyennes"

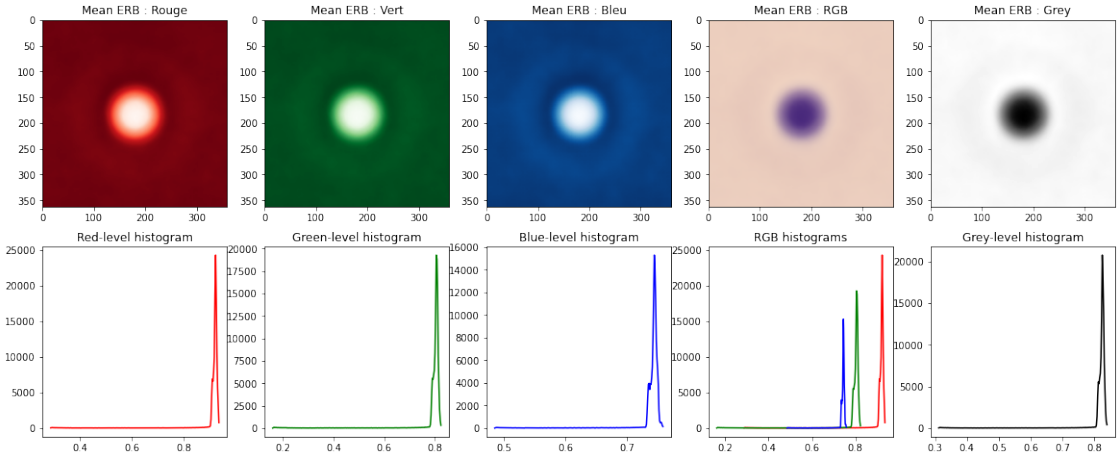
On va maintenant représenter une cellule "moyenne" pour chaque type de cellule :

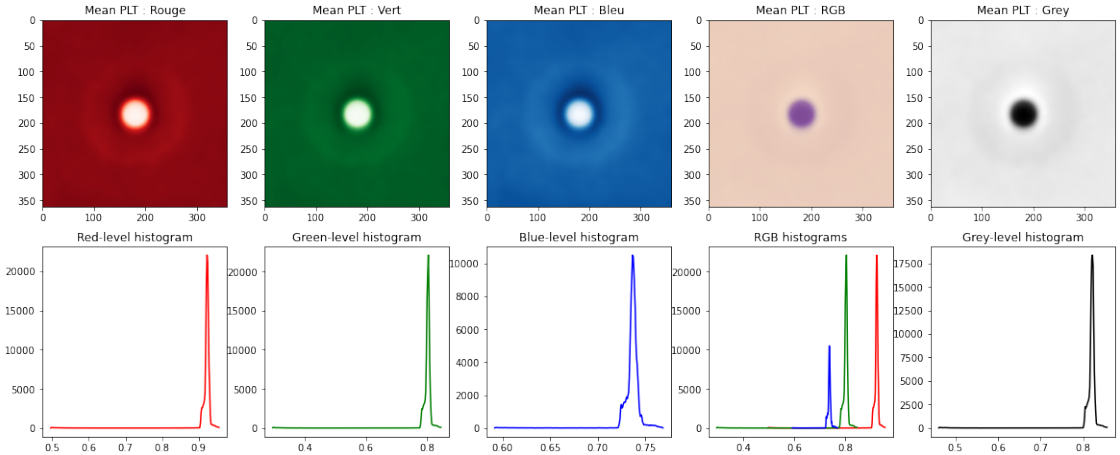
```
Number of cells in MO : 1420
```

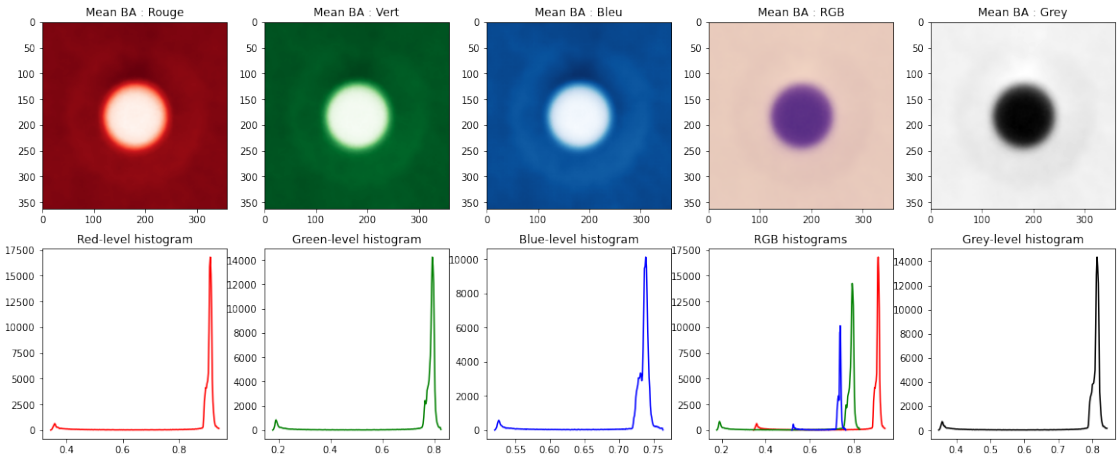


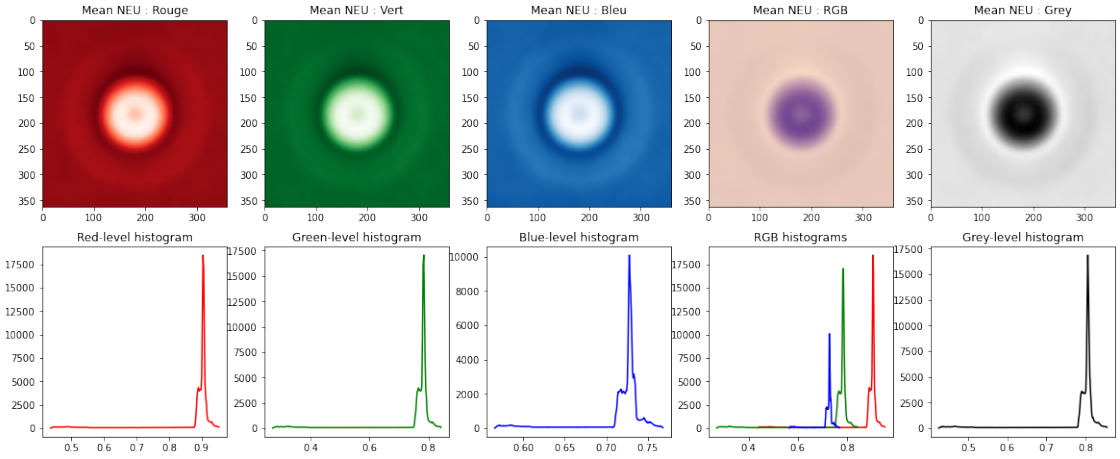

```
Number of cells in MO : 1420
Number of cells in ERB : 1551
Number of cells in PLT : 2348
Number of cells in BA : 1218
Number of cells in NEU : 3329
Number of cells in LY : 1214
Number of cells in EO : 3117
Number of cells in IG : 2895
```

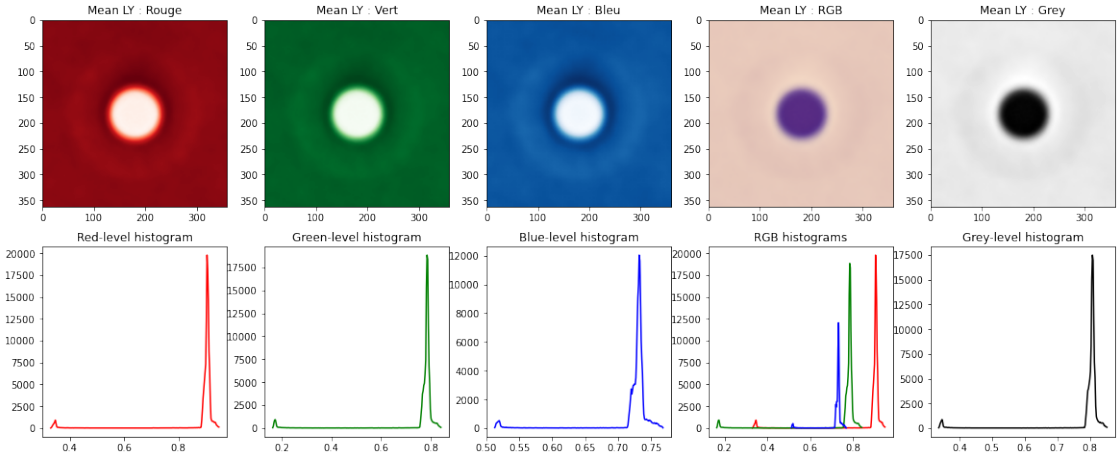


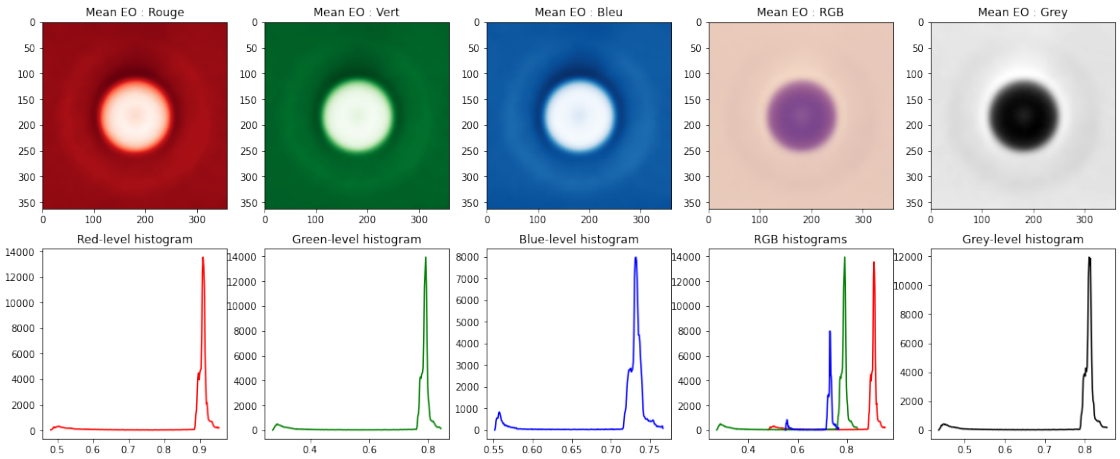

Mean ERB : Rouge
Mean ERB : Vert
Mean ERB : Bleu
Mean ERB : RGB
Mean ERB : Grey
Red-level histogram
Green-level histogram
Blue-level histogram
RGB histograms
Grey-level histogram

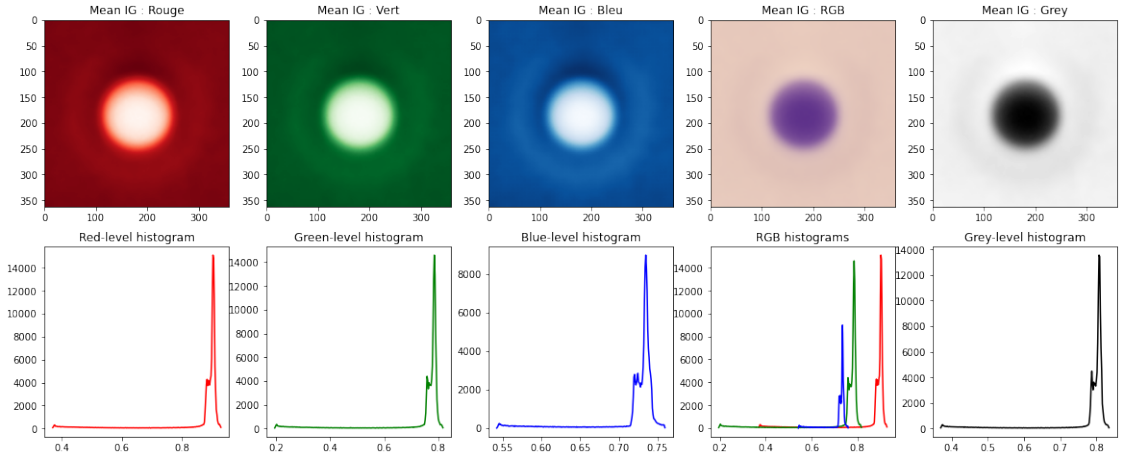

Mean PLT : Rouge
Mean PLT : Vert
Mean PLT : Bleu
Mean PLT : RGB
Mean PLT : Grey
Red-level histogram
Green-level histogram
Blue-level histogram
RGB histograms
Grey-level histogram


Mean BA : Rouge
Mean BA : Vert
Mean BA : Bleu
Mean BA : RGB
Mean BA : Grey
Red-level histogram
Green-level histogram
Blue-level histogram
RGB histograms
Grey-level histogram


Mean NEU : Rouge
Mean NEU : Vert
Mean NEU : Bleu
Mean NEU : RGB
Mean NEU : Grey
Red-level histogram
Green-level histogram
Blue-level histogram
RGB histograms
Grey-level histogram


Mean LY : Rouge
Mean LY : Vert
Mean LY : Bleu
Mean LY : RGB
Mean LY : Grey
Red-level histogram
Green-level histogram
Blue-level histogram
RGB histograms
Grey-level histogram

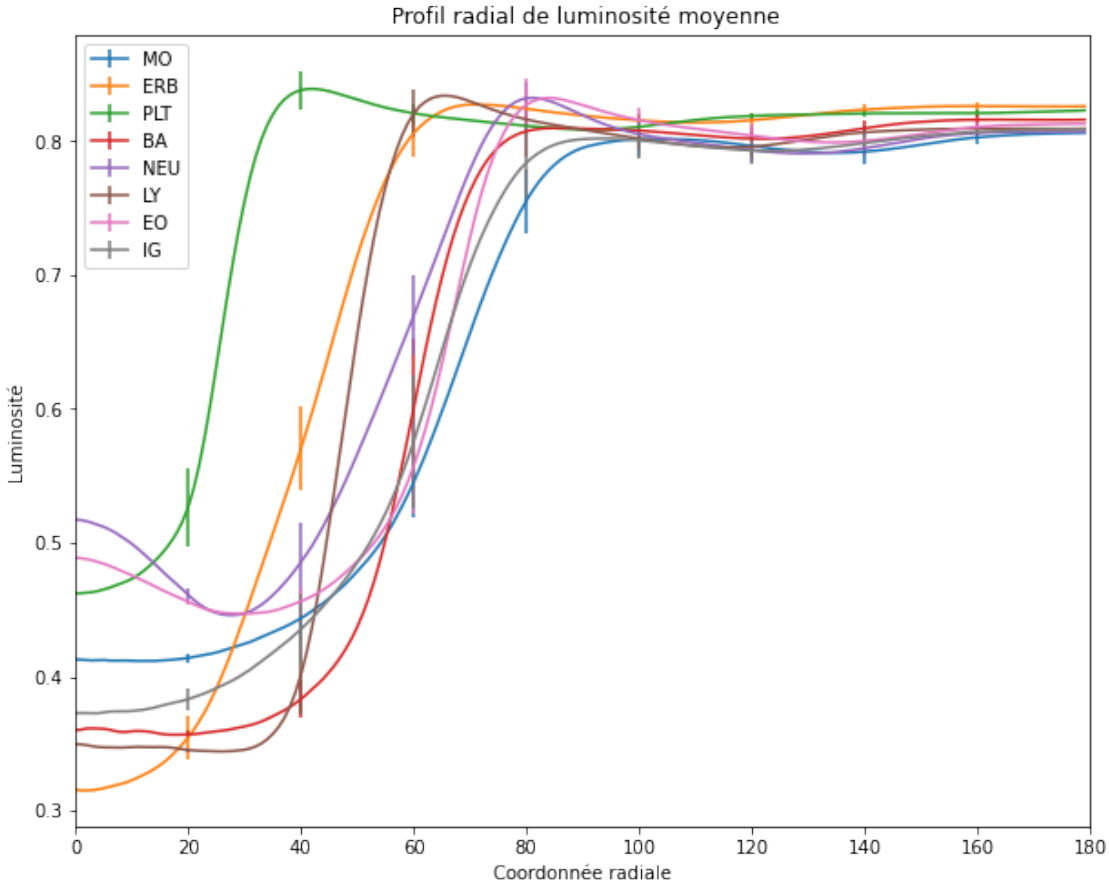

Mean EO : Rouge
Mean EO : Vert
Mean EO : Bleu
Mean EO : RGB
Mean EO : Grey
Red-level histogram
Green-level histogram
Blue-level histogram
RGB histograms
Grey-level histogram



- La cellule moyenne (première ligne de chaque sous ensemble), quelque soit sa catégorie, a une **symétrie cylindrique forte** comme on pouvait l'espérer : une cellule n'étant pas un disque parfait mais plutôt un "patatoïde", on en déduit qu'en moyenne il n'y a pas d'orientation particulière pour un type de cellule par rapport aux autres. La cellule moyenne se trouve bien au centre de l'image, les globules rouges ont été lissés et n'apparaissent presque plus (leur répartition est donc bien aléatoire), de même que les agglomérats qu'on a pu voir dans les outliers.
- Les images moyennes des monocytes (MO) et des basophiles (BA) se ressemblent beaucoup. Les autres cellules se différencient bien les unes des autres, soit par leur diamètre, soit par la structure à l'intérieur de la cellule moyenne.
- On remarque l'existence d'une sorte d'auréole de surbrillance autour de chaque cellule moyenne. Cette auréole correspond à la région de l'espace située au contact direct de la membrane cellulaire : sur la plupart des images du dataset, la cellule d'intérêt n'est pas en contact direct avec les globules rouges, il y a toujours un petit espace. La présence de cet anneau nous permet de remonter à une taille caractéristique, en pixel, de chaque type de cellule.
- Les écarts par rapport à la symétrie cylindrique sont visibles sur la deuxième ligne, représentant l'écart-type de luminosité pour chaque pixel de l'image. On peut déceler, pour la plupart des catégories de cellules, une légère modulation selon la variable angulaire (lancer plot_mean_color_cell).
- La position et la hauteur des pics RGB varie selon le type de cellule, ce à quoi on pouvait s'attendre du fait des disparités entre chaque type de cellule (noyau plus ou moins sombre lorsqu'il y en a un, occupant la totalité ou seulement une partie de la cellule, zones "roses" des éosinophiles etc…)

Ces différences sont relatives à la nature même de chaque type de cellule et apparaissent sur des propriétés moyennes des images : cela constitue un signal positif quant à la capacité d'utiliser un algorithme de machine-learning pour trier ces images.

On peut aller un peu plus loin et tracer le profil radial de luminosité pour chacune des cellules moyennes :



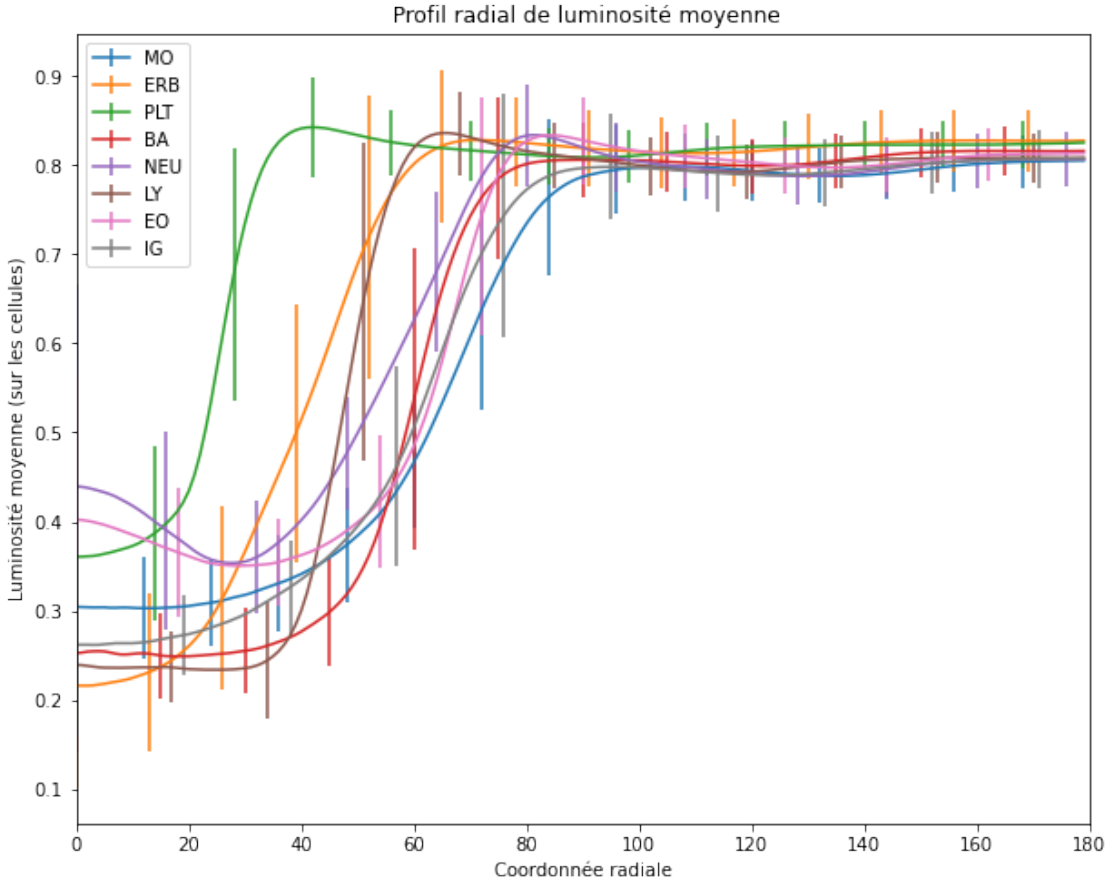
La figure ci-dessus représente l'évolution de la luminance $L(r)$ en fonction de la distance r par rapport au centre de l'image (donc par rapport au centre de la cellule). Cette luminance a été moyennée sur la variable angulaire θ : les barres d'erreurs représentent l'écart-type de $L(r)$ par rapport aux valeurs de θ. Chaque courbe correspond à une des 8 cellules moyennes calculées plus haut.

Deux zones apparaissent distinctement sur chaque courbe : - **zone I :** entre $r = 0$ et $r = r_{max}$ tel que $L(r_{max})$ est maximale (l'anneau de surbrillance autour de chaque cellule moyenne). Elle correspond à l'intérieur de la cellule. La luminosité augmente globalement avec le rayon, jusqu'au bord de la cellule, défini par le pic de luminosité en r_{max}. Un minima local plus ou moins prononcé d'un type de cellule à l'autre est présent sur certaines classes de cellules. Les plaquettes, les érythroblastes et les granulocytes immatures en sont dépourvus. **Le profil de luminosité est assez différent d'un type de cellule à l'autre et les plaquettes sortent clairement du lot,** du fait de leur petite taille par rapport aux autres cellules. **Les deux profils les plus semblables sont ceux des granulocytes immatures et des monocytes.** L'image moyenne est globalement invariante par rotation dans cette zone (les cellules sont globalement circulaires et il n'y a pas d'orientation privilégiée d'une image à l'autre).

- **zone II :** entre r_{max} et $r \simeq 180$: un plateau de luminosité constante, invariante par rotation, d'où la faible importance des barres d'erreur sur chacune des 8 courbes. Cette portion du

profil correspond au fond de l'image, avec les globules rouges. Dans un cadre idéal, il faudrait que les profils des 8 classes se confondent dans cette zone, ce qui n'est pas le cas : malgré tout, les valeurs sont proches d'une cellule moyenne à l'autre (5% au plus fort). Plus important, les **écarts** entre les différentes courbes sont plus importants dans la zone I que dans la zone II, donc **plus importants au centre des images qu'à leur périphérie**, y compris pour lorsqu'on compare les types de cellules dont les zones I sont les plus proches : les différences de profils en zone I entre IG et MO sont les plus faibles, mais leurs profils en zone II se confondent pratiquement.

On peut donc espérer qu'un algorithme de machine-learning privilégiera les différences au centre des images pour les classer, donc qu'il n'y a pas de biais significatif de luminosité d'une classe de cellule sanguine à l'autre.



Cette fois-ci, les barres d'erreurs représentent l'écart-type au sein de la classe de cellules considérée. On voit qu'au niveau du plateau de la zone II, les barres d'erreur d'une courbe recouvrent aussi les autres courbes, ce qui signifie qu'il n'y a pas de différences notables de luminosité entre classes de cellule pour la zone II, on peut donc conclure qu'il n'y a pas de biais de luminosité.

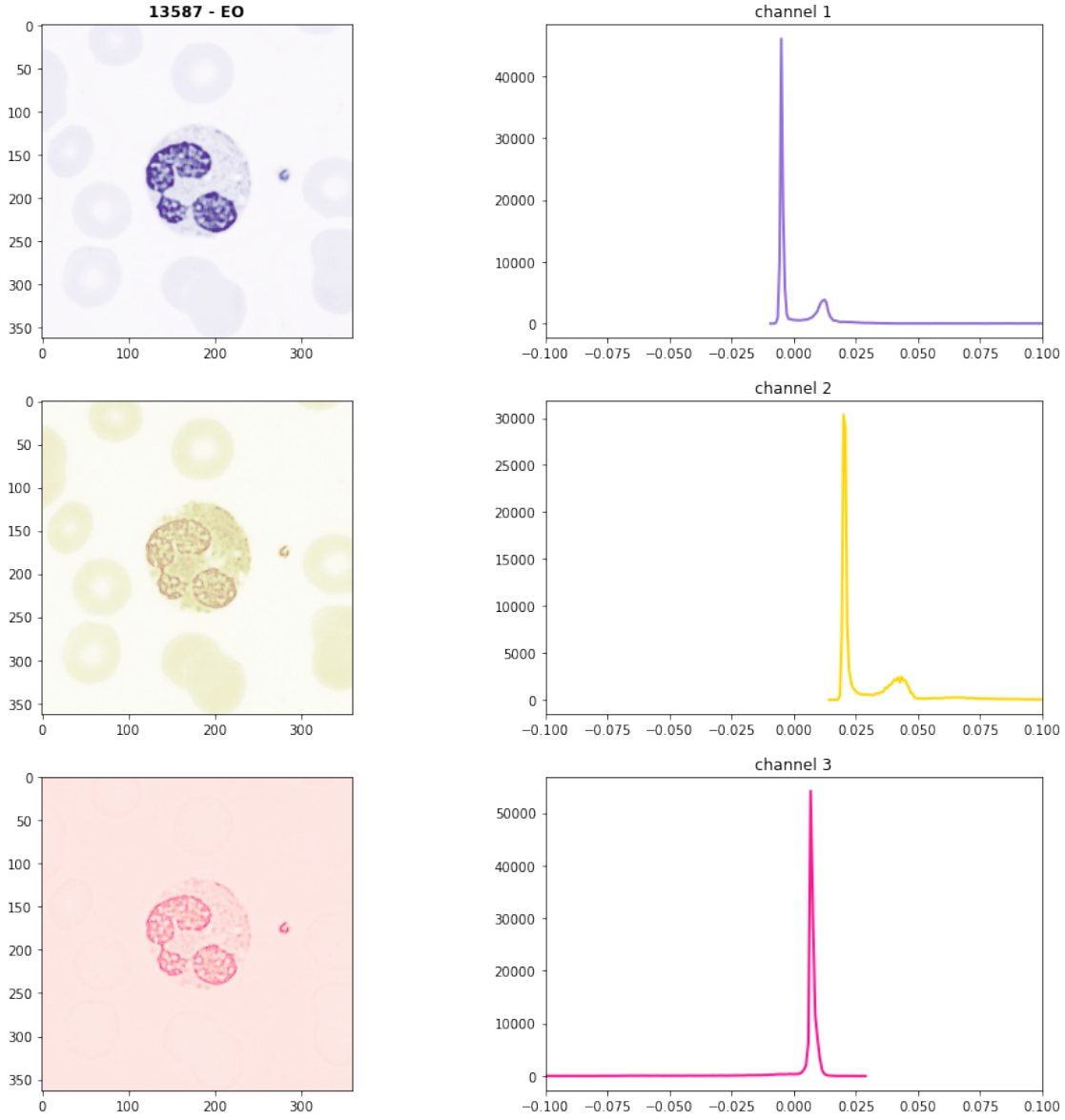
Dans la zone I, les choses sont plus complexes. Pour éviter de détailler au cas par cas, retenons que les disparités d'une classe à l'autre sont plus significatives que pour la zone II, certaines classes étant bien séparées des autres.

Ainsi, les différences de luminosité observées entre les distributions de chaque classe de cellules semblent surtout liées à la taille de la cellule et à celle de son noyau, donc à des caractéristiques qui sont pertinentes pour le classement des images.

6 Déconvolution

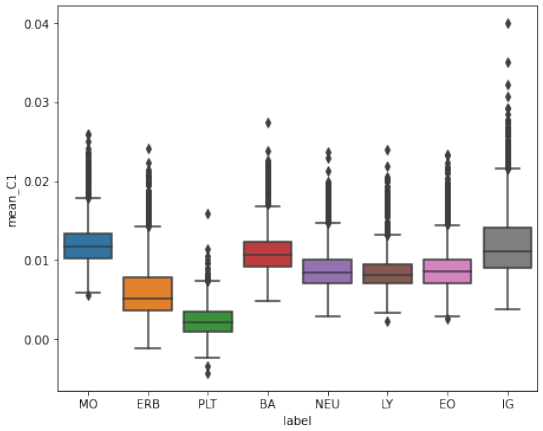
On va utiliser la déconvolution pour décomposer les images sur trois nouveaux canaux de couleur (C1, C2 et C3). La déconvolution des couleurs a été développé pour l'analyse histologique et la librairie utilise l'algorithme développé par G. Landini.
La déconvolution des couleurs est basée sur la loi de Beer-Lambert et peut séparer les images en trois canaux représentant l'absorbance de chaque coloration histologique. Nous avons choisit une matrice de conversion de la coloration `Methyl Blue + Ponceau Fuchsin` qui est proche du MGG original et sépare bien les globules rouges.

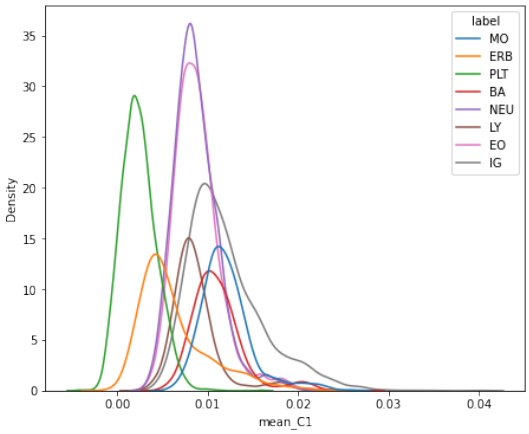
Plot d'un exemple de color deconvolution

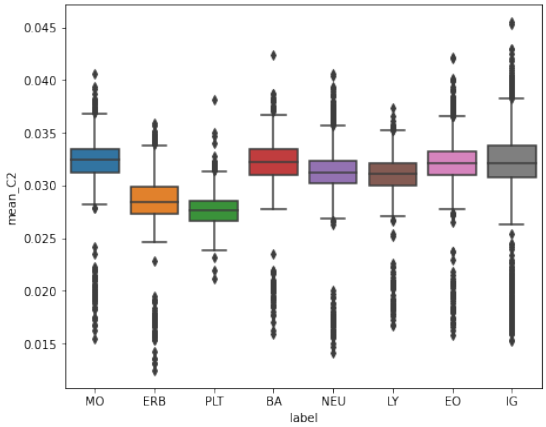


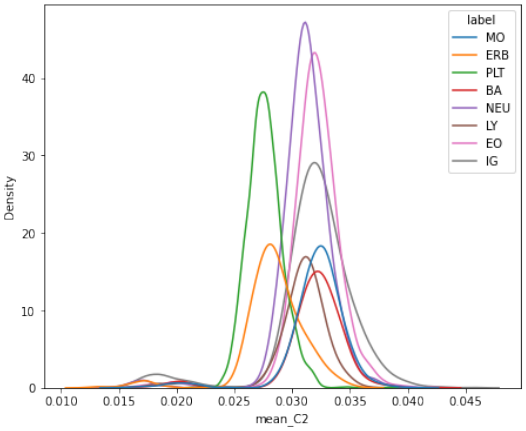
Cette technique permet de séparer les images en trois composantes principales: - le noyau ressort plus sur le canal 1 (en blue) - les globules rouges sont très présents sur le canal 2 en jaune - le fond (bkg) est associé au canal 3 en rose

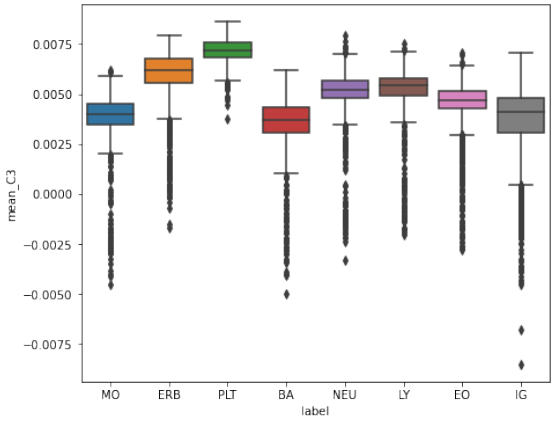
6.0.1 Distribution des expositions moyennes par canaux de déconvolution

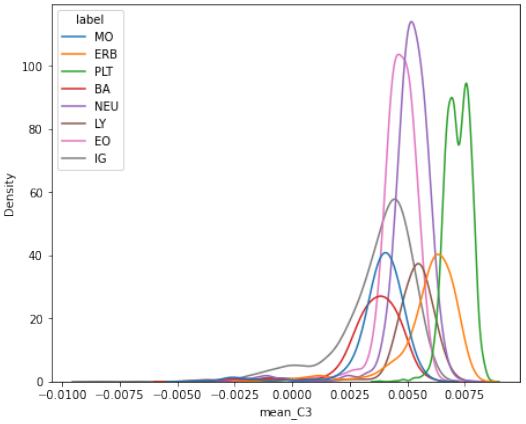






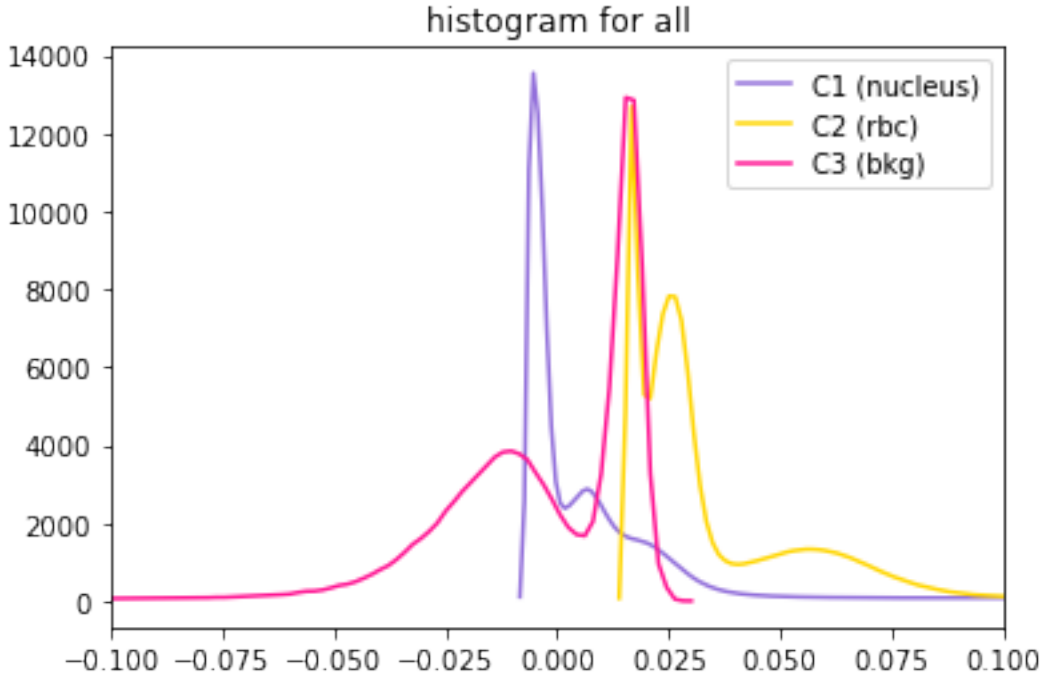


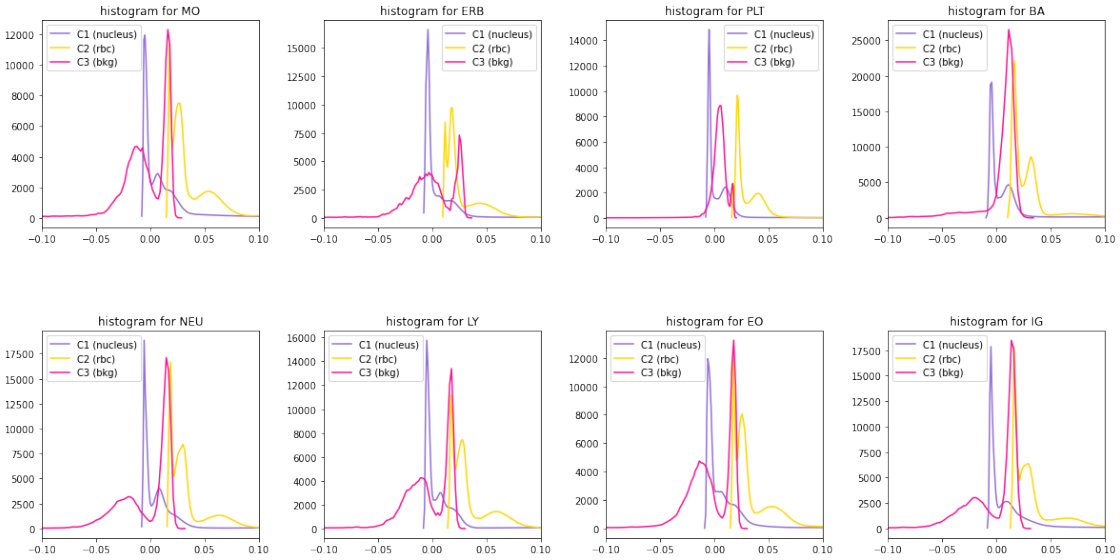




Comme précedemment, les plaquettes (PLT) et les érythroblastes (ERB) présentent une intensité plus faible pour le canal **'noyau'** et plus haute pour le canal **'bkg'**. La différence sur le canal 'rbc-red blood cell' est plus faible, ce qui soutient un nombre similaire de globules rouges moyens en concordance avec l'analyse des cellules moyennes.
La différence de noyau et background peut s'expliquer par la taille des cellules comme précédemment montré par l'analyse radiale. toutes les classes présentent des outliers, essentiellement de faible luminosité.

6.0.2 Distribution des histogrammes d'exposition par type de cellules





Les histogrammes révèlent encore une fois des différences significatives entre les plaquettes (PLT) et les érythroblastes(ERB) d'une part, et d'autre part les autres cellules, confirmant ce qui a pu être dit dans la section précédente. Les basophiles (BA) présentent eux aussi un shift dans la canal background ce qui peux s'expliquer par leurs tailles et la présence d'agglomérats.

7 Conclusion

Compte tenu des différences assez visibles entre les **plaquettes** et les autres cellules (taille plus petite, absence de noyau), on s'attend à ce que le futur modèle puisse les **détecter assez précisément**.

Les données dont nous disposons pour les cellules saines sont de qualité et ne **présentent pas de biais évident**. Il faut toutefois ajouter qu'elles ont été enregistrées avec **le même analyseur, et procéssée dans le même laboratoire avec le même protocole de coloration** : c'est une autre forme de biais, qui pourrait induire un **overfitting** de notre modèle. Lors de la phase de test du modèle et si ce problème apparaît, on pourra essayer d'entraîner le modèle en ajoutant aux données des cellules malades ou provenant d'autres sources.