DataScientest

CLASSIFICATION DES CELLULES SANGUINES

Rapport 1 :

EXPLORATION, DATA VISUALISATION ET PRE-PROCESSING DES DONNÉES

Jessica Planade, Julia Canac, Richard Bonfils, Frédéric Navez

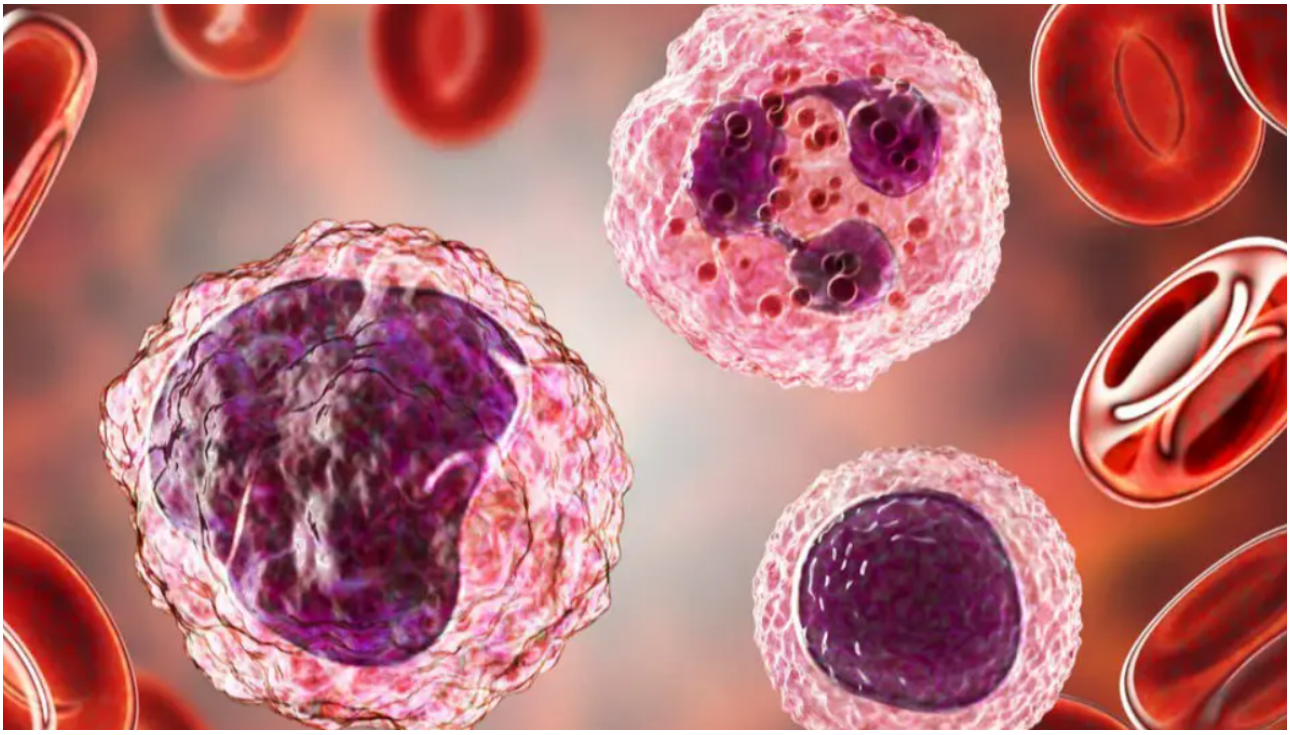


Table des Matières

Pynt of Blood

Introduction au projet

Contexte

Le diagnostic d'un cancer nécessite plusieurs examens dont l'un des plus importants est l'examen histologique. Un prélèvement ou une biopsie est effectuée et une lame est observée par microscopie, ce qui permet de confirmer ou non la présence d'anomalie, identifier les cellules ou cellules touchées et poser définitivement le diagnostic.
Un dépistage réalisé tôt permet de meilleures chances de survie. Cependant, le dépistage est un processus lourd, long et pouvant être sujet aux erreurs. L'intelligence artificielle peut aider à l'analyse d'image. Un modèle bien entraîné peut obtenir des résultats au moins aussi fiable qu'un professionnel de santé et dans un temps réduit, par l'analyse de plusieurs milliers d'images en simultané. Cette aide, la confirmation du diagnostic étant laissée au professionnel, est d'autant plus utile dans le contexte de tension des ressources en personnel hospitalier.

Dans le cas de l'hématologie, qui concerne ce projet, la numération et formule sanguine est réalisée de manière routinière par des automates utilisant les technologies d'impédance électrique et de cytométrie de flux. En cas de résultats anormaux, l'observation de lames est souvent nécessaire pour identifier et / ou confirmer l'anormalité.

Ce projet cible principalement les cancers du sang (leucémies). L'importance du modèle qui sera mis en place va résider, dans un premier temps, dans sa capacité à différencier le type des cellules présentes dans le sang, puis dans un second temps, à déterminer si une cellule est cancéreuse ou non.

Hypothétiquement, en plus de ces deux objectifs, un modèle bien entraîné à reconnaître des cellules saines et malades pourrait mettre en lumière des paramètres encore inconnus ou simplement non encore quantifiés entre caractéristiques morphologiques des cellules et propriétés (type de cellule, état sain ou malade, éventuellement âge?, etc.).

Objectifs

L'objectif primaire est de choisir et d'entraîner un modèle adapté à la distinction des différentes modalités de la catégorie « cellule sanguine » à partir de photographies de bonne qualité (cellules à identifier relativement centrées, image peu/pas bruitée) d'échantillons préparés pour analyse (noyaux cellulaires colorés). On privilégiera le dataset Barcelone pour cela.

L'objectif secondaire serait d'arriver à entraîner le modèle suffisamment bien pour lui permettre de prendre en main d'autres sets de données que les 3 initialement disponibles pour ce projet.

Ce projet vise à développer un modèle de computer vision afin d'identifier les différents types de cellules du sang. Des expériences ont été menées sur un ensemble de 17092 images de cellules sanguines avec leurs sous-types. Un cadre basé sur des algorithmes de Convolutional Neural Network (CNN) sera construit, par la suite, pour classer automatiquement les cellules sanguines. En amont, l'utilisation de modèle de machine learning nous permettra de donner un premier aperçu et d'orienter notre travail vers de l'apprentissage profond.

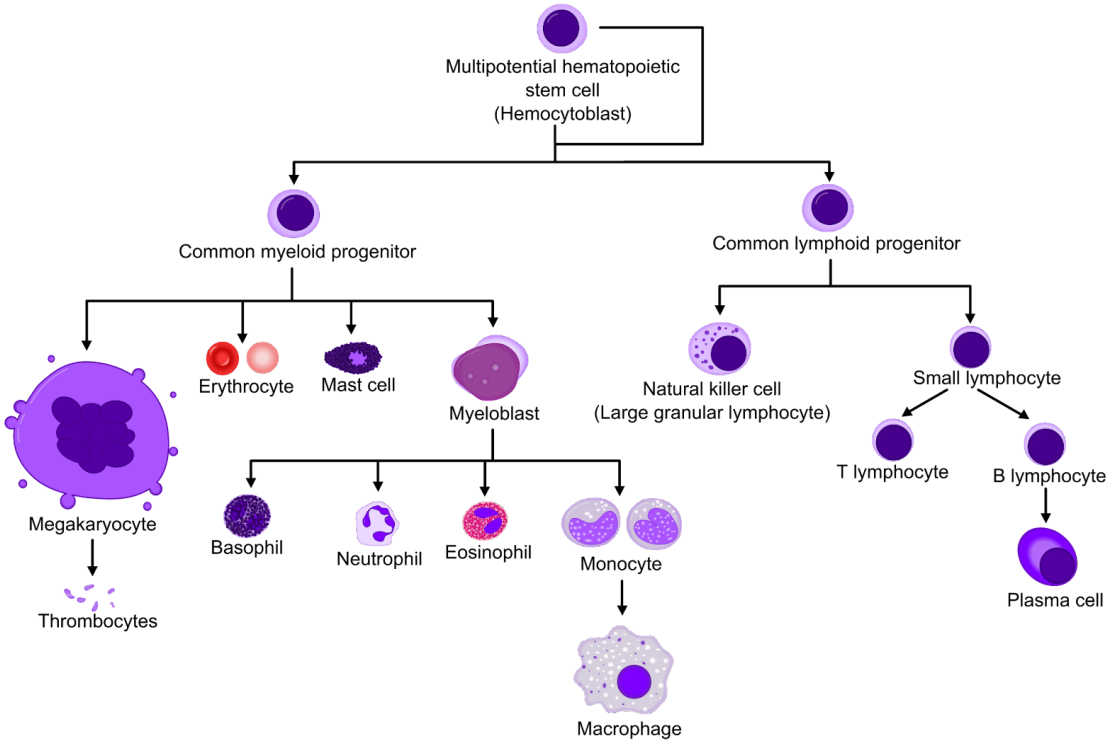
Compréhension et manipulation des données

Cadre

Selon les objectifs primaires et secondaires, des dataset différents vont être utilisés.

1. Notions générales d'hématologie

Il nous paraît utile de résumer les différentes catégories de cellules sanguines existantes et les différentes lignées de cellules :



hematopoièse: *source Myélocyte - Wikipedia (Myélocyte — Wikipédia)*

Les myélocytes sont un stade du développement normal des granulocytes, qui sont un type de globules blancs. Les granulocytes comprennent les neutrophiles, les éosinophiles et les

basophiles, qui sont responsables de la lutte contre les infections et font partie de la réponse immunitaire du corps.

Les myélocytes sont considérés comme le stade du développement des granulocytes qui vient après les promyélocytes, qui sont encore plus tôt dans le processus de développement. Les myélocytes sont caractérisés par la présence de granules spécifiques dans leur cytoplasme, qui contiennent des enzymes et d'autres substances qui aident la cellule à fonctionner.

D'autre part, les granulocytes immatures (mentionnés dans le dataset 1 suivant) sont une catégorie plus large qui comprend non seulement les myélocytes, mais aussi des stades antérieurs du développement des granulocytes, tels que les promyélocytes et les métamyélocytes. Les granulocytes immatures sont généralement identifiés par leur apparence au microscope, où ils peuvent montrer des caractéristiques telles qu'une taille de cellule plus grande, une structure nucléaire moins condensée et des granules moins développés dans le cytoplasme par rapport aux granulocytes matures.

Dans certains contextes cliniques, le terme "granulocytes immatures" peut être utilisé de manière interchangeable avec "myélocytes", mais strictement parlant, les myélocytes sont un stade spécifique du développement des granulocytes, tandis que les granulocytes immatures englobent une gamme plus large de stades, y compris les promyélocytes et les métamyélocytes.

Par ailleurs, nous indiquons ci-dessous les principales caractéristiques morphologiques des différentes catégories de cellules:

- les granulocytes (petits grains dans le cytoplasme, autour du noyau)
    - neutrophils:
        - **couleur: plus pale** que les globules rouges (Red Blood Cells, RBC) qui l'entourent
        - noyau scindé en **3 parties arrondies accrochées les unes aux autres**
        - **2x taille RBC**
    - eosinophil:
        - plus rouge
        - noyau scindé en **2 partie rondes** régulières
    - basophil:
        - **violet très foncé**
        - **2 x plus grand que les RBC**
        - noyau difficilement différenciable du fait de la couleur foncée
- Les agranulocytes (pas de présence de petits grains, diffèrent par leur taille)
    - lymphocytes:
        - violet
        - taille similaire RBC
        - présence d'un halo
    - monocytes:

- noyau en forme de rein
- taille gigantesque, 4-5 x celle des RBC

Bien qu'il soit attendu que les modèles d'apprentissage puissent détecter par eux même ces caractères specifiques (de couleur, aspect du noyau, présence ou non de granules dans le cutoplasme, présence d'un halo autour de la cellule, nombre de composants du noyau, taille relative par rapport aux RBC), nous les gardons en mémoire pour référence et future considération.

2. Premier dataset : 'Barcelona Mendeley data'.

Ce dataset provient d'un article scientifique "*"A dataset for microscopic peripheral blood cell images for development of automatic recognition systems"*. C'est un dossier contenant des images au format .jpg et de taille 363x360 en RGB. Les images, de taille normalisées, sont réparties dans 8 dossiers différents, correspondant aux catégories de cellules suivantes: basophiles, éosinophiles, érythroblastes, granulocytes immatures (IG), lymphocytes, monocytes, plaquettes et neutrophiles.

Les IG (Immature Granulocytes) sont les précurseurs des granulocytes, caractérisées par la présence de granules dans leur cytoplasme et jouant un rôle en cas d'inflammation ou d'infection. Ils contiennent les myélocytes.
Ce dataset permet d'étudier toutes les cellules de la lignée des globules blancs.

Toutes les images sont issues de patients sains (absence de maladies, infections, allergies). Ce dataset va donc permettre l'entraînement d'un modèle à la reconnaissance des cellules du sang

3. Second set : Acute Promyelocytic Leukemia

Ce dataset contient un fichier au format .csv "master.csv" avec 5 colonnes. La première correspond aux identifiants de patient, la seconde au diagnostic, la troisième à la cohorte, la quatrième à l'âge, et la dernière au sexe.
La colonne diagnostic contient uniquement des diagnostics de cancer qualifié de APL pour Acute Promyelocytic Leukemia ou AML pour Acute Myelocytic Leukemia. Ces cancers touchent des cellules pro-myéloïdes ou myéloïdes.

En plus de ce fichier .csv, il est également fourni une banque de 25915 images réparties en 23 catégories :

'Arifact' 'Band neutrophils' 'Basophil' 'Blast, no lineage spec', 'Eosinophils' 'Erythroblast' 'Giant thrombocyte' 'Lymphocyte', 'Lymphocyte, variant' 'Metamyelocyte', 'Monocyte' 'Myelocyte', 'Patient_100', 'Plasma cells', 'Prolymphocyte', 'Promonocyte', 'Promyelocyte', 'Segmented neutrophils', 'Smudge cells', 'Thrombocyte aggregation', 'Unidentified' 'Unsigned slides', 'Young Unidentified'.

Ce dataset va donc permettre d'entraîner un modèle à la détection des cellules saines ou cancéreuses. La catégorie 'Arifact' comprend 26 images considérées comme des artéfacts qu'il pourrait être très intéressant, dans un second temps, d'entraîner le modèle dessus afin de diminuer le nombre de faux positifs.
Un total de 10127 images non annotées (25 "Unidentified", 10100 "Unsigned slides", et 7 "Young Unidentified") pourraient être classées par le modèle une fois celui-ci entraîné de manière satisfaisante sur le reste du set, sauf à être atypiques.

4. **Troisième set: Leukemia dataset (Milan)**

Ce dataset rassemble des images de cancers lymphoblastiques. Ce type de cancer est diagnostiqué par la présence trop importante de lymphocytes. Il est attendu que ce dataset ne comprenne que des images de lymphocytes

Sont disponibles 108 images au format .jpg, associées chacune à un fichier texte au format .xyc reportant les coordonnées des barycentres des blastes. Ces images sont notées XXX_1.jpg ou XXX_0.jpg pour qualifier sur l'image est issu d'un patient malade ou non.

Sont également disponibles 260 images au format .tif qui représentent au centre de chacune des images un lymphocyte cancéreux ou non. Comme pour les images au format .jpg, les patients sains ou malades sont indiqués par respectivement 0 ou 1 à la fin du nom de l'image.

Ce dataset va donc permettre d'entraîner le modèle à la détection des lymphocytes sains ou cancéreux.

Pertinence

Les données des dataset contiennent énormément de variables de part le nombre de pixels RGB présents sur chacune des photos. Cependant, grâce aux étapes de pré-processing, il va être possible de ne sélectionner que les pixels importants dans la reconnaissance des cellules puis dans la reconnaissance des cellules malignes.

Selon l'étape du projet, on peut définir plusieurs variables cibles. Dans un premier temps les variables cibles vont être les types cellulaires présents dans les images. Puis les variables cibles seront les cellules IG malignes ou non et les lymphocytes sains ou cancéreux.

Dans l'optique de mettre au point un modèle performant, il est nécessaire de réaliser chaque étape de pré-processing, de data visualisation, de machine learning et de deep learning de manière consciencieuse. Nous avons donc choisi de nous concentrer sur le data set 1 pour la construction

d'un modèle de reconnaissance des cellules sanguines Nous incorporerons le data set 2 pour la reconnaissance de cellules malades ensuite, si le temps de travail restreint par le format bootcamp le permet. Ainsi, le reste du document va être axé sur le dataset 1.

Dans ce dataset, certaines catégories comportent beaucoup plus d'images que d'autres : il y aura sans doute un biais lors de l'apprentissage, à cause de la différence des tailles d'échantillons.

Enfin, d'un point de vue technique, nous semblons nous confronter aux limites de Google Colab gratuit avec un lag important lorsque plusieurs membres du groupe collaborent vs posent simplement leur dernières versions de notebooks.

Pre-processing

Le premier set de données sur lequel nous avons choisi de travailler est très propre : pas de doublon, pas d'images non annotées, cellules marquées par coloration dans les échantillons avant photographie. Mais les images sont nombreuses et lourdes (360x360x3 pixels en général). De plus, sur certaines images plusieurs cellules, de type différent qui plus est, sont visibles. Enfin, une image est endommagée et doit être éliminée du set de données.

Quelques étapes de pré-processing sont donc tout de même nécessaires :

1. Lors de la mise en commun des données de travail, nous avons constaté que des problèmes d'upload des données avaient pu générer la création de copies de certaines images. Nous avons donc ajouté à notre code des cellules de nettoyage visant à supprimer les copies « ordinateurs » porteuses d'un numéro entre parenthèses dans le nom de fichier. Cette étape est une sécurité et ne gaspille pas trop de temps de traitement.
2. L'image problématique est contenue dans le dossier "neutrophil", donc son élimination se fait dans le code à l'aide de la méthode pop grâce à une vérification automatique du type de fichier contenu dans ce dossier spécifique. Une élimination des fichiers non lisibles lors du chargement des images sur tous les dossiers chargés pourra facilement être mise en place en généralisant, lorsque nous voudrons incorporer de nouveaux sets, au cas où le set de données entrées comporterait plus de fichiers corrompus, mais pour l'instant le code (plus rapide) disponible convient.
3. Des dictionnaires contenant chacun la liste des noms des images ont été créés, de même qu'un dataframe de 17092 lignes contenant 6 colonnes : le nom des images, le code de type cellulaire, la famille cellulaire, la hauteur de l'image, la largeur de l'image, le code de la sous-catégorie de type cellulaire (permettant de re-découper certaines familles de cellules).

Ci-dessous une capture d'écran de l'affichage des dernières lignes de ce tableau :

```
df_barc.tail()
```

	image	code_g	cell_type	height	width	code_spe
17087	SNE_995183.jpg	NEU	neutrophil	363	360	SNE
17088	SNE_99568.jpg	NEU	neutrophil	363	360	SNE
17089	SNE_995695.jpg	NEU	neutrophil	363	360	SNE
17090	SNE_995874.jpg	NEU	neutrophil	363	360	SNE
17091	SNE_999519.jpg	NEU	neutrophil	363	360	SNE

4. Un dictionnaire contenant une image moyenne, par catégorie, de toutes les autres, a ensuite été généré. Pour se faire il a fallu redimensionner certaines images qui étaient de taille différente des autres : la méthode resize de cv2 a été utilisée.

```
df_barc[df_barc.height>363]
```

	image	code_g	cell_type	height	width	code_spe
30	BA_127671.jpg	BA	basophil	369	366	BA
32	BA_128084.jpg	BA	basophil	369	366	BA
70	BA_162483.jpg	BA	basophil	369	366	BA
74	BA_165215.jpg	BA	basophil	369	366	BA
91	BA_178345.jpg	BA	basophil	369	366	BA
...	...	...	...	...	...	...
1678	NEUTROPHIL_941537.jpg	NEU	neutrophil	369	366	NEUTROPHIL
1679	NEUTROPHIL_971896.jpg	NEU	neutrophil	369	366	NEUTROPHIL
1680	NEUTROPHIL_982898.jpg	NEU	neutrophil	369	366	NEUTROPHIL
1681	NEUTROPHIL_985581.jpg	NEU	neutrophil	369	366	NEUTROPHIL
1682	NEUTROPHIL_993818.jpg	NEU	neutrophil	369	366	NEUTROPHIL

250 rows × 6 columns

5. Une étape de redécoupage des images sur la base d'une détection des cellules et d'une découpe plus proche pourrait permettre de diminuer la taille des fichiers à manipuler tout en éliminant certains artefacts. Pour le moment nous n'avons pas implémenté cette idée mais une première étape de seuillage sur des images type de chaque catégorie a dores et déjà permis de mesurer la taille moyenne de la surface projetée des cellules, en pixel, de chaque type cellulaire. Il est ainsi possible d'extraire un paramètre de nouvelle taille d'image par catégorie que l'on pourrait appliquer au données pour diminuer la taille des tableaux à utiliser et améliorer la qualité de l'apprentissage. L'étape de détection des objets serait encore nécessaire afin de pouvoir positionner correctement le cadre de re-découpe

sur chaque image. Il n'est pas certain que nous aurons le temps de mettre en place cette étape.

6. Une analyse en composantes principales est en cours de réalisation sur l'ensemble des images du dataset afin de générer un fichier de travail plus léger contenant des features importantes pour les étapes d'entraînement (réduction des données). Il sera créé en fonction des résultats des dataframes au format csv pour chacune des catégories d'image.
7. Pour finir, un premier set de dataframes de travail a été réalisé. Il s'agit de 8 dataframes sauvegardés dans 8 fichiers au format .csv, contenant chacun une ligne de pixel par image ainsi que, dans la dernière colonne, le code de la catégorie de cellule annotée (de manière à pouvoir concaténer les dataframes si besoin). Chaque image a été redimensionnée à 224 x 224 x 3 px. Il sera éventuellement remplacé ultérieurement par un set contenant les données réduite suivant un modèle intelligent.

Nous envisageons également de rééquilibrer le nombre d'images par catégorie de cellule au moment de l'entraînement du modèle, soit par suppression d'un certain nombre d'images dans les catégories sur-représentées, soit en jouant sur le poids de chaque image dans le modèle. Le cas échéant, les dataframes de travail pourront donc être redéfinies.

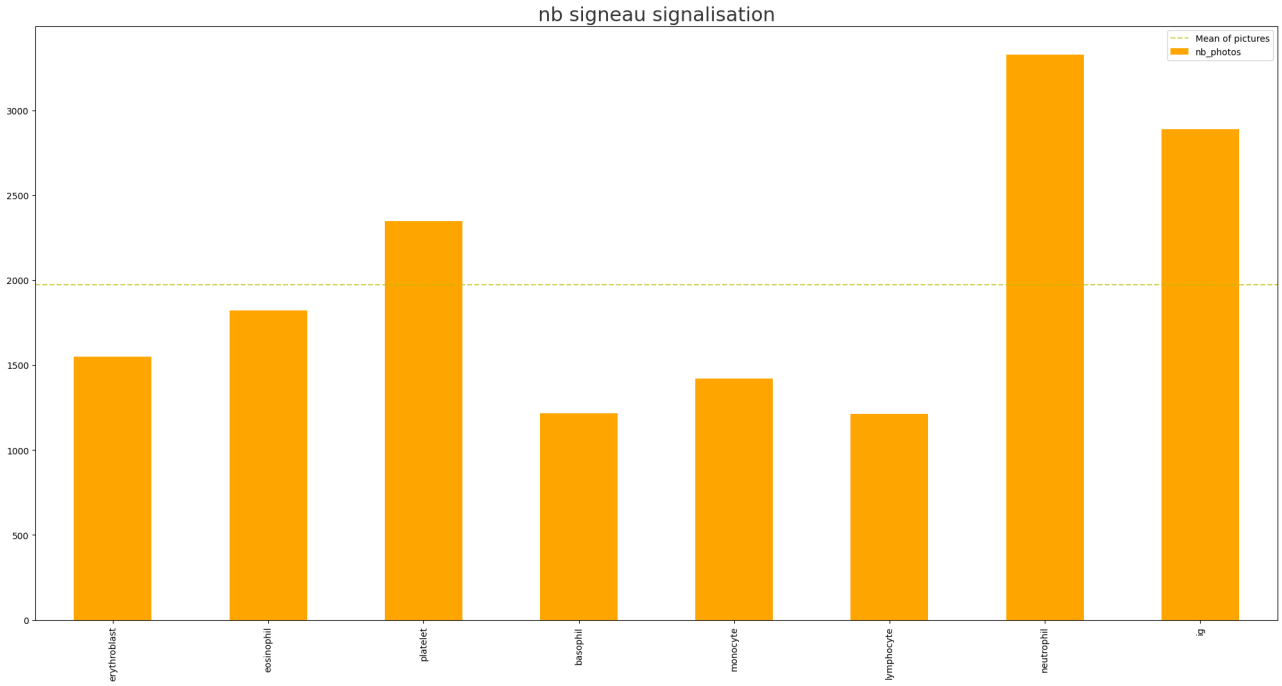
Visualisations et Statistiques

Le dataset Mendeley_Barcelona contient 8 classes de cellules saines:

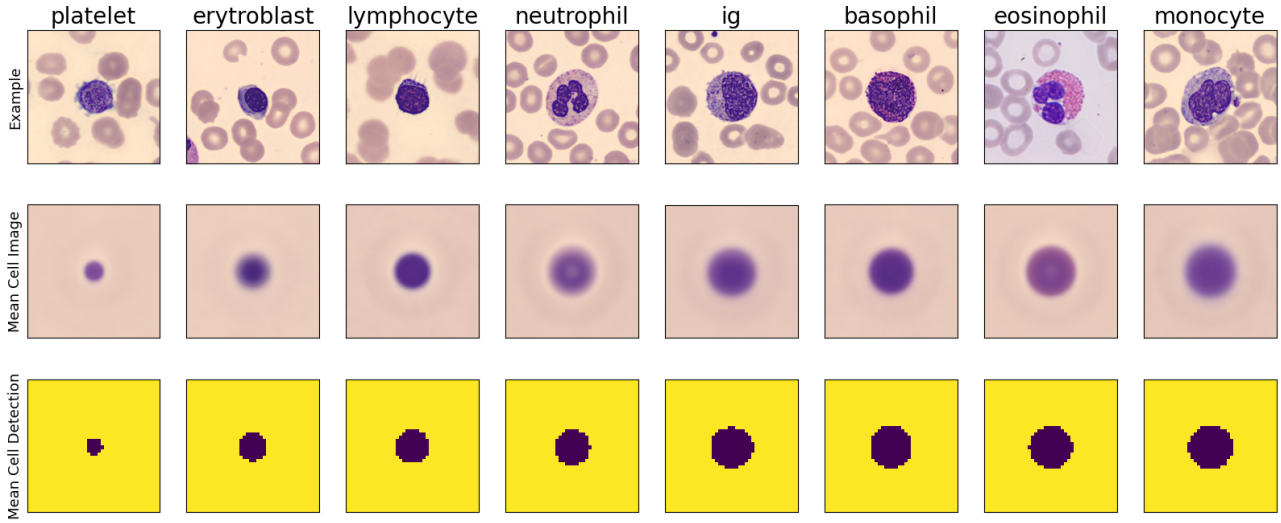
- Basophile (BA),
- Eosinophile (EO),
- Erythrophile (ER),
- Immature Granulocyte (IG),
- Lymphocyte (LM),
- Monocyte (MON),
- Plaquette (PLA)
- Neutrophile (NEU).

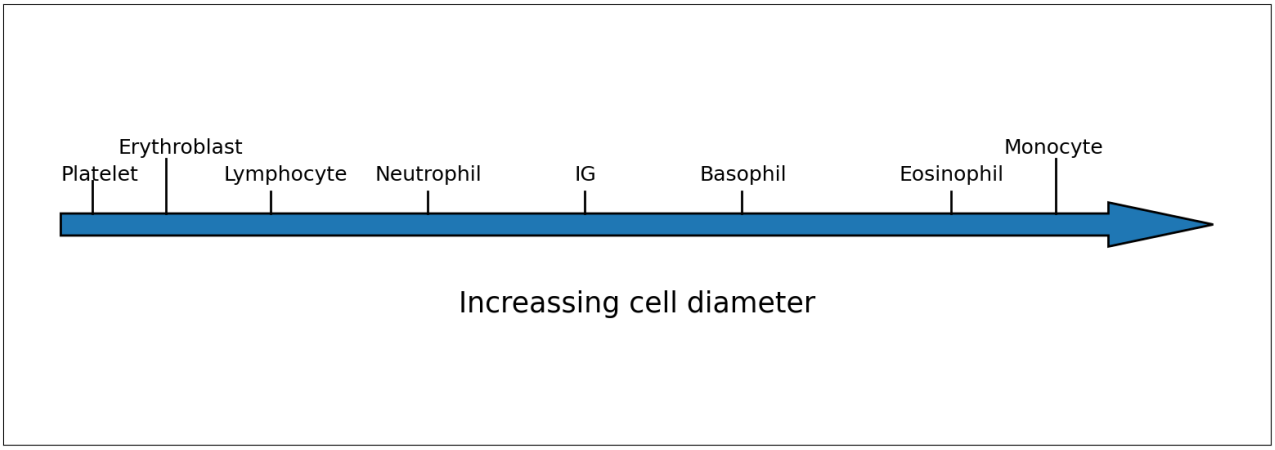
Au total, il y a 17092 images de cellule avec 3330 images de NEU, 1823 images d'EO, 2890 images d'IG, 2348 d'images de PLA, 1551 d'images d'ER, 1420 d'images de MON, 1218 d'images de BA et 1214 d'images de LYM.
En moyenne, il y a par classes 1974 images avec un écart-type de 747..

Comme dit précédemment, il y a un déséquilibre des classes. Le graphique ci-dessous illustre ce phénomène.



Concernant les cellules, selon le type cellulaire nous allons observer des formes et des tailles de cellules et noyaux différents. A partir des images moyennées, il a été calculé la taille des cellules en fonction du nombre de pixels. Ci dessous, sont présentes trois figures, la première représente une cellule de chaque catégorie prise au hasard, la seconde représente les cellules “moyennées” et la dernière est une frise qui classe les cellules en fonction de leur taille.

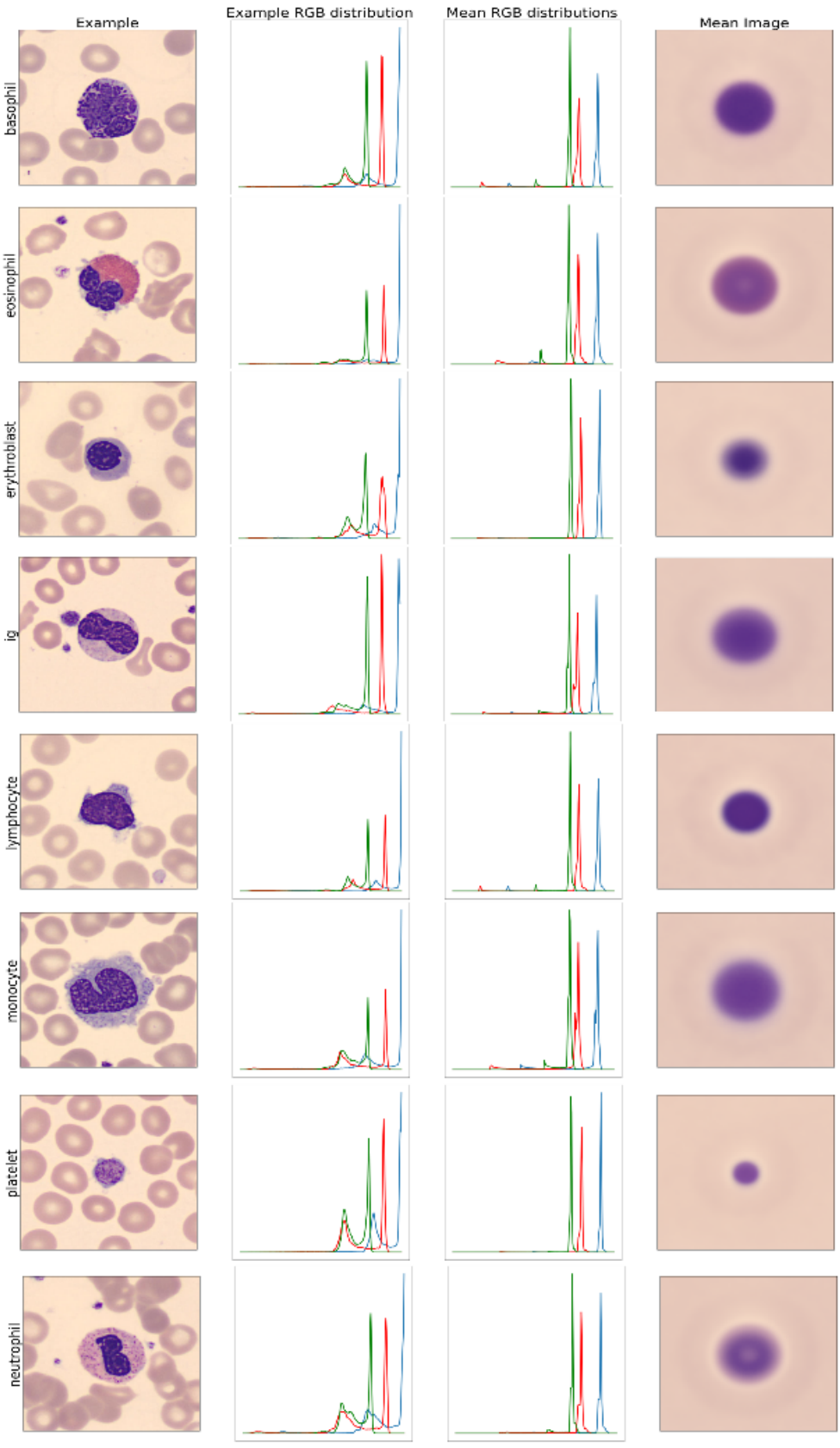




Enfin, il a aussi été réalisé une analyse sur la répartition RGB des images, premièrement sur les images brutes, en prenant une cellule par type. Une deuxième analyse a ensuite été réalisée sur les images “moyennées”.

A partir de ces graphiques on peut voir que selon le type cellulaire on a une répartition différente des couleurs RGB (en terme de hauteurs de pics, séquence mais aussi parfois distance entre les pics) avec parfois des sous-pics au sein d’une même couleur. Sur certains graphiques, notamment sur ceux des IG et éosinophile, on peut voir qu’il y a une sorte de bruit de fond qui pourrait être dû au fond de chaque image.

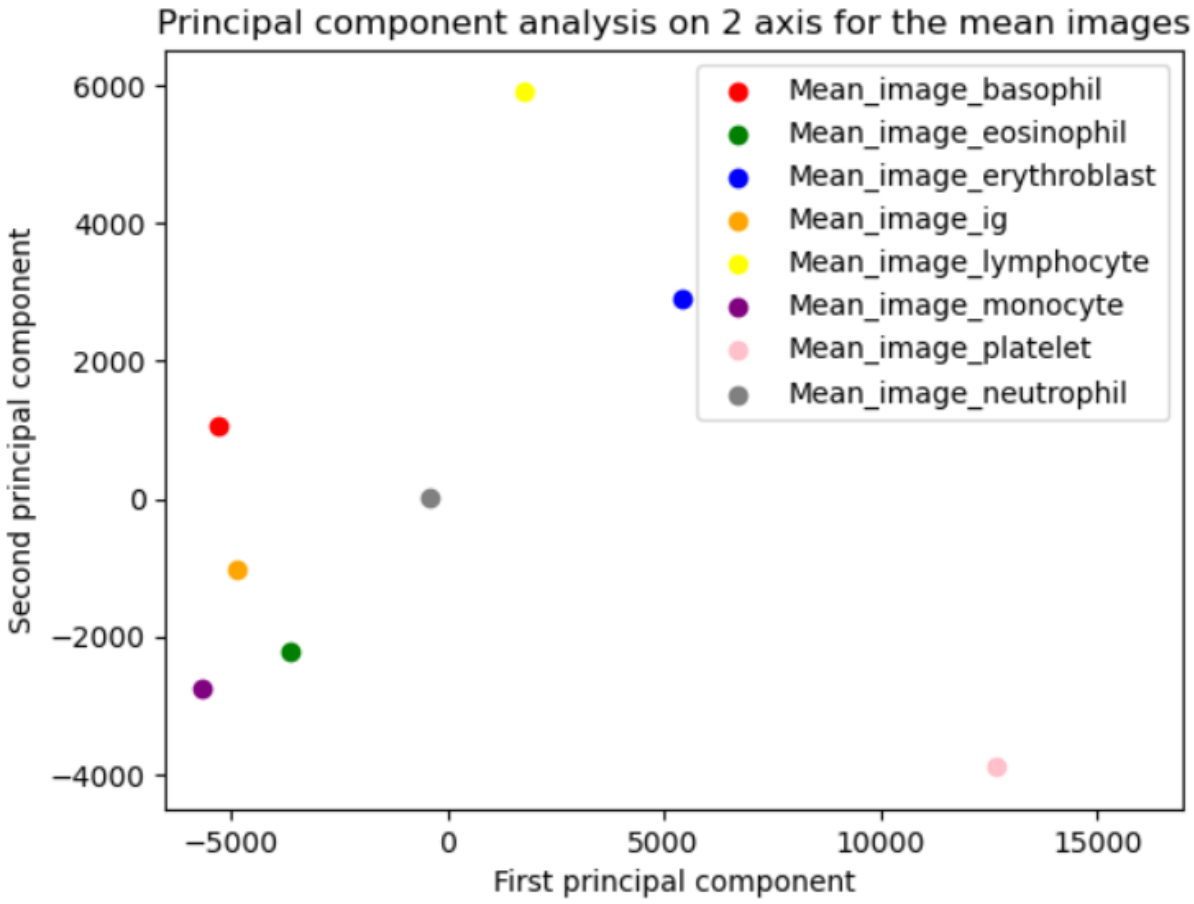
Des différences subsistent sur les images moyennées entre certains types cellulaires.


Example
Example RGB distribution
Mean RGB distributions
Mean Image
basophil
eosinophil
erythroblast
ig
lymphocyte
monocyte
platelet
neutrophil

Analyse en composantes principales

L'analyse en composantes principales (ACP) est une méthode de réduction de dimension qui transforme des variables corrélées en variables décorrélées. Il s'agit de résumer l'information contenue dans une large base de données en un certain nombre de variables synthétiques appelées composantes principales.

Cette représentation a été réalisée sur la moyenne de chaque classe de cellules sanguines. Elle est satisfaisante puisque chacune des classes est bien séparée.



Une analyse plus fine sera réalisée dans la suite du projet par l'utilisation d'autres algorithmes comme l'Isomap ou la tSNE.