



"Optimisation de l'Imagerie Médicale par IA :

Approches Avancées pour la Classification des Cellules Sanguines".

(projet bcds_blood_cells)

Réalisé par :

_Kipédène COULIBALY,

_Elie B. DURAN,

_Joseph LIEBER,

_Sergey SASNOUSKI

Sous la direction de Aziz AGREBI

Remis le 13 septembre 2023

Juin 2023 Bootcamp Data Scientist

Table des matières

Résumé	4
Introduction.....	5
- Présentation du contexte et de l'importance de la classification des cellules sanguines	5
- Objectifs de la modélisation	5
Collecte des données.....	7
- Sources des données et méthodes de collecte	7
- Description des types de cellules sanguines à classifier	7
- Description des types de cellules sanguines à classifier	9
- Hématopoïèse	10
- Rôle de chaque type de cellule dans le sang.....	11
Exploration des données	13
- Exploration du jeu de données.....	13
Observation initiale des images.....	13
Standardisation de la taille des images	13
Distribution des classes	13
Analyse en coordonnées polaires.....	14
Prétraitement des données : mise en œuvre du modèle U-Net	18
- Vue d'ensemble de la chaîne de traitement.....	18
- Nettoyage préliminaire du jeu de données.....	19
- Enjeux de la segmentation	19
- Raisons du choix d'architecture U-Net.....	19
- Structure du modèle U-Net.....	19
- Harmonisation de la taille des images	20
- Segmentation manuelle	20
- Augmentation des images.....	21
- Répartition des données en ensembles d'entraînement, de validation et de test	21
- Choix des hyperparamètres	22
- Métriques utilisées pour évaluer le modèle	22
- Entraînement du modèle et techniques pour éviter le surapprentissage	23
- Performance du modèle sur l'échantillon de test	23
- Interprétation des résultats et identification des domaines de réussite	24

- Limitations du modèle et perspectives d'amélioration	25
- Nettoyage complémentaire des images	25
Modèle de classification : ResNet-34	27
- Enjeux de la classification	27
- Raisons du choix de ResNet-34	27
- Architecture du modèle ResNet-34	27
- Répartition des données en ensembles d'entraînement, de validation et de test	28
- Choix des hyperparamètres du modèle	28
- Métriques utilisées pour évaluer le modèle	29
- Entraînement du modèle et techniques pour éviter le surapprentissage	29
- Performances du modèle sur l'échantillon 'Golden Test'	30
Modèle à 8 classes.....	30
Modèle à 11 classes.....	31
- Interprétation des résultats et identification des domaines de réussite	33
- Suggestions d'améliorations et de travaux futurs	33
Conclusion	35
- Importance de ce travail dans le contexte médical et clinique	35
- Potentiel d'utilisation du modèle dans des applications réelles	35
Table des illustrations	36
Bibliographie.....	38
Annexes	39
- Signature radiale de chaque catégorie de cellule	39
- Matrice de confusion de la classification basée sur la signature radiale des cellules	43
- Résultat de l'estimation de la segmentation U-Net en 9 classes.....	44
- Exemple de cellules mal classées par le modèle Resnet-34 à 11 classes	44

Résumé

L'avènement de l'imagerie médicale moderne a ouvert des portes dans le diagnostic et la surveillance des maladies. Particulièrement, la classification des cellules sanguines est vitale pour le diagnostic de diverses maladies hématologiques, comme la leucémie. Dans ce projet, nous explorons la puissance combinée de l'imagerie médicale et de l'intelligence artificielle pour discriminer onze catégories de cellules sanguines. Pour mener à bien cette mission, nous avons utilisé un jeu d'images produites par le Laboratoire central de l'Hôpital Clinic de Barcelone, représentant un ensemble de 17 092 images de cellules sanguines de patients sains.

Un prétraitement rigoureux est essentiel pour éviter le bruit et les artefacts d'une image médicale. Notre équipe a veillé à éliminer les anomalies comme les clichés présentant plusieurs cellules ou des débris cellulaires et à masquer le fond de l'image pour éliminer les éléments non pertinents. Cela a été fait grâce à un modèle U-net pré-entraîné sur un échantillon de 400 images pour lesquelles nous avons réalisés des masques manuellement, échantillon augmenté de 2800 images par rotation et effet miroir. Nos résultats sont prometteurs : l'architecture U-Net, avec son mécanisme d'encodeur et de décodeur, a enregistré une IoU¹ moyenne de 0.98. L'application du modèle U-Net a été complétée par la mise en œuvre d'une procédure de nettoyage automatisée écartant les cas aberrants, notamment les images représentant plusieurs cellules. Plusieurs stratégies d'amélioration sont envisageables pour cette étape : l'augmentation des données, l'exploration d'autres architectures et une collaboration accrue avec des experts en hématologie.

On a ensuite entraîné un modèle de classification ResNet34, qui permet d'atteindre une précision de 98% (pour un modèle à 8 classes) et 90% (pour un modèle à 11 classes). Une analyse détaillée a toutefois révélé des performances inégales entre les classes. Une série d'expérimentations, y compris l'ajustement des hyperparamètres et la comparaison avec d'autres architectures constituent des pistes d'amélioration. De façon plus spéculative, nous nous sommes interrogés sur l'intérêt d'une transformation des images de cellules en coordonnées polaires comme étape préalable à la modélisation.

Les erreurs de classification interviennent parfois entre des catégories de cellules biologiquement très proches, par exemples des cellules à différent stade de leur développement. Cela permet de s'interroger sur la tangibilité d'une frontière nette entre les catégories de cellules et donc la pertinence de l'étiquetage. On chercherait en vain à améliorer le modèle si l'ensemble des erreurs de classification s'expliquaient par des imprécisions d'étiquetage. Nous suggérons qu'une comparaison d'étiquetage des mêmes images, par plusieurs professionnels opérant de façon indépendante, serait ici pertinente.

Nos efforts soulignent l'énorme potentiel des techniques d'apprentissage automatique en matière d'imagerie médicale. Une suite à donner à ce travail serait d'utiliser la catégorisation des cellules sanguines réalisée comme une aide au diagnostic lors de l'analyse des extraits sanguins. Même s'il reste du chemin à parcourir, nous espérons que notre système pourra contribuer aux avancées dans le domaine de l'imagerie biomédicale.

¹ IoU : Intersection over Union. Pour un segment donné (ici l'intérieur d'une cellule, en opposition au fond de l'image) est définie comme le rapport entre les pixels dans l'intersection entre le segment prédit et la segment réel et l'union des pixels qui appartiennent soit au segment prédit soit au segment réel.

Introduction

- Présentation du contexte et de l'importance de la classification des cellules sanguines

Les progrès technologiques du XXI^e siècle ont radicalement changé le paysage médical, ouvrant la porte à des innovations en matière de diagnostic, de traitement et de recherche clinique. L'un des aspects les plus prometteurs de cette révolution est l'utilisation des données médicales, notamment l'imagerie médicale, comme source inestimable pour la recherche et l'innovation. Ces données, principalement visuelles, permettent non seulement d'affiner les diagnostics, mais aussi d'ouvrir la voie à la découverte de nouvelles thérapies.

Néanmoins, la mise en œuvre de ces avancées technologiques est loin d'être simple. Des études récentes ont souligné un enjeu majeur dans le domaine de l'IA et du big data : une grande partie de l'effort est consacrée à la préparation des données. Ce processus méticuleux est essentiel pour garantir la qualité, la confidentialité et la conformité des données, tout en respectant les normes éthiques et réglementaires. Ainsi, bien que l'IA offre un potentiel immense, la préparation des données peut devenir un frein à l'innovation.

Face à ces défis, la numérisation des données médicales et la nécessité d'une infrastructure robuste pour gérer d'importants volumes de données sont devenues primordiales. La transformation numérique vise à augmenter l'efficacité opérationnelle, à réduire les coûts et à promouvoir l'innovation. Dans les sciences de la vie, l'imagerie médicale s'est imposée comme un élément clé, alimentant des initiatives visant à accélérer la découverte de nouveaux traitements et dispositifs.

Au sein de cette vaste étendue de l'imagerie médicale se trouve la classification des cellules sanguines, un domaine qui, bien que technique, possède un intérêt sociétal majeur. Notre projet, à visée pédagogique, se penche sur ce secteur. Pour les contributeurs de ce projet, il s'agit non seulement d'une opportunité d'apprentissage, mais également d'une première étape vers l'accès à cette branche prometteuse de l'IA.

- Objectifs de la modélisation

L'ambition de ce projet est d'exploiter les techniques avancées du *Deep Learning* pour reconnaître divers types de cellules sanguines. La configuration quantitative des cellules sanguines dans un frottis joue un rôle primordial dans le diagnostic de pathologies majeures, la leucémie étant un exemple saillant où le diagnostic se concentre sur le rapport de lymphocytes. Dans le contexte de ce projet, étant donné le temps imparti, nos efforts ont été ciblés sur l'optimisation de la reconnaissance de onze types spécifiques de cellules sanguines. Cette focalisation découle de l'exploitation d'un jeu de données conséquent, comportant 17 092 images provenant de patients indemnes de toute affection. Concevoir un instrument capable de déchiffrer les caractéristiques des cellules à partir de frottis sanguins ne se contente pas d'augmenter le potentiel diagnostique, mais présente également un intérêt pour les applications de recherche avancée.

Le présent rapport de modélisation s'articulera autour de plusieurs axes essentiels. Nous débuterons par une immersion dans nos méthodes de collecte de données, en mettant en avant la typologie des cellules sanguines étudiées. Cette exploration sera suivie d'une analyse détaillée des étapes de prétraitement des données, englobant la normalisation, l'augmentation des images et mise en œuvre du modèle U-Net permettant de masquer le fond des images. Puis, nous présenterons le modèle de classification, justifierons notre choix d'architecture de modèle (ResNet-34) et celui des hyperparamètres utilisés pour entraîner le modèle. En conclusion, nous effectuerons une évaluation rigoureuse du modèle, s'appuyant sur des métriques adaptées pour déceler et interpréter ses performances.

Collecte des données

- Sources des données et méthodes de collecte

Notre jeu de données regroupe 17 092 images de cellules colorées selon le protocole May Grünwald-Giemsa et obtenues grâce à l'analyseur CellaVision DM96 au Laboratoire central de l'Hôpital Clinic de Barcelone. Elles proviennent d'individus sains, sans infections ou pathologies hématologiques/oncologiques et non sous traitement pharmacologique lors de la collecte.

Chaque image représente, en principe, une cellule unique dont la catégorie a été annotée par des pathologistes cliniques experts : neutrophiles (à noyaux en bâtonnets et noyaux segmentés), éosinophiles, basophiles, lymphocytes, monocytes, granulocytes immatures (promyélocytes, myélocytes, métamyélocytes), thrombocytes, érythroblastes et plaquettes.

Ce jeu de données étiqueté est destiné à entraîner et tester des modèles d'apprentissage supervisés visant à prédire la catégorie d'une cellule sanguine à partir de son image. C'est un ensemble de cette envergure unique pour les cellules sanguines périphériques normales, se positionnant comme une référence pour l'évaluation de modèles.

Les images, en principe d'une taille de 360 x 363 pixels, sont fournies au format jpg.

- Description des types de cellules sanguines à classifier

Neutrophiles: De forme généralement sphérique, les neutrophiles possèdent un noyau segmenté, souvent comparé à une perle enfilée ou à des formes lobées connectées. Le cytoplasme contient de petits granules neutrophiles qui sont difficiles à distinguer à faible grossissement.

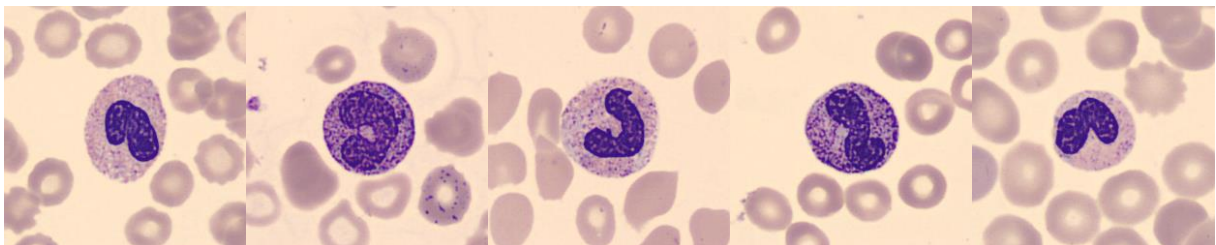


Figure 1 : Neutrophiles: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone

Éosinophiles: Tout comme les neutrophiles, les éosinophiles sont généralement sphériques. Ils se distinguent par un noyau bilobé et un cytoplasme contenant de grands granules, souvent de forme ovale, qui prennent une teinte rosâtre à l'éosine.

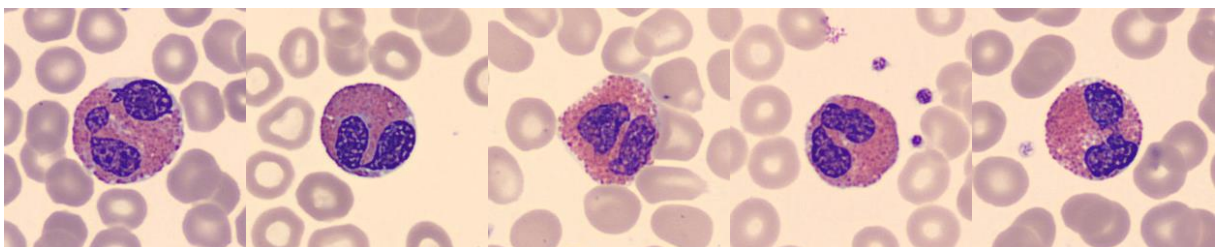


Figure 2 : Eosinophiles: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone

Basophiles: Ces cellules, également de forme sphérique, ont un noyau lobé qui est souvent masqué par les granules cytoplasmiques. Ces granules, de forme irrégulière, prennent une couleur pourpre foncé ou bleuâtre lorsqu'ils sont teints.

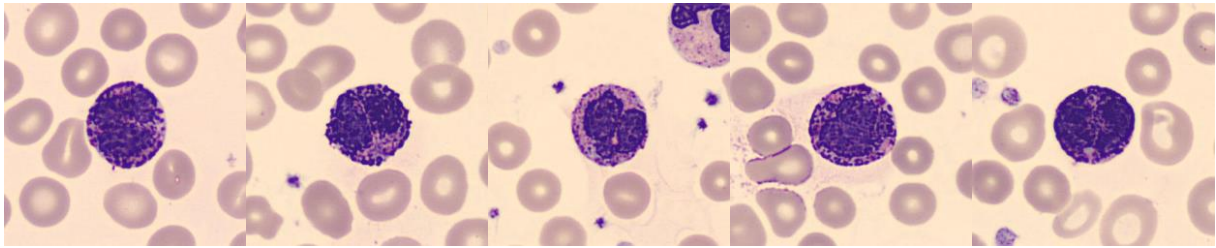


Figure 3 : Basophiles: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone

Lymphocytes: Les lymphocytes sont caractérisés par un grand noyau rond et un cytoplasme réduit, faisant penser à une fine bordure autour du noyau. Leur forme peut être comparée à un disque.

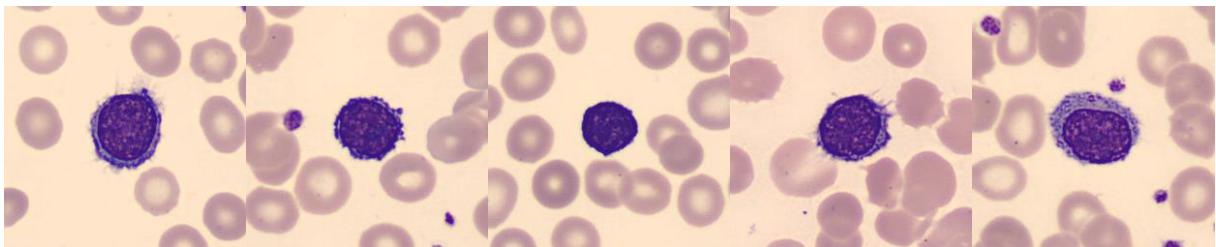


Figure 4 : Lymphocytes: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone

Monocytes: Ces cellules ont une forme ovale à elliptique avec un noyau en forme de fer à cheval ou de rein. Leur cytoplasme est abondant avec un aspect vacuolé et une couleur grisâtre.

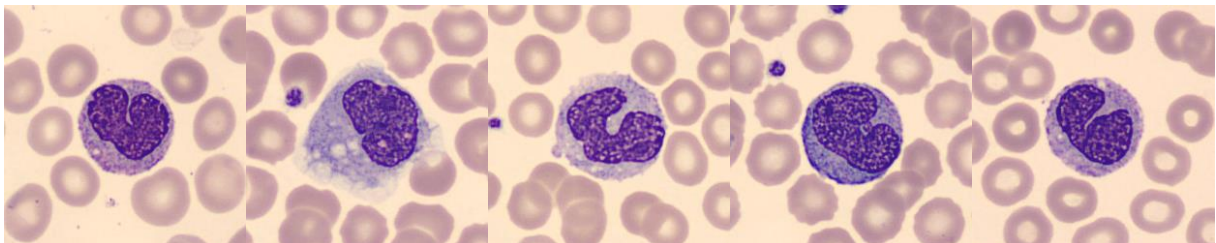


Figure 5 : Monocytes: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone

Granulocytes immatures (Promyélocytes, Myélocytes, Métamyélocytes). Promyélocytes: Sphériques avec un noyau rond à ovale et un cytoplasme contenant des granules primaires. Myélocytes: Ressemblent aux neutrophiles avec un noyau sphérique et un cytoplasme contenant des granules secondaires. Métamyélocytes: Ils ont un noyau en forme de fer à cheval et un cytoplasme similaire aux myélocytes.

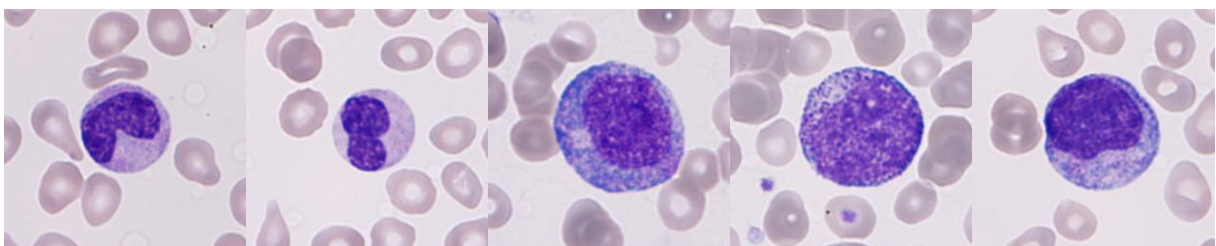


Figure 6 : Granulocytes immatures: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone

Érythroblastes: Précurseurs des érythrocytes, ils sont ronds avec un grand noyau central. Au fur et à mesure de leur maturation, le noyau diminue de taille et le cytoplasme devient plus éosinophile.

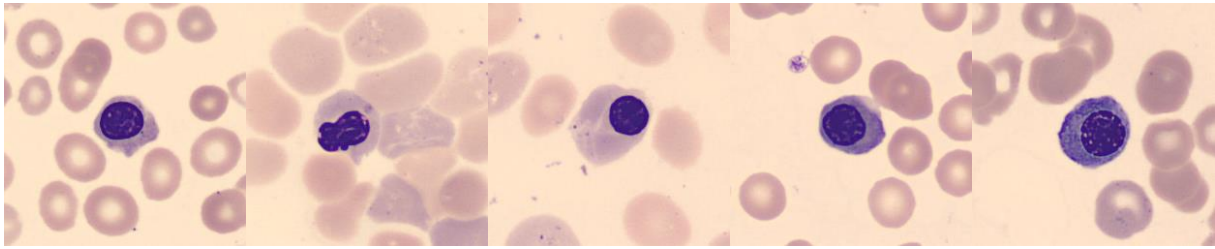


Figure 7 : Erythroblastes: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone

Plaquettes ou Thrombocytes: De forme irrégulière, généralement discoïdes ou ovales, les plaquettes sont beaucoup plus petites que les autres cellules sanguines. Leur structure est dépourvue de noyau, mais elles contiennent de nombreux granules.

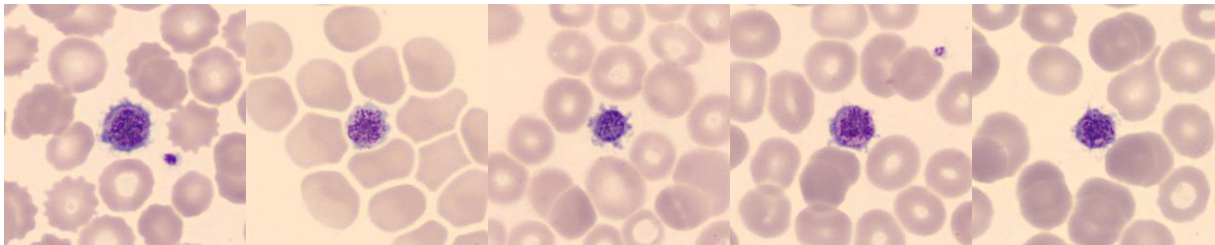


Figure 8 : Plaquettes: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone

NB : La morphologie des cellules sanguines peut varier en fonction des conditions physiologiques, des techniques de préparation d'échantillons et des méthodes de coloration.

La richesse descriptive des différentes catégories cellulaires, bien que cruciale pour la compréhension biologique, introduit une complexité notable pour la modélisation et la classification automatisée. Cette complexité souligne l'importance d'approches robustes et précises pour assurer une identification correcte et fiable des cellules dans notre projet.

La problématique de notre projet repose sur les défis inhérents à la classification des catégories cellulaires sanguines via un modèle de machine learning. Nombre d'entre elles, tels que les neutrophiles, éosinophiles et basophiles, partagent une forme sphérique, augmentant ainsi le risque de confusions lors de la classification. D'autres similitudes, comme celles des noyaux de certaines cellules (par exemple, métamyélocytes et monocytes) ou des caractéristiques cytoplasmiques, accentuent cette complexité. De subtiles différenciations, telle que la teinte ou la taille des granules, pourraient être moins discernables, en particulier à faible grossissement. Par ailleurs, des variables telles que les variations individuelles ou les possibles erreurs dans le processus de coloration peuvent introduire d'autres niveaux d'incertitude.

- Description des types de cellules sanguines à classifier

Les concentrations des différents types de cellules dans le sang sont très différentes, les plus nombreuses étant les érythrocytes (globules rouges) qui ne font pas partie de notre jeu de données.

Chez un individu sain, la concentration sanguine des cellules immatures (érythroblastes et granulocytes immatures) est très faible ; la plupart de ces cellules se trouvent dans la moelle osseuse.



Erythrocytes	4 500 000 à 6 000 000
Plaquettes	150 000 à 450 000
Neutrophiles	3 000 à 7 000
Lymphocytes	1 000 à 2 000
Monocytes	200 à 1 000
Eosinophiles	50 à 500
Basophiles	20 à 300
Erythroblastes	Très faible
Granulocytes Immatures	Très faible

Figure 9 : Concentration de chaque catégorie de cellule dans un microlitre de sang, pour un patient sain

- Hématopoïèse

La façon dont les cellules se forment n'est pas indifférentes à notre problème de classification. Dans le jeu de données mises à disposition, certaines cellules sont immatures et constituent une étape transitoire vers la formation d'une cellule appartenant à une autre catégorie. A cet égard, il serait intéressant de se renseigner auprès d'un expert en hématologie pour en connaître d'avantage sur le processus de maturation des cellules sanguines : s'agit-il d'un processus continu, qui rendrait difficile la délimitation stricte entre les catégories de cellules matures et immatures ou d'un processus au cours duquel la transition d'état entre les cellules immatures et matures intervient en un temps très bref, avec des caractéristiques distinctes visibles sur les images et ne laissant aucune place à l'ambiguïté ?

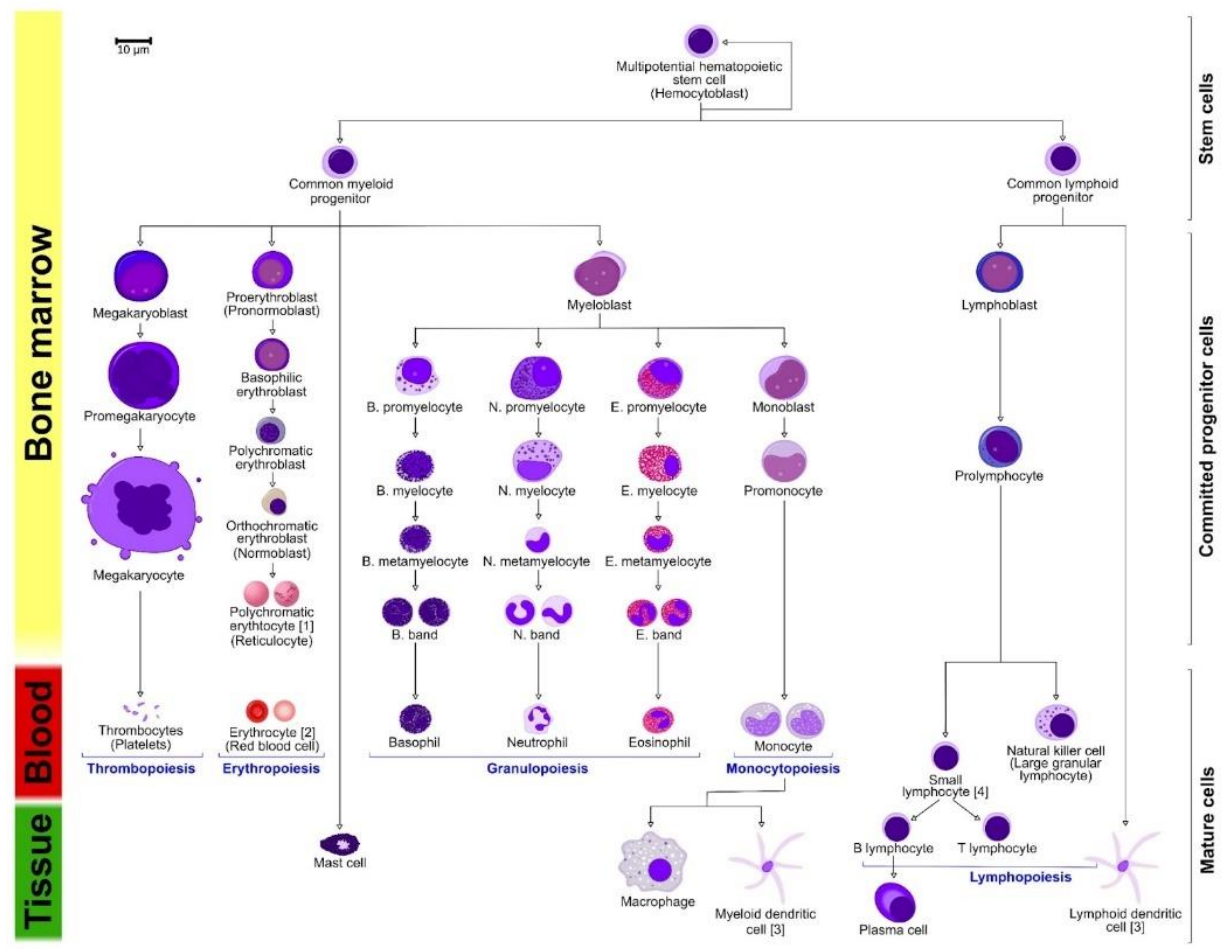


Figure 10: Hématopoïèse

- Rôle de chaque type de cellule dans le sang

Les cellules sanguines jouent un rôle primordial dans diverses fonctions de l'organisme, allant de la défense immunitaire à la distribution d'oxygène aux tissus. Afin de mieux comprendre et classer ces cellules, elles ont été regroupées en huit catégories distinctes, chacune ayant ses propres caractéristiques et fonctions.

Neutrophiles: Les neutrophiles sont le type le plus courant de globules blancs et jouent un rôle vital en tant que première ligne de défense de l'organisme contre les infections. Ils ont la capacité d'engloutir et de détruire les bactéries et autres micro-organismes nuisibles.

Éosinophiles: Les éosinophiles sont essentiels dans la réponse aux infections parasitaires et aux réactions allergiques. Contenant des granules spécifiques, ils libèrent leurs contenus en réponse à certaines infections pour neutraliser les menaces.

Basophiles: Les basophiles, bien qu'étant les moins courants parmi les globules blancs, ont un rôle significatif dans les réactions inflammatoires, en particulier les réactions allergiques. Ils sont responsables de la libération d'histamine, une substance chimique impliquée dans les symptômes allergiques.

Lymphocytes: Les lymphocytes se trouvent au cœur de la réponse immunitaire adaptative. Ils se déclinent en lymphocytes T et B, qui ciblent spécifiquement les pathogènes et produisent des anticorps, respectivement.

Monocytes: Les monocytes, lorsqu'ils migrent vers les tissus, se transforment en macrophages. Ces cellules jouent un rôle clé dans la lutte contre certaines bactéries, engloutissant et détruisant les envahisseurs.

Granulocytes immatures (Promyélocytes, Myélocytes, Métamyélocytes): Ces cellules représentent des stades préliminaires dans la maturation des granulocytes. En progressant à travers ces phases, ils mûrissent pour devenir soit des neutrophiles, éosinophiles, soit des basophiles.

Érythroblastes: Précurseurs des érythrocytes ou globules rouges, les érythroblastes sont essentiels à la production d'hémoglobine. Cette protéine est responsable du transport de l'oxygène à travers le sang.

Plaquettes ou Thrombocytes: Ces fragments cellulaires jouent un rôle vital dans la coagulation du sang. Lorsqu'une lésion survient, ils s'agrègent pour former un caillot, empêchant ainsi un saignement excessif.

La distinction et la compréhension précise de ces cellules sont cruciales en hématologie, car elles fournissent des informations inestimables sur la santé globale et les éventuelles pathologies d'un individu.

Exploration des données

- Exploration du jeu de données

Observation initiale des images

L'ensemble initial de données, obtenu à partir du laboratoire central de l'hôpital clinique de Barcelone, contenait 17 092 images de cellules issues de frottis sanguins. Les observations préliminaires ont révélé, dans un petit nombre de cas, la présence de plusieurs cellules par images, ce qui constitue une anomalie manifeste (tout au moins dans le cas où ces cellules appartiennent à différentes catégories), puisque chaque image ne comporte qu'un libellé unique de catégorie de cellule. On peut supposer que la cellule la plus centrale sur l'image est celle qui a fait l'objet de la catégorisation, mais dans certains cas, aucune cellule n'occupe le centre de l'image².

On remarque également dans le jeu de données certaines images dont la teinte de fond est différente, suggérant que le protocole de colorisation de May Grünwald-Giemsa, s'il constitue une pratique de référence, ne garantit pas une stricte homogénéité de la teinte des images mises à disposition.

Par ailleurs certaines images ont été identifiées comme étant corrompues ou illisibles. C'est le cas notamment de l'image étiquetée ".DS_169665.jpg" dans la classe neutrophile.

Standardisation de la taille des images

Il a été observé que les tailles des images variaient entre les classes. La taille majoritaire, 363*360, a été choisie comme standard pour toutes les images.

Distribution des classes

² On peut être surpris (et déçu) qu'un jeu de donnée de référence constitué en vu de la classification des cellules sanguines par *Computer Vision* comporte de telles anomalies. Mais ces anomalies ont au moins l'avantage de nous donner une idée de la qualité des données à attendre si l'on envisageait une mise en œuvre pratique de notre modèle auprès de praticiens hospitaliers et, sous réserve que l'on sache les éliminer, sont peut-être paradoxalement un atout pour une approche réaliste de la qualité de notre prédiction.

Par la suite, notre approche a consisté à traiter ces anomalies en appliquant deux stratégies opposées, l'une pour l'entraînement du modèle l'autre pour le test. Pour l'entraînement du modèle, on privilégiera un nettoyage le plus complet possible du jeu de donnée afin d'éviter tout biais lié à un artefact. En revanche, pour le test du modèle, la sincérité de l'évaluation de la performance exige que nous ne mettions en œuvre que les procédures de nettoyage des données qui n'utilisent pas le libellé de la catégorie de la cellule.

Le jeu de données n'est pas équilibré et n'est pas non plus représentatif de la proportion des différentes catégories de cellules dans le sang d'un patient sain.

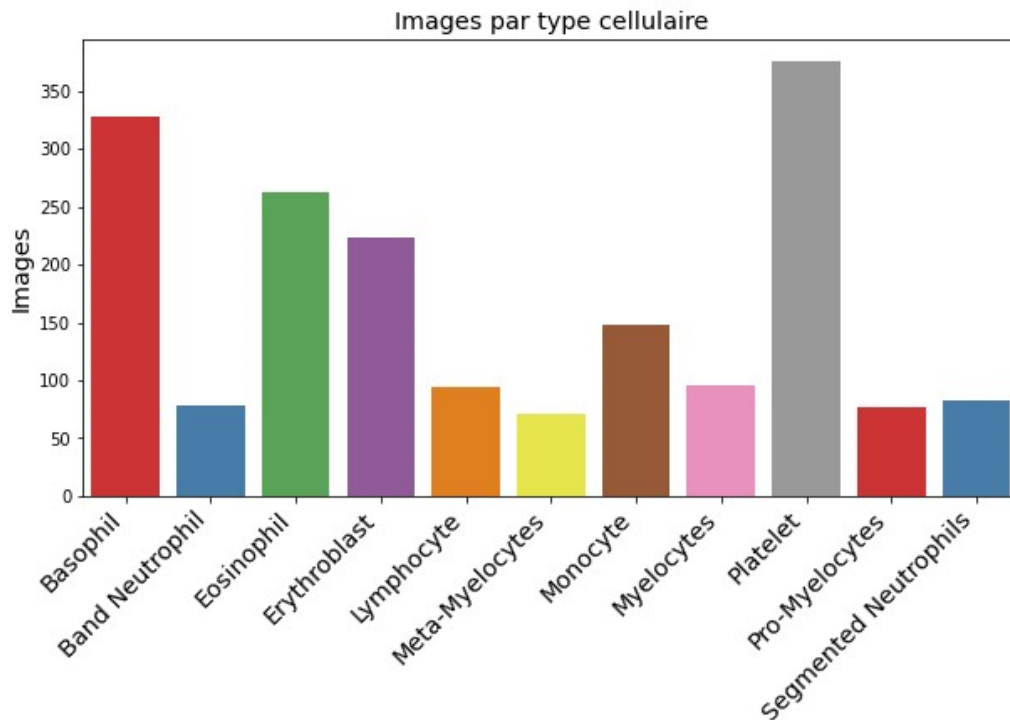


Figure 11: Distribution des catégories de cellules dans le jeu de données

Analyse en coordonnées polaires

- Justification de la transformation en coordonnées polaires

Les images sont initialement codées en coordonnées cartésiennes avec une dimension d'intensité RGB. La transformation polaire est suggérée par le fait que les cellules ont une forme ronde ou oblongue et qu'elles n'ont ni haut, ni bas, ni dessous, ni dessus. A la différence des photographies à l'échelle macroscopique où, par exemple, les lignes verticales et horizontales des bâtiments rendent très intuitive l'usage des coordonnées cartésiennes, une représentation en coordonnées polaires semble une représentation plus naturelle des cellules sanguines.

- Centrage des images

Avant de passer à la transformation polaire, il est essentiel de centrer correctement les images. Un algorithme spécifique a été développé pour ce centrage, se fondant sur une équipartition à gauche et à droite de l'image, et en haut et en bas, des pixels sombres, qui représentent généralement les cellules d'intérêt³.

- Analyses post-transformation : signature radiale des cellules et catégories de cellules

³ Une méthode plus élégante, consisterait à utiliser une annotation manuelle du centre des cellules et un algorithme d'apprentissage automatique de ce centre.

Suite à la transformation en coordonnées polaire, une attention particulière a été portée sur la composante radiale. Nous avons analysé l'intensité moyenne de chaque couleur en fonction de sa distance au centre de l'image. C'est ce que nous appelons la *signature radiale* d'une cellule (ou catégorie de cellule).

Intensité de la couleur en fonction de la distance au centre de l'image - platelet

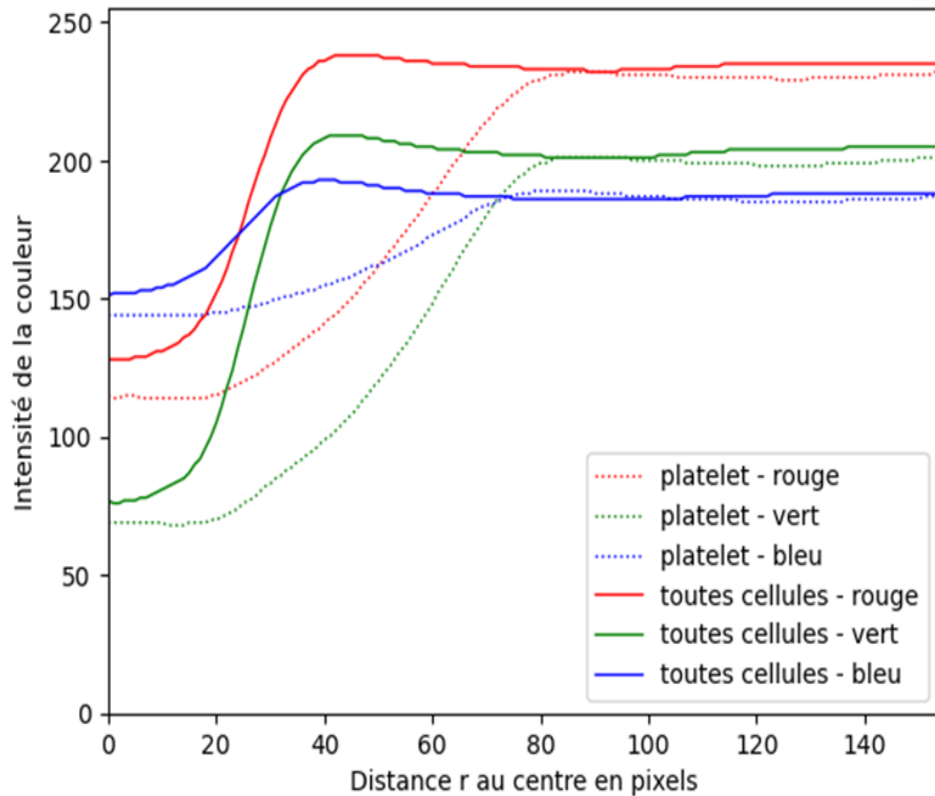


Figure 12: Signature radiale d'une catégorie de cellule : les plaquettes sanguines

Il ressort de ces analyses que chaque catégorie de cellule se caractérise par une signature radiale qui lui est propre⁴.

- Classification des cellules selon leur distance radiale à la classe la plus proche.

La caractérisation d'une cellule donnée par sa signature radiale et son rapprochement avec la signature radiale de la catégorie la plus proche (en suivant une distance euclidienne⁵) constituent un premier essai de classification des cellules. Ce modèle permet une classification correcte dans 62% des cas (pour la classification en 8 classes), performance correcte mais très en deçà du attendu pour une utilisation clinique⁶.

- Repérage d'images aberrantes

⁴ Les graphiques représentant les signatures radiales de chaque catégories de cellules figurent en annexe de ce rapport.

⁵ D'autres distances sont peut-être mieux adaptés que la distance euclidienne.

⁶ La matrice de confusion de ce modèle est présentée en annexe de ce document.

La *signature radiale* d'une image nous permet également de repérer des images aberrantes. Pour ce faire, on calcule la distance entre la signature radiale d'une image et sa classe de rattachement réelle. Les images les plus distantes de leur classe de rattachement constituent des suspects.

Distribution des distances entre la 'réduction radiale' de l'image et celle de sa classe

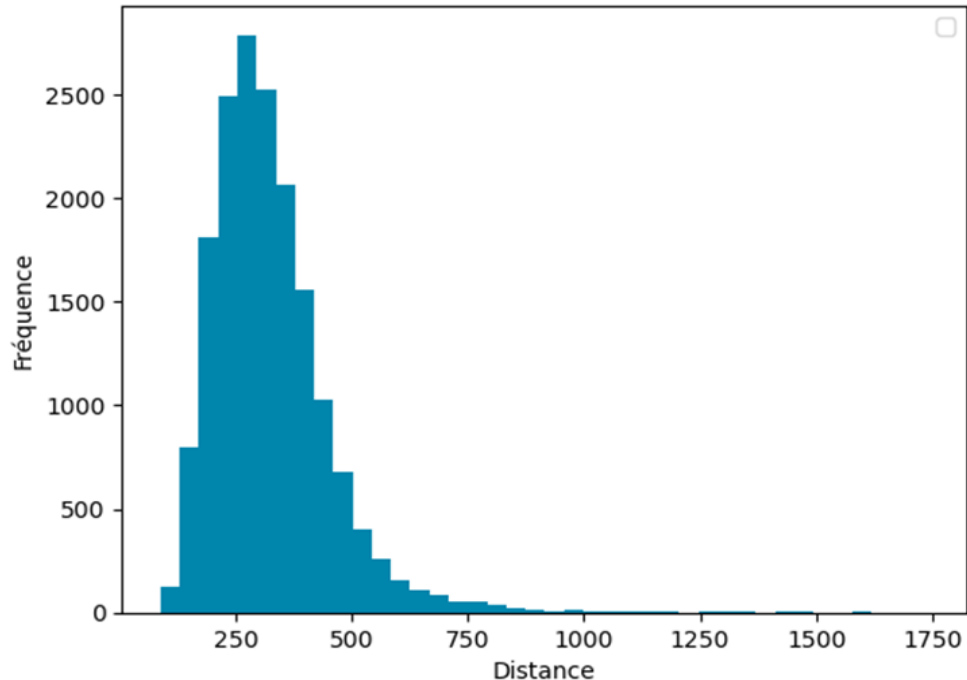


Figure 13: Histogramme de la distance des images à leur classe de rattachement, selon leur signature radiale

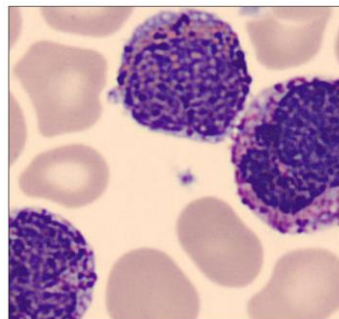


Figure 14: Un suspect démasqué : l'image comporte 3 cellules !

- Utilisation de la transformation polaire en entrée d'un modèle de *Deep Learning*

La méthode de transformation en coordonnées polaires a montré son potentiel pour l'analyse et le nettoyage de notre jeu de données. Plusieurs compléments ou améliorations pourraient être apportées à cette approche, notamment le choix d'une autre distance ou l'exploitation de la composante angulaire.

Au vu de ce succès, il serait tentant d'explorer les apports d'une transformation polaire, en entrée d'un réseau de convolution⁷.

⁷ Il est intéressant d'observer qu'une transformation polaire accorde un poids accru aux pixels au centre de l'image, ce qui assemble un argument en faveur de cette transformation, car les pixels centraux sont ceux qui

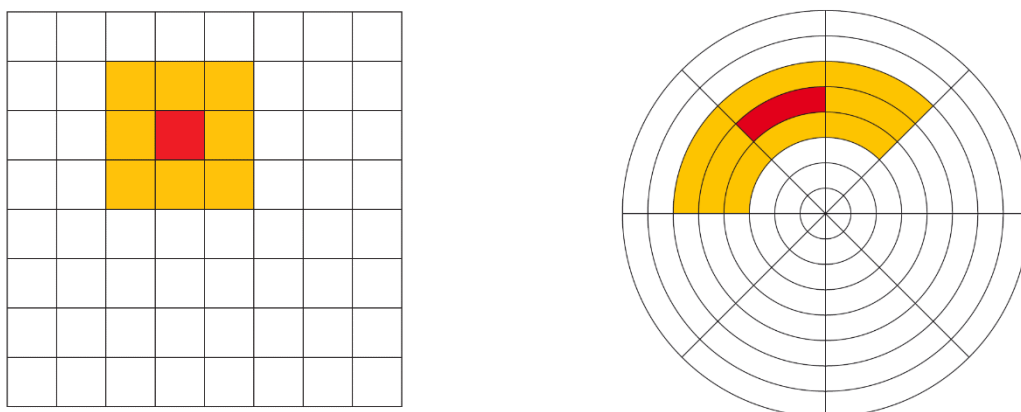


Figure 15: Représentation d'une convolution en coordonnées cartésienne et polaire

Nous n'avons pas exploré cette voie plus avant pour deux raisons. En premier lieu, le niveau de performance atteint pas un réseau de convolution basé sur une représentation en coordonnées cartésiennes s'est avéré avoir des performances pleinement satisfaisantes, rendant une approche alternative inutile. Par ailleurs la transformation en coordonnées polaires comporte des défis méthodologiques spécifiques : par exemple, celui du calcul du gradient en coordonnées polaires.

Conscient que nous n'avons pas démontré son potentiel, nous suggérons néanmoins de façon intuitive que la transformation polaire pourrait être utile pour traiter certains types de problème de *Computer Vision*, particulièrement à l'échelle cellulaire⁸.

portent le plus d'information (la périphérie de l'image est l'environnement de la cellule). Un autre argument potentiel est une meilleure interprétabilité des résultats en coordonnées polaire (repérage de caractéristiques).

⁸ Si nous suggérons l'intérêt potentiel d'une représentation polaire pour catégoriser les cellules sanguines, nous sommes loin de penser que la représentation polaire serait utile pour améliorer la performance de la plupart des modèles de *Computer Vision*. A l'échelle macroscopique, et encore d'avantage en ce qui concerne la plupart des productions humaines (et spécifiquement les bâtiments), la présence de lignes de droites et d'angles droits plaide pour une plus grande pertinence d'une représentation cartésienne.

En ce qui concerne l'échelle cellulaire, il n'est pas certain que la représentation polaire soit plus efficace que la représentation cartésienne. Peut-être, d'ailleurs, y aurait-il intérêt à associer ces deux modes de représentation dans un modèle hybride plutôt qu'à les envisager comme des substituts exclusifs.

Prétraitement des données : mise en œuvre du modèle U-Net

- Vue d'ensemble de la chaîne de traitement

Dans l'univers complexe et riche de l'imagerie biomédicale, la précision et la spécificité de l'analyse sont primordiales. Face à cette exigence, nous avons choisi d'adopter une approche modulaire, décomposant notre pipeline de traitement en deux étapes distinctes, mais complémentaires : la segmentation et la classification. La première étape, la segmentation, vise à délimiter avec précision les structures d'intérêt dans nos images et à en masquer le fond. Pour ce faire, nous avons jeté notre dévolu sur l'architecture U-Net, reconnue pour sa puissance dans le domaine de la segmentation d'images biomédicales. La seconde étape, la classification, a pour mission d'identifier et de catégoriser ces structures. Ici, la force d'un modèle pré-entraîné, en l'occurrence le ResNet-34, se révèle être un atout inestimable, capitalisant sur des connaissances antérieures pour affiner et accélérer notre analyse. La complémentarité de ces deux modèles, l'un axé sur la précision spatiale et l'autre sur l'identification, nous offre une vision holistique et approfondie des images biomédicales à notre disposition.

Avant d'aborder dans le détail chaque étape d'analyse, nous partageons une cartographie des analyses effectuée et de notre chaîne de traitement. C'est à cet objectif que répond le schéma suivant, qui ne vise ni à expliquer ni à justifier dans le détail nos choix. Nous invitons le lecteur à s'y référer lors de la lecture de ce document.

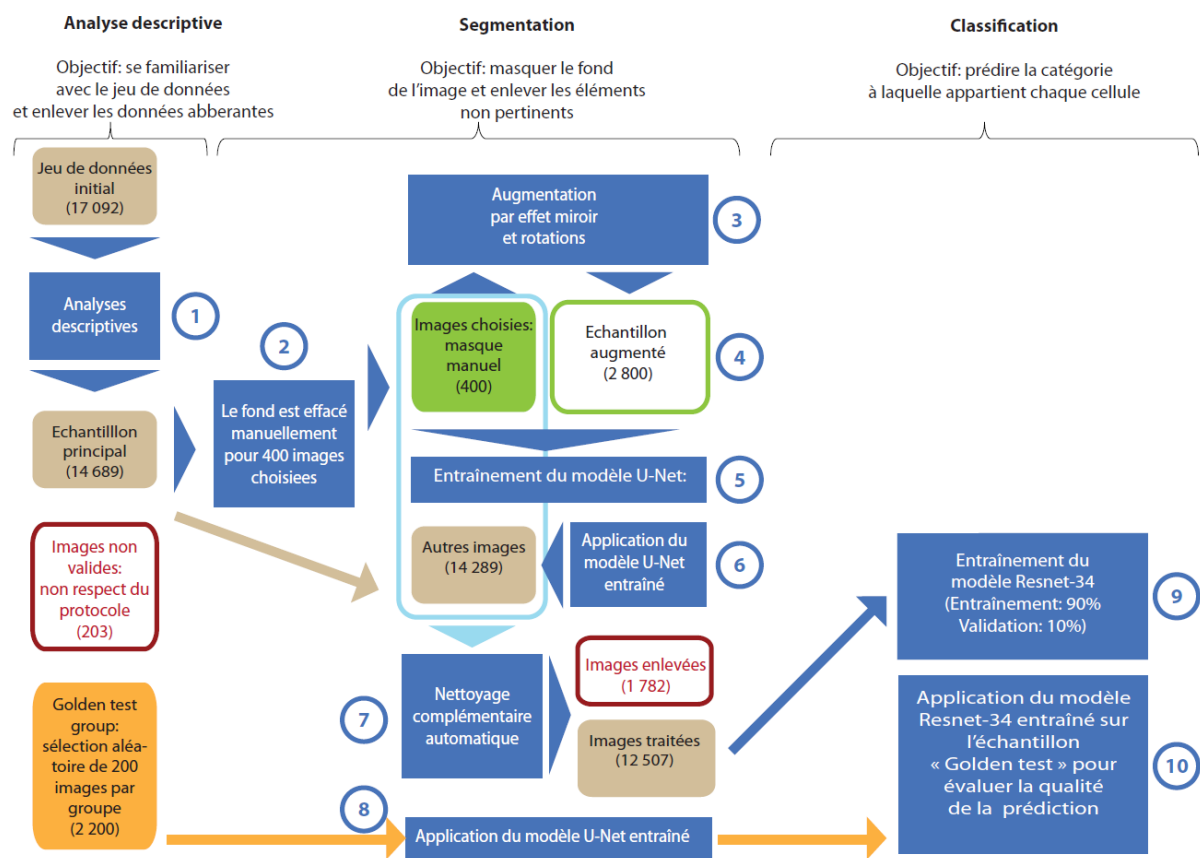


Figure 16: Chaîne de traitement

Nous attirons particulièrement l'attention sur l'échantillon « Golden Test », extrait dès le départ de façon aléatoire et qui comporte 200 images par catégorie cellulaire. Cet échantillon garantit la sincérité de notre classification.

- Nettoyage préliminaire du jeu de données

Un premier assainissement manuel a été réalisé. Cette intervention visait à garantir l'intégrité des données avant tout traitement automatisé ultérieur. Il s'agit d'images manifestement aberrantes (présence de multiples cellules sur l'image, teintes très différentes) qui n'ont pas leur place dans ce jeu de données, ni en entraînement, ni en test. Pour ces 203 images, on considère que le protocole correct de photographie des cellules n'a pas été bien respecté.

- Enjeux de la segmentation

Pour les images retenues, nous souhaitons masquer les éléments périphériques de l'image (fond) afin d'éliminer tout risque de biais dans notre analyse.

- Raisons du choix d'architecture U-Net

Plusieurs facteurs ont orienté notre décision vers U-Net. Tout d'abord, son succès antérieur dans la segmentation d'images biomédicales le rendait particulièrement attrayant pour notre contexte. De plus, sa capacité à travailler avec un nombre limité d'images d'entraînement correspondait bien à notre situation, où nous disposons d'un nombre initial réduit d'images segmentées manuellement. Enfin, l'architecture unique de U-Net, qui combine les cartes de caractéristiques à différentes résolutions, garantit que même les détails les plus fins de nos images cellulaires, cruciaux pour une classification précise, soient capturés avec fidélité. Enfin, avec U-Net, nous bénéficions de l'expérience acquise sur d'autres jeux d'images médicales, via du *Transfer Learning*.

U-net s'est avéré être l'outil idéal pour répondre aux exigences et défis spécifiques de la phase de segmentation de notre projet.

- Structure du modèle U-Net

L'architecture U-net est caractérisée par sa forme en U, comprenant une partie contractante suivie d'une partie expansive. La partie contractante est composée de couches convolutives, permettant de capturer les caractéristiques essentielles des images, tandis que la partie expansive reconstruit l'image à partir de ces caractéristiques.

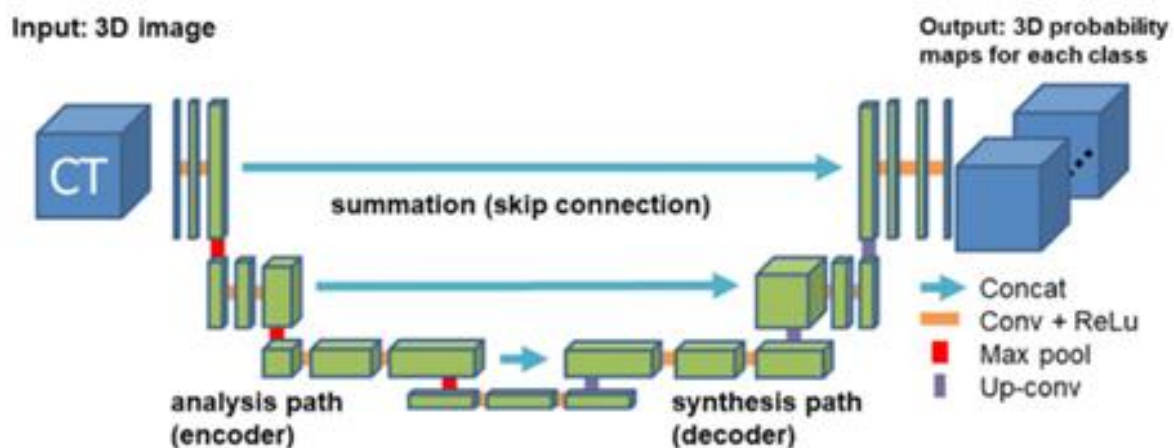


Figure 17: Schéma architectural de U-Net

- Harmonisation de la taille des images

Face à des résolutions d'images disparates, nous avons décidé d'harmoniser ces dimensions pour faciliter leur traitement par réseau neuronal. Après analyse, nous avons choisi une résolution standard de 352 x 352 pixels. Ce choix ne fut pas anodin : 352 est le chiffre le plus proche et divisible par 32, une condition pour la compatibilité avec notre architecture U-net.

- Segmentation manuelle

Pour pouvoir entraîner U-Net, la première étape de notre processus de prétraitement a été la segmentation manuelle. Nous avons segmenté 400 images choisies parmi celles ne comportant qu'une seule cellule et réparties équitablement entre chaque type cellulaire, en utilisant l'outil *Sefexa*. Ce choix visait à obtenir un ensemble de données initial bien défini sur lequel s'appuyer pour les étapes ultérieures. La segmentation manuelle a consisté à qualifier chaque pixel de l'image selon son appartenance à deux segments : soit la cellule (noyau ou cytoplasme), soit le fond de l'image⁹.

⁹ Signalons que nous avons aussi testé une segmentation en 9 classes, constituées par le fond et chaque type de cellule, et dont les résultats sont présentés en annexe. Cette approche n'a pas été privilégiée pour une classification directe des cellules. En effet, l'échantillon utilisé pour la segmentation (3200 images dont 2800 issues d'augmentation) étant moins riche que celui de la classification, les résultats attendus de la classification sont meilleurs. Malgré cette limitation, les résultats de la segmentation en 9 classes s'avèrent être excellents compte tenu du petit nombre d'images entraînées, confirmant la capacité du modèle U-Net à apprendre avec relativement peu d'images.

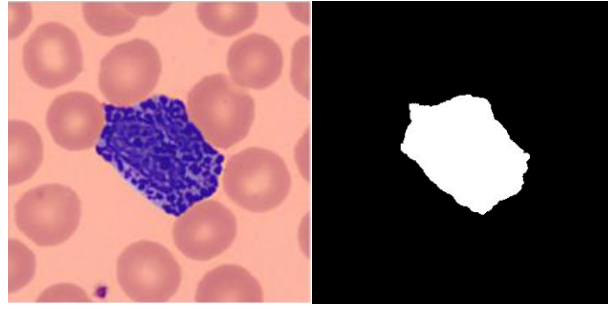


Figure 18: Un exemple d'images segmentées en deux classes : la cellule d'intérêt et le fond de l'image

- Augmentation des images

Pour renforcer la robustesse de notre modèle et augmenter notre ensemble d'entraînement, nous avons opté pour des techniques d'augmentation d'images non destructives. En appliquant à l'image des rotations à 90° et des effets miroir nous avons pu élargir notre base initiale de 400 images à un total de 3 200 images. Pour garantir la cohérence, chaque masque associé aux images originales a subi les mêmes transformations, assurant ainsi une correspondance parfaite avec les images transformées¹⁰.

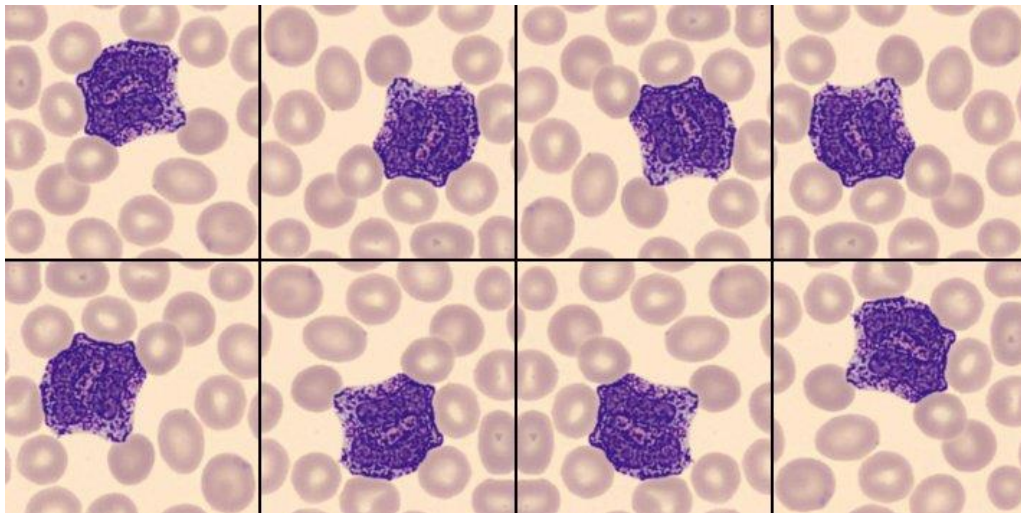


Figure 19: Processus d'augmentation d'images ; la cellule en haut à droite représente l'image initiales, les autres ses transformées

- Répartition des données en ensembles d'entraînement, de validation et de test

Pour mener à bien notre tâche de segmentation, nous avons disposé d'un ensemble de 3200 images classées en deux catégories : les cellules et le fond. Ces images ont été soigneusement réparties pour garantir une formation robuste et une évaluation fiable de notre modèle. 60% de ces images (soit 1920 images) ont été allouées à l'entraînement, assurant ainsi une base solide pour la calibration de notre

¹⁰ Nous aurions pu étendre encore d'avantage notre jeu de données en opérant des rotations d'angles différents. En revanche nous avons délibérément écarté les transformations étirant la cellule dans une certaines direction, ou les zoom / dé-zooms, car nous avons considéré que ces transformations étaient plus risquées. A la réflexion, le procédé d'étirement semble risqué pour la phase de classification (la déformation d'une cellule peut prêter à confusion sur sa forme) mais semble moins gênant pour l'étape de segmentation.

modèle. Les 20% suivants (640 images) ont servi de validation, permettant une évaluation intermédiaire et régulière des performances du modèle tout au long du processus d'apprentissage. Les 20% restants, également 640 images, ont été réservés pour le test du modèle de segmentation, afin d'évaluer l'efficacité du modèle dans des conditions réelles et inédites¹¹.

- Choix des hyperparamètres

Notre modèle U-Net vise à prédire, pour chaque pixel, si elle appartient à la cellule ou au fond de l'image. Le modèle de segmentation consiste donc en un modèle de classification binaire au niveau de chaque pixel, c'est pourquoi nous choisissons **comme fonction d'activation la fonction sigmoïde**.

Notre **taux d'apprentissage**, déterminant la taille des pas dans la mise à jour des poids pendant l'entraînement, a été fixé à une constante de **0.0001**. Cette valeur permet un apprentissage stable tout en garantissant que le modèle explore suffisamment l'espace des solutions. La taille de batch est fixée à 8 et le nombre d'époques d'entraînement à 30.

Quant à **notre fonction de perte**, nous avons adopté une **combinaison de 'dice_loss' et de 'BinaryFocalLoss'**.

Pour le modèle final **notre choix d'optimiseur** s'est arrêté sur SGD. Ce choix repose sur plusieurs facteurs clés. Tout d'abord, sa simplicité et sa compréhensibilité font de SGD un algorithme d'optimisation bien établi et souvent utilisé comme point de départ. De plus, avec une base de 3200 images, sa capacité à mettre à jour fréquemment les poids le rend particulièrement efficace, favorisant une convergence rapide. L'avantage de SGD est aussi son aptitude à éviter les minimums locaux grâce à son approche stochastique. Même avec un learning rate fixé à 0.0001, il offre une adaptabilité, et sa nature stochastique est avantageuse pour explorer exhaustivement l'espace de fonctions de coût complexes comme celle associée à Dice_loss et BinaryFocalLoss. De plus, il est reconnu pour son efficacité avec des architectures comme U-net et permet un ajustement fin des poids pour optimiser le modèle, surtout dans les phases avancées de l'entraînement. En dépit de l'émergence d'optimiseurs plus sophistiqués, SGD demeure une référence solide, en particulier pour des architectures comme U-Net combinées à des fonctions de perte complexes.

- Métriques utilisées pour évaluer le modèle

La performance de notre modèle de segmentation U-Net basé sur l'architecture ResNet-34 a été évaluée à l'aide de **l'Intersection over Union score (IoU-Score)**. C'est une métrique incontournable pour les tâches de segmentation, permettant de mesurer le degré de correspondance entre la prédiction du modèle et la vérité terrain. Une valeur d'IoU proche de 1 est indicative d'une segmentation de haute précision. Bien que l'IoU-Score soit la métrique principale, des métriques complémentaires comme la précision, le rappel et le F1-score sont également pertinentes pour une évaluation plus nuancée.

¹¹ Pour éviter tout biais statistique dans la mesure de la performance du modèle, il faudrait veiller à ce que les images augmentées soient classées dans le même groupe que les images ayant servies à leur augmentation.

- Entraînement du modèle et techniques pour éviter le surapprentissage

Afin de prévenir le surapprentissage et de garantir une généralisation efficace de notre modèle à de nouvelles données, plusieurs stratégies ont été adoptées. En premier lieu, la répartition des données joue un rôle essentiel dans la prévention du surapprentissage. 60% des images (1920) ont été allouées à l'entraînement, 20% (640 images) à la validation, et les 20% restants (640 images) au test. Cette séparation garantit que le modèle est évalué sur des images inédites, fournissant ainsi une évaluation fiable de sa capacité à généraliser. En regardant l'évolution des performances du modèle (IoU et fonction de perte) sur l'échantillon de validation et en les comparant à celle de l'échantillon d'entraînement, on peut s'assurer du contrôle du surapprentissage : il y aurait surapprentissage si ces métriques étaient bien plus favorables sur l'échantillon d'entraînement par rapport à l'échantillon de validation. Nous observons sur la figure suivante que le modèle est très satisfaisant à cet égard, l'écart étant limité.

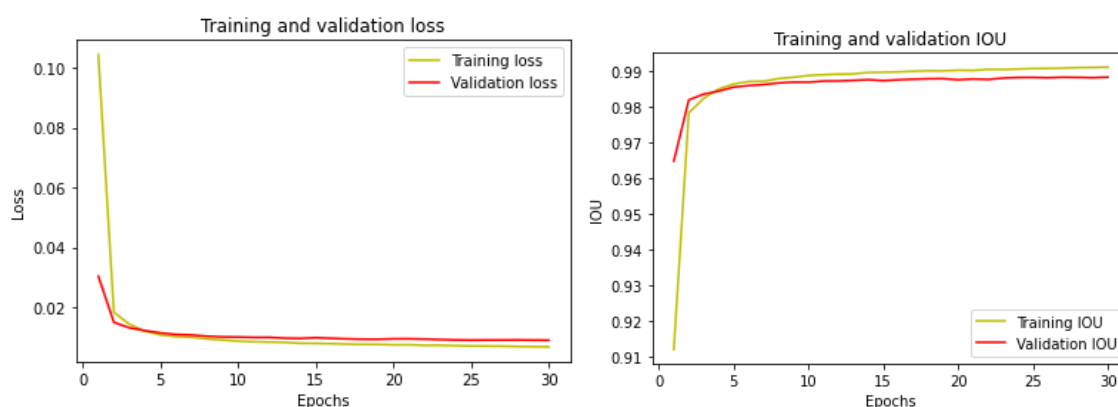


Figure 20: : Courbes d'apprentissage du modèle U-Net en 2 classes

Nous avons aussi employé un mécanisme de contrôle spécifique, 'ModelCheckpoint', qui surveille la métrique IoU sur l'ensemble de validation. Grâce à celui-ci, le modèle ne sauvegarde que les instances surpassant un score IoU de 0.9, assurant ainsi que seuls les modèles les plus performants sont retenus.

Bien que des techniques comme le *dropout* et l'augmentation de données soient couramment utilisées pour régulariser les modèles, notre approche s'est principalement concentrée sur la répartition judicieuse des données et la surveillance attentive de la performance du modèle grâce à 'ModelCheckpoint'.

- Performance du modèle sur l'échantillon de test

L'IoU du segment 'cellule' sur l'échantillon de test est de **0,979** ce qui constitue un résultat très satisfaisant. Plus précisément : 16 777K pixels de l'ensemble des images de test ont été correctement classées comme 'fond' (background) et 11 777K pixels ont été correctement identifiés comme 'cellules'.

Les éléments hors diagonale, 125K et 128K, correspondent aux erreurs de classification des pixels, respectivement pixels classés à tort en 'fond' et en 'cellule'¹².

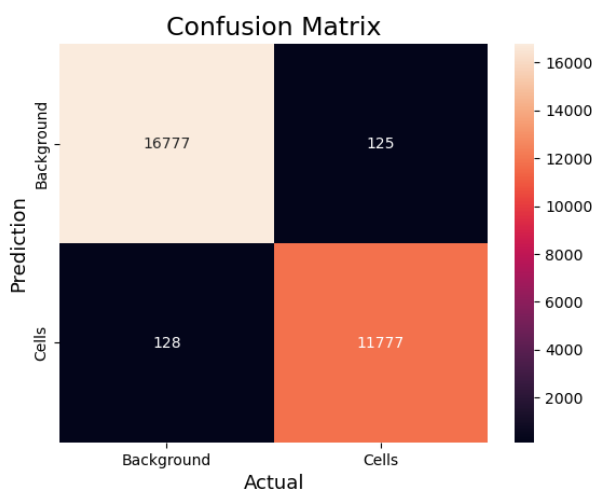


Figure 22: Matrice de confusion normalisée pour la classification des pixels avec U-net. Une unité correspond à 1000 pixels.

La qualité de la classification effectuée par le modèle U-Net est corroborée par l'observation directe de quelques cellules de l'échantillon test.

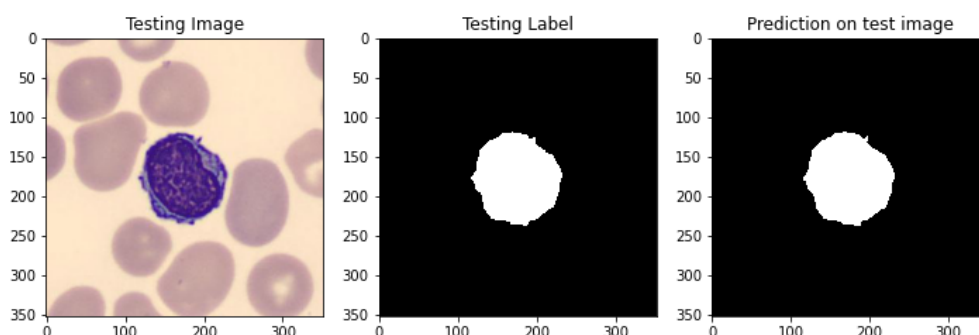


Figure 21: Exemple de prédiction de U-Net sur une image de l'échantillon test comportant une seule cellule

- Interprétation des résultats et identification des domaines de réussite

Le modèle de segmentation basé sur l'architecture U-Net a démontré une efficacité remarquable compte tenu du petit nombre d'images de l'échantillon d'entraînement (1920, dont 87.5% d'images issue de l'augmentation). De telles performances peuvent être attribuées à une combinaison de choix appropriés en matière d'optimisation, d'architecture de réseau et de techniques de régularisation. La segmentation est particulièrement délicate dans le domaine biomédical, étant donné la complexité des structures et la variabilité entre les échantillons. Le succès de notre modèle dans une telle tâche est donc un signe positif de son applicabilité dans des scénarios réels.

¹² Nous sommes confortés dans notre optimisme par l'excellente performance du modèle de segmentation en 9 classes, dont les résultats sont présentés en annexe, qui a un IoU moyen de 0.946.

- Limitations du modèle et perspectives d'amélioration

Pour améliorer davantage le modèle, plusieurs avenues peuvent être explorées. Premièrement, une extension de l'augmentation des données, notamment en introduisant des rotations différentes de 90° degrés, pourrait aider le modèle à mieux généraliser. L'introduction d'augmentations comportant des zooms ou étirement d'image pourrait également être testé mais doit s'accompagner d'une grande prudence, ce type de transformation pouvant, par leur nature, comporter certains dangers. Deuxièmement, une analyse approfondie des erreurs de segmentation pourrait fournir des indices sur les domaines spécifiques à améliorer. Cette analyse fine pourrait révéler des tendances d'erreurs que le modèle commet, guidant ainsi les efforts d'optimisation. Enfin, la recherche d'autres architectures de réseau ou la combinaison de plusieurs modèles en une approche d'ensemble pourraient aider à améliorer la robustesse et la précision de la segmentation. Une collaboration avec des experts du domaine fournirait des enseignements précieux sur les caractéristiques spécifiques à prioriser ou sur des techniques de prétraitement des images qui pourraient améliorer les performances du modèle.

Néanmoins, la principale réserve que l'on peut faire quant à l'usage de U-Net ne tient pas à la performance du modèle, mais au rôle qu'on lui a assigné. Rappelons que nous souhaitons obtenir, pour chaque image, une cellule unique dont on masquerait le fond. Or U-Net ne remplit pas exactement ce rôle : U-Net qualifie pour chaque pixel son segment d'appartenance : fond ou cellule. Dans le cas où les images comportent plusieurs cellules, U-Net ne masque donc pas les cellules périphériques. En conclusion : l'application d'un modèle U-Net pour masquer le fond des images constitue une étape très utile, mais incomplète pour le nettoyage de notre jeu de données.

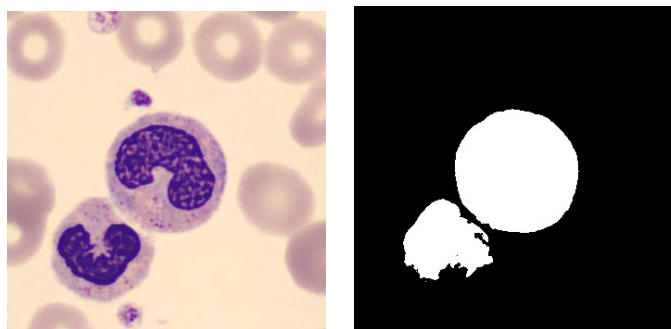


Figure 22: Exemple de prédiction de U-Net sur une image de l'échantillon test comportant deux cellules

- Nettoyage complémentaire des images

U-Net ne permettant pas de nettoyer complètement notre jeu de données, trois étapes complémentaires visent à garantir la pertinence et la fiabilité des données traitées en entrée de notre modèle de classification.

- Élimination des images multicellulaires

Toutes les images contenant plus d'une cellule ont été retirées. Cette étape permet de garantir la simplicité et la clarté de notre analyse, en se focalisant sur des images montrant une seule cellule.

- Suppression des débris cellulaires

Les images présentant des débris cellulaires ou des fragments d'autres cellules ont été éliminées. Ces artefacts peuvent entraîner des interprétations erronées ou biaiser nos modèles de traitement, il est donc essentiel de les retirer pour maintenir la pureté de notre jeu de données.

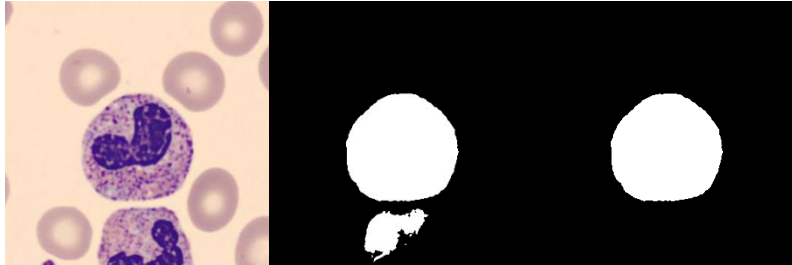


Figure 23: Etapes de nettoyage des images (doubles et débris)

- Comblement des espaces vides dans les masques

Les espaces vides présents dans les masques ont été comblés pour garantir une représentation plus uniforme et cohérente des cellules dans l'ensemble du jeu de données.



Figure 24: Exemple de comblement des espaces vides dans les masques

Cette approche méthodique du nettoyage garantit que notre jeu de données est à la fois propre et robuste, prêt à être utilisé pour des analyses ultérieures. Il a été réalisé par une procédure automatique.

Il faut noter que certaines étapes de cette procédure utilisent l'annotation de la catégorie de la cellule. Ce nettoyage n'est donc pertinent que pour préparer l'entraînement (et la validation) de notre modèle de classification. En revanche, ce nettoyage ne sera pas (et ne pourra pas) être utilisé lors d'une mise en œuvre pratique de notre prédiction sur de nouvelles images, contrairement à l'application du masque issu de U-Net.

Plus important encore, cette procédure ne devra pas être appliquée à notre échantillon 'Golden Test' visant à évaluer la performance finale de notre modèle de classification, au risque de compromettre la sincérité de nos résultats.

Modèle de classification : ResNet-34

- Enjeux de la classification

Nous en arrivons à l'étape finale de notre projet qui consiste à déterminer, à partir d'une image de cellule sanguine sa catégorie d'appartenance. Deux catégorisations sont présentées, l'une comportant 8 catégories de cellules, l'autre 11 catégories. Chacune de ces catégorisations donne lieu à la réalisation d'un modèle de prédiction distinct¹³.

Dans la configuration à 8 classes, les types de cellules sont agrégés en basophiles, neutrophiles, éosinophiles, érythroblastes, granulocytes immatures, lymphocytes, monocytes, et plaquettes. Pour la variante à 11 classes, les neutrophiles se subdivisent en neutrophiles à noyaux en bâtonnets et noyaux segmentés tandis que les granulocytes immatures se décomposent en métamyélocytes, myélocytes et promyélocytes.

- Raisons du choix de ResNet-34

Plusieurs éléments ont guidé notre choix vers ResNet-34. Sa renommée dans les domaines de la classification et de la reconnaissance d'images a naturellement retenu notre attention. L'un des principaux atouts de Res-Net34 est sa capacité à résoudre le problème de la disparition du gradient dans les architectures profondes grâce à ses connexions résiduelles, ce qui permet une formation efficace et stable sur de grands volumes de données. De plus, étant donné que nous avons besoin d'un modèle pré-entraîné pour optimiser la classification à partir d'un ensemble de données limité, ResNet-34, avec ses poids déjà formés sur des ensembles de données massifs comme *ImageNet*, semblait être le choix judicieux. Sa conception innovante, qui emploie des connexions d'identité pour faciliter le flux d'information et la rétention des détails essentiels, assure que toutes les nuances et variations de nos images cellulaires soient fidèlement représentées et distinguées. ResNet-34 s'est imposé comme la pierre angulaire de notre approche pour aborder la complexité inhérente à notre projet de classification cellulaire.

- Architecture du modèle ResNet-34

L'architecture ResNet-34 est remarquable par son concept de "*residual learning*". Elle se compose d'une série de blocs résiduels où chaque bloc contient des couches convolutives et est suivi d'un "*shortcut*" ou connexion d'identité. Plutôt que d'apprendre une représentation entière à chaque étape, ResNet apprend les résidus ou différences entre les entrées et sorties de chaque bloc. Ces connexions d'identité permettent de contourner ou de "sauter" certaines couches, facilitant ainsi la propagation arrière et l'entraînement sur de plus grandes profondeurs sans rencontrer le problème de la disparition du gradient.

¹³ En d'autres termes, la catégorisation à 8 classes ne constitue pas une catégorisation agrégée issue de notre modèle à 11 classes, bien que les 8 classes agrègent hiérarchiquement les 11 classes. Il est à noter que cette façon de faire peut induire des incohérences entre les résultats de la prédiction du modèle à 8 classes et celui du modèle à 11 classes. Dans une optique d'usage du modèle par des professionnels de la santé on pourra envisager de ne communiquer que l'un des deux modèles, pour éviter une perception d'incohérence dans des cas marginaux.

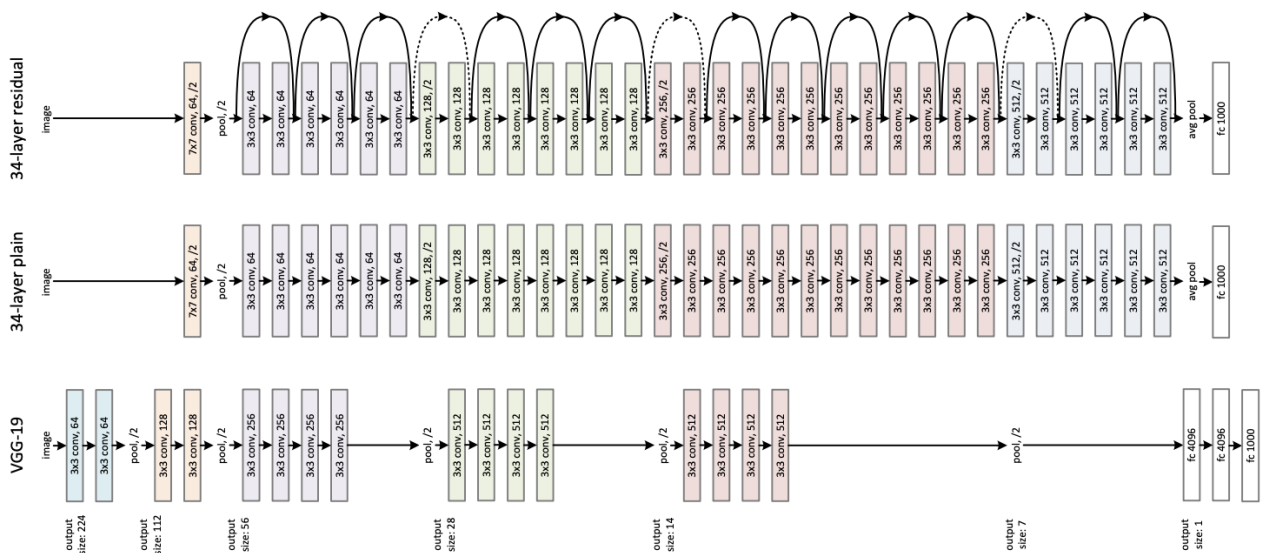


Figure 25: Schéma architectural ResNet34 et comparaison avec deux autres architectures plus simples

Nous adaptons le modèle ResNet-34 à nos besoins spécifiques en ajoutant une couche en entrée de dimension (352,352,3) et en le faisant suivre d'une couche de *Global Average Pooling* et d'une couche dense avec une activation 'softmax' pour classer les différents types cellulaires.

- Répartition des données en ensembles d'entraînement, de validation et de test

Pour mener à bien notre tâche de classification, nous avons disposé d'un ensemble de 12 507 images réparties en deux groupes : un groupe de training représentant 90% des images sélectionnées aléatoirement et un groupe de validation formé sur les 10% restantes. Il est à noter que cette subdivision est commune aux modèles à 8 et 11 classes. Le groupe de test sera lui constitué par les 2200 images du groupe 'Golden Test' que l'on avait sélectionnées au départ à cette fin, et qui ont au préalable été segmentées par U-Net afin d'en masquer le fond¹⁴.

- Choix des hyperparamètres du modèle

Concernant l'optimisation, nous avons adopté la méthode du **Stochastic Gradient Descent (SGD)** avec un **taux d'apprentissage initial de 0,01**, une **décroissance du taux à 0.01/20** et un **Momentum de 0.9** pour accélérer la convergence vers le minimum global.

Lors de l'entraînement, le modèle a été formé sur un ensemble de données composé de 11 256 images de cellules uniques, segmentées en plusieurs types cellulaires. Ces images ont été alimentées par **lots de 64 images** à la fois sur une durée totale de **20 époques**.

¹⁴ En revanche, aucun nettoyage dépendant de la classe à prédire n'est appliqué au groupe 'Golden Test' afin de préserver la sincérité de l'évaluation de la performance du modèle.

- Métriques utilisées pour évaluer le modèle

L'évaluation de notre modèle de classification s'est appuyée sur trois métriques : la **précision**, le **rappel** et le **score F1**. Chacune de ces métriques offre une perspective unique sur la performance du modèle.

- **Précision**. Elle indique la justesse des prédictions positives du modèle. Une précision élevée signifie que le modèle a correctement identifié un grand nombre de vrais positifs tout en minimisant les faux positifs.

- **Rappel**. Elle montre la capacité du modèle à détecter toutes les instances positives possibles. Un rappel élevé signifie que le modèle a capturé un grand nombre de vrais positifs par rapport aux faux négatifs.

- **F1-score**. C'est la moyenne harmonique de la précision et du rappel. Il est particulièrement utile lorsque les classes sont inégales, car il prend en compte à la fois les faux positifs et les faux négatifs.

- Entraînement du modèle et techniques pour éviter le surapprentissage

Pour prévenir le surapprentissage et assurer une généralisation efficace de notre modèle de classification ResNet-34 à de nouvelles données, nous avons mis en place des mécanismes rigoureux de régularisation. En premier lieu, le surapprentissage est contrôlé par la présence du groupe de validation (1 251 images) et au suivi des performances de ce groupe au cours des époque, en comparaison du groupe de test, selon les métriques de perte (*loss*) et de précision (*accuracy*). Si les tendances que suivent ces métriques sur les échantillons d'entraînement et de validation s'écartent de façon significative et durable à la faveur de l'échantillon d'entraînement, cela traduit un surapprentissage du modèle.

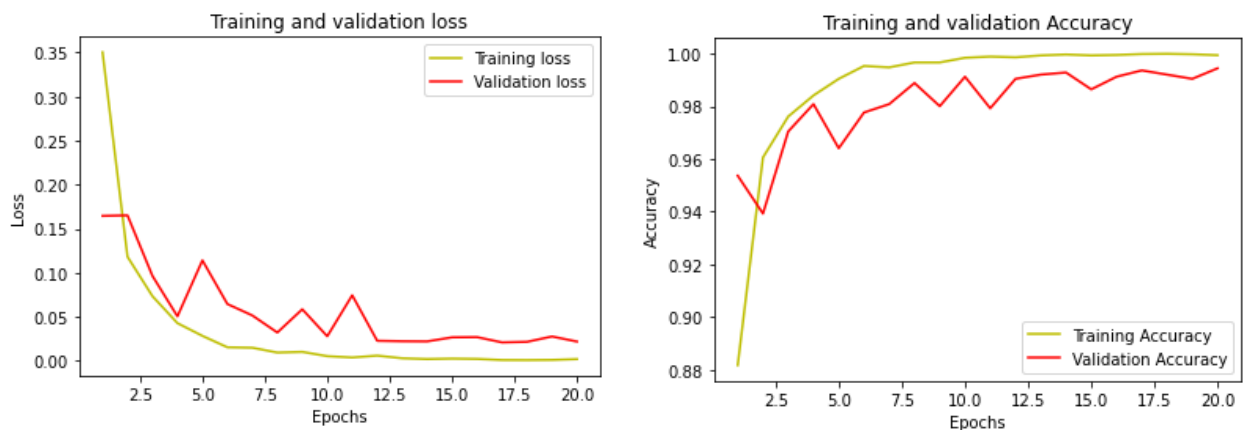


Figure 26: Courbes d'apprentissage du modèle Resnet-34 en 8 classes

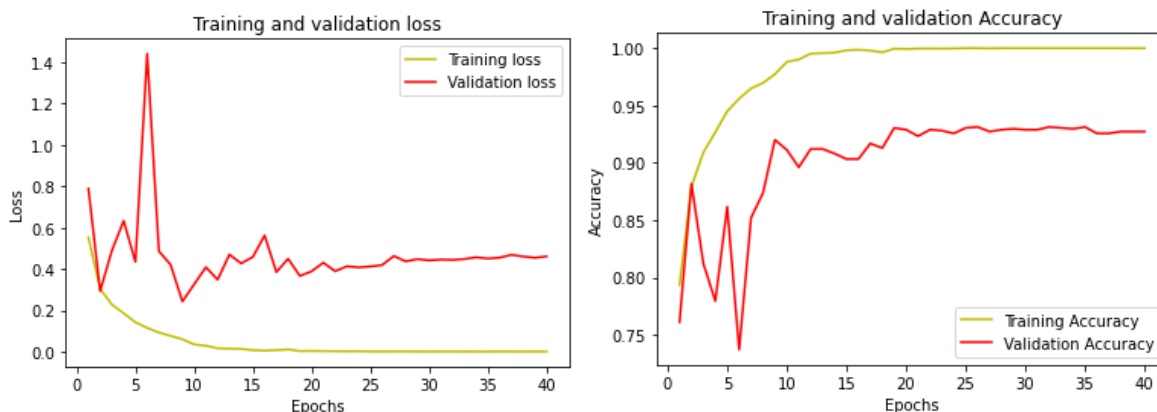


Figure 27: Courbes d'apprentissage du modèle Resnet-34 en 11 classes

Malgré nos efforts de prévention, nous constatons un surapprentissage, assez limité pour le modèle à 8 classes mais significatif pour le modèle à 11 classes.

Un mécanisme de surveillance via `'ModelCheckpoint'` a été utilisé. Ce dernier a permis de suivre la performance du modèle sur un ensemble de validation. Ainsi, seuls les modèles surpassant des critères prédéfinis de performance ont été sauvegardés, assurant la rétention des versions les plus prometteuses.

Bien que l'usage de techniques telles que le dropout ou l'augmentation de données soit courant pour éviter le surapprentissage, notre démarche a mis l'accent sur une distribution judicieuse de notre jeu de données et une surveillance stricte de la performance du modèle grâce à `'ModelCheckpoint'`.

- Performances du modèle sur l'échantillon 'Golden Test'

Le modèle a été formé sur l'ensemble d'entraînement avec des résultats prometteurs, comme le montre l'évolution rapide des courbes d'apprentissage. Cependant, pour une évaluation juste et impartiale, il est crucial d'examiner les performances sur l'ensemble de test. Le modèle à 8 classes a atteint une précision globale de 98.2% sur l'échantillon de test tandis que le modèle à 11 classes atteint la précision de 89.8% sur le même échantillon. Ce sont donc les catégories de cellules très proches biologiquement et qui sont agrégées dans le modèle à 8 classes qui s'avèrent les plus difficiles à distinguer, résultat qui semble intuitif.

Les matrices de confusion des deux modèles permettent d'affiner l'analyse en ayant une idée plus précise sur les classes pour lesquels la classification est correcte et celles pour lesquels on observe des erreurs de classification.

Modèle à 8 classes

La classification est excellente pour l'ensemble des catégories de cellules, à l'exception des neutrophiles et dans une moindre mesure des monocytes, confondus avec des granulocytes immatures. Les scores restent très élevés sur ces trois classes. Il est intéressant de noter que si l'on se réfère à l'hématopoïèse, certains granulocytes immatures donneront lieu à la génération de neutrophiles et monocytes. Cette remarque nous oriente vers une potentielle difficulté d'annotation de cellules qui seraient dans un stade intermédiaire de maturité entre des granulocytes matures et immatures, plutôt qu'en une défaillance de la capacité de prédiction de notre modèle.

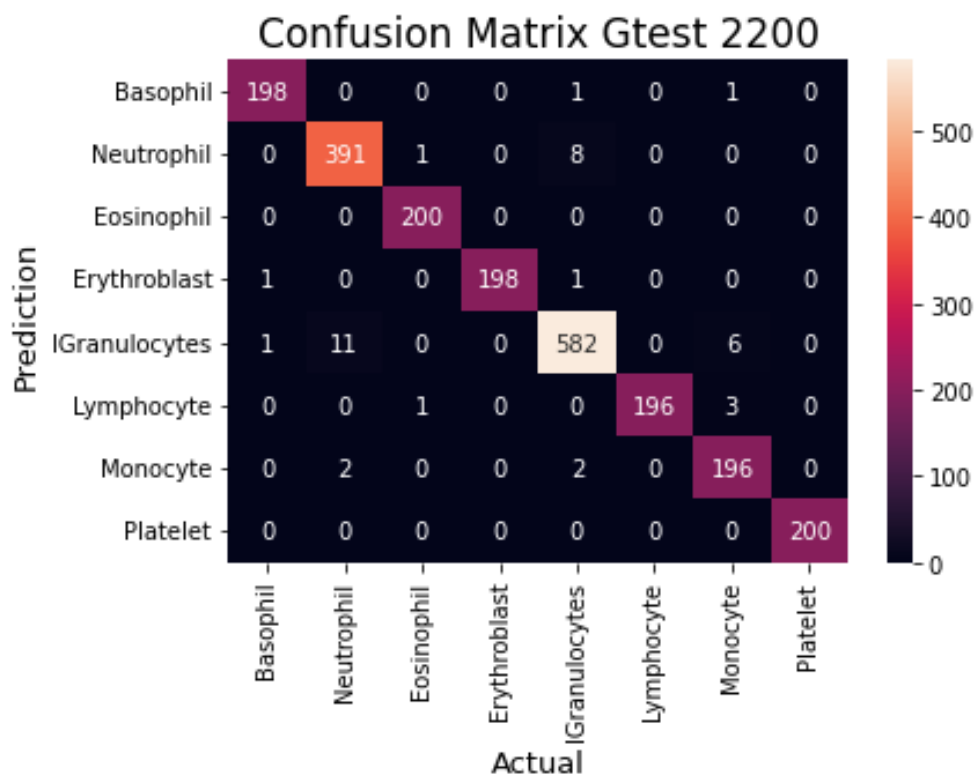


Figure 28: Figure 30 : Matrice de confusion du modèle Resnet-34 en 8 classes, sur l'échantillon Golden Test

	precision	recall	f1-score	support
Basophil	0.990	0.990	0.990	200
Neutrophil	0.968	0.978	0.973	400
Eosinophil	0.990	1.000	0.995	200
Erythroblast	1.000	0.990	0.995	200
IGranulocytes	0.980	0.970	0.975	600
Lymphocyte	1.000	0.980	0.990	200
Monocyte	0.951	0.980	0.966	200
Platelet	1.000	1.000	1.000	200
accuracy			0.982	2200
macro avg	0.985	0.986	0.985	2200
weighted avg	0.982	0.982	0.982	2200

Figure 29: Métriques qualifiant la qualité de la prédiction du modèle Resnet-34 en 8 classes, sur l'échantillon Golden Test

Modèle à 11 classes

Comparé au modèle à 8 classes, le modèle à 11 classes semble moins performant. On relève notamment un classement de certains myélocytes en leur forme immature, les promyélocytes, quand les métamyélocytes sont confondus avec leur propre formes immature, les myélocytes (et réciproquement). Notons que là encore, ces confusions sont tout à fait cohérentes dans une perspective d'hématopoïèse.

Le modèle a également des difficultés à distinguer entre les deux types de neutrophiles : mais là encore les neutrophiles à noyaux en bâtonnets sont les formes immatures des neutrophiles à noyaux segmentés.

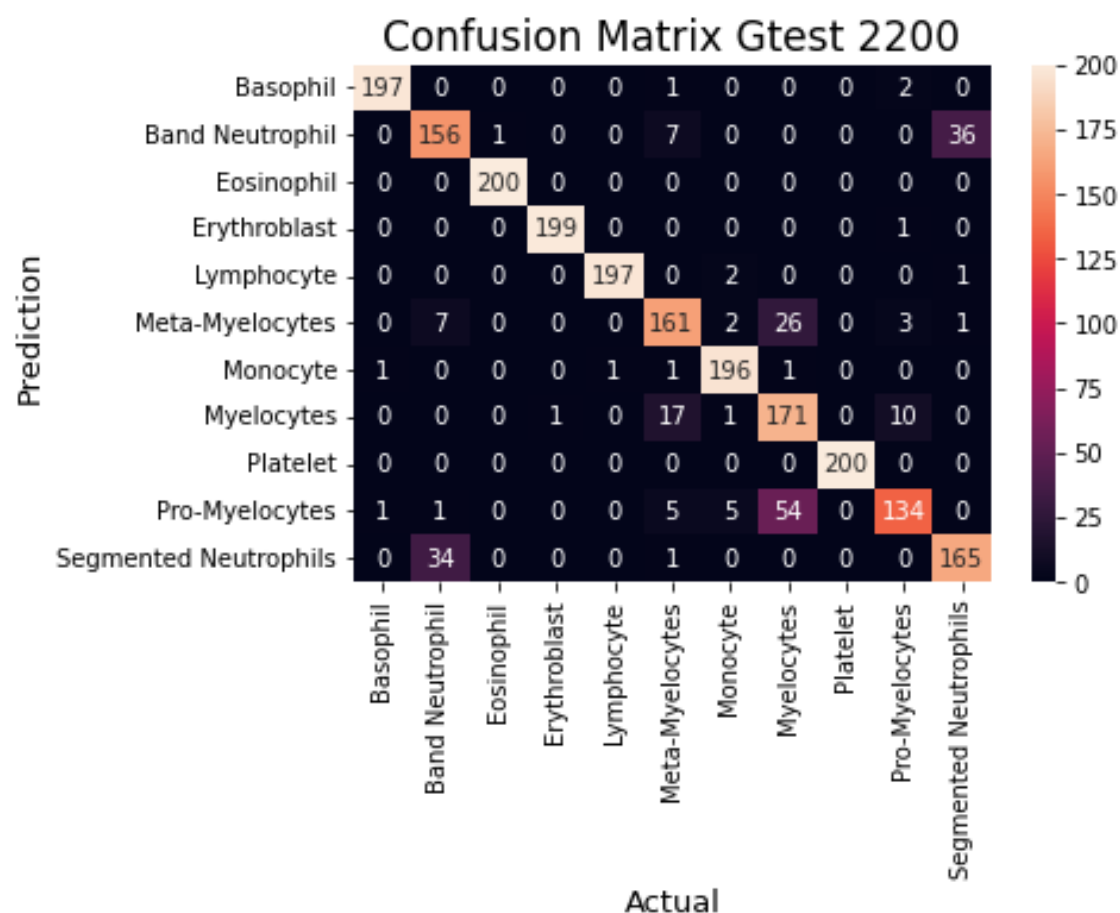


Figure 30: Matrice de confusion du modèle Resnet-34 en 11 classes, sur l'échantillon Golden Test

11 classes	precision	recall	f1-score	support
Basophil	0.990	0.985	0.987	200
Band Neutrophil	0.788	0.780	0.784	200
Eosinophil	0.995	1.000	0.998	200
Erythroblast	0.995	0.995	0.995	200
Lymphocyte	0.995	0.985	0.990	200
Meta-Myelocytes	0.834	0.805	0.819	200
Monocyte	0.951	0.980	0.966	200
Myelocytes	0.679	0.855	0.757	200
Platelet	1.000	1.000	1.000	200
Pro-Myelocytes	0.893	0.670	0.766	200
Segmented Neutrophils	0.813	0.825	0.819	200
accuracy			0.898	2200
macro avg	0.903	0.898	0.898	2200
weighted avg	0.903	0.898	0.898	2200

Figure 31: Métriques qualifiant la qualité de la prédiction du modèle Resnet-34 en 11 classes, sur l'échantillon Golden Test

- Interprétation des résultats et identification des domaines de réussite

S'il est attendu que les erreurs rencontrées de prédiction portent sur des catégories de cellules dont la proximité biologique se traduit par une similitude morphologique, le fait que ces erreurs concernent principalement des formes plus ou moins matures de même type de cellules nous amène à nous interroger sur le caractère continu ou discontinu des processus de transformation des formes cellulaires. Soit il existe des critères distincts permettant de qualifier la maturité d'une cellule sans ambiguïté, soit ce processus laisse place à une zone grise où la distinction du type de cellule, même par des experts, s'avère délicate et subjective¹⁵.

En d'autres termes, l'imprécision de notre modèle peut tenir à deux facteurs distincts : 1. une imperfection inhérente notre modèle de *Deep Learning* 2. une difficulté dans l'étiquetage préalable des cellules en catégories. Cette question est importante car elle détermine où il faudra concentrer ses efforts pour améliorer la prédiction du modèle : il serait vain de tester de nouveaux modèles de *Deep Learning* si l'ensemble des erreurs de classification pouvaient être attribués à des difficultés d'étiquetage.

- Suggestions d'améliorations et de travaux futurs

Afin d'éclaircir la question précédente¹⁶, il est tentant de regarder quelques cellules mal classées et de les mettre en regard de cellules des classes prédites et réelles, pour vérifier s'il est possible de les distinguer visuellement. Le représentation ci-dessous représente un exemple de classification ambiguë (les cellules choisies sont sélectionnées aléatoirement).

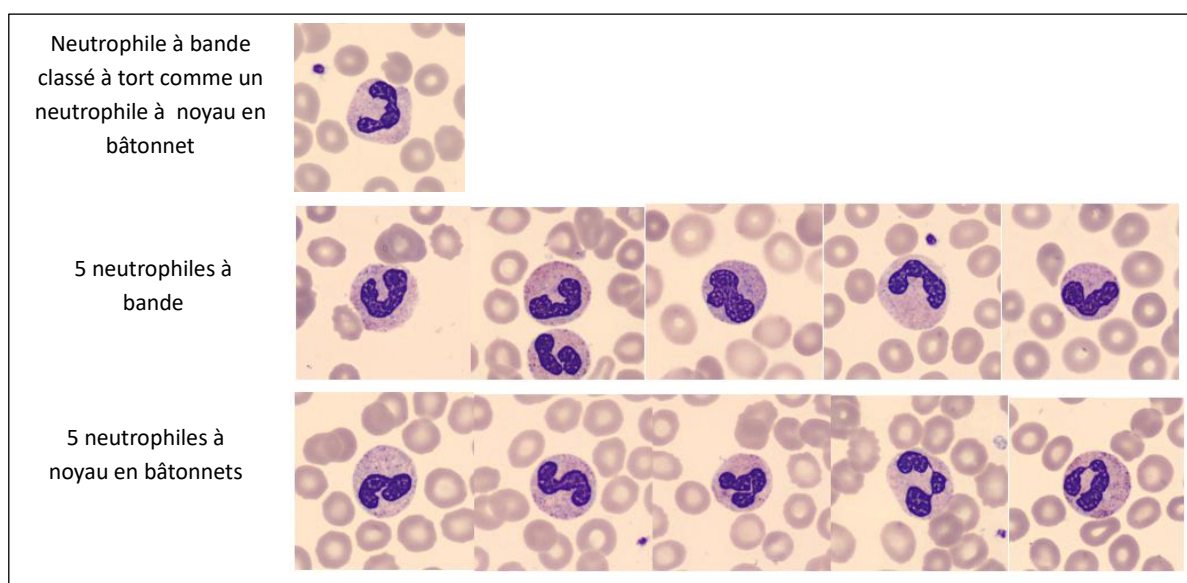


Figure 32: Comparaison d'une cellule mal classifiée avec des cellules de sa classe labélisée et des cellules de sa classe prédite

¹⁵ Osons une comparaison avec une catégorisation hypothétique de jeunes adultes en deux groupes : adolescents et adultes. Saurions-nous correctement les reconnaître à l'aide de photographies ? Certes on peut considérer que l'anniversaire des 18 ans constitue la limite objective du passage à l'âge adulte mais n'est-il pas plus pertinent de considérer que cette catégorisation a un caractère un peu artificiel et comporte une zone grise ?

¹⁶ Des exemples de masques de cellules mal classifiées sont donnés en annexe, afin de permettre au lecteur averti de forger sa propre opinion.

Cette approche visuelle pourrait être complétée par un avis d'experts sur la caractère continu ou discontinu des catégories de cellules. Néanmoins, établir avec certitude la part d'imprécision qui s'explique par un défaut de qualité du modèle ou une ambiguïté de l'étiquetage semble ardu. Nous suggérons que seul un étiquetage des mêmes images par plusieurs professionnels opérant de façon indépendantes permettrait d'éclairer la question, permettant de quantifier les cas de catégorisation divergente.

Outre la question de la qualité des libellés, il reste de nombreuses pistes d'amélioration du modèle. On peut essayer de régler le problème de surapprentissage en élargissant le nombre d'images en intégrant d'autres jeux de données ou en procédant à des augmentations cellulaires par rotation, translation et effet miroir¹⁷. Affiner les hyperparamètres du modèle constitue une autre piste d'amélioration potentielle, ainsi que la comparaison avec d'autres architectures de réseau de neurones : essayer les transformers ou pourquoi pas, de poursuivre plus avant une approche directe par U-Net faisant suite à notre segmentation en 9 classes, faisant l'économie d'une étape de classification (la classe étant obtenue par le segment le plus fréquent des pixels de la cellule) ; cette approche, à son plein potentiel, aurait comme préalable d'effectuer un plus grand nombre de masques de façon supervisé.

¹⁷ Les transformations de type zoom ou étirement des cellule si nous paraissent sensibles, susceptibles d'introduire des biais.

Conclusion

- Importance de ce travail dans le contexte médical et clinique

Les tâches de segmentation et de classification des cellules sont essentielles dans le domaine médical, en particulier en ce qui concerne le diagnostic, la recherche et le traitement de nombreuses maladies. Un modèle précis, comme celui que nous avons développé, peut révolutionner la manière dont les cliniciens évaluent et traitent les échantillons, en offrant des analyses plus rapides, plus objectives et potentiellement plus précises que les évaluations manuelles. De telles avancées ont le potentiel d'accélérer les diagnostics, d'identifier des sous-types cellulaires clés pour la recherche, et d'augmenter l'efficacité des interventions thérapeutiques.

- Potentiel d'utilisation du modèle dans des applications réelles

Compte tenu de leurs performances impressionnantes, ces modèles ont un potentiel énorme pour être intégrés dans des systèmes médicaux et cliniques réels. Ils pourraient être utilisés dans des laboratoires pour l'analyse automatisée d'échantillons, aidant ainsi à alléger la charge de travail des techniciens et pathologistes. De plus, avec une formation et une validation supplémentaire sur des ensembles de données plus vastes et diversifiés, ces modèles pourraient être adaptés à diverses applications médicales, de la recherche en laboratoire à la prise de décision clinique en temps réel. Leur adaptabilité et leur performance font de ces modèles des outils prometteurs pour la transformation numérique en cours dans le domaine de la santé.

Ce travail représente une avancée significative dans le domaine de la modélisation pour la segmentation et la classification cellulaire. Avec des améliorations continues et une validation appropriée, ces modèles peuvent jouer un rôle significatif dans l'avancement des soins médicaux et de la recherche biomédicale.

Table des illustrations

Figure 1 : Neutrophiles: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone.....	7
Figure 2 : Eosinophiles: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone.....	7
Figure 3 : Basophiles: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone.....	8
Figure 4 : Lymphocytes: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone.....	8
Figure 5 : Monocytes: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone.....	8
Figure 6 : Granulocytes immatures: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone.....	9
Figure 7 : Erythroblastes: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone.....	9
Figure 8 : Plaquettes: images extraites des collections du Laboratoire Central de l'Hôpital Clinic de Barcelone.....	9
Figure 9 : Concentration de chaque catégorie de cellule dans un microlitre de sang, pour un patient sain	10
Figure 10: Hématopoïèse	11
Figure 11: Distribution des catégories de cellules dans le jeu de données.....	14
Figure 12: Signature radiale d'une catégorie de cellule : les plaquettes sanguines	15
Figure 13: Histogramme de la distance des images à leur classe de rattachement, selon leur signature radiale	16
Figure 14: Un suspect démasqué : l'image comporte 3 cellules !	16
Figure 15: Représentation d'une convolution en coordonnées cartésienne et polaire	17
Figure 16: Chaîne de traitement.....	18
Figure 17: Schéma architectural de U-Net	20
Figure 18: Un exemple d'images segmentées en deux classes : la cellule d'intérêt et le fond de l'image	21
Figure 19: Processus d'augmentation d'images ; la cellule en haut à droite représente l'image initiales, les autres ses transformées.....	21
Figure 20: : Courbes d'apprentissage du modèle U-Net en 2 classes.....	23
Figure 21: Exemple de prédiction de U-Net sur une image de l'échantillon test comportant une seule cellule	24
Figure 22: Exemple de prédiction de U-Net sur une image de l'échantillon test comportant deux cellules.....	25
Figure 23: Etapes de nettoyage des images (doubles et débris)	26
Figure 24: Exemple de comblement des espaces vides dans les masques	26
Figure 25: Schéma architectural ResNet34 et comparaison avec deux autres architectures plus simples.....	28
Figure 26: Courbes d'apprentissage du modèle Resnet-34 en 8 classes	29
Figure 27: Courbes d'apprentissage du modèle Resnet-34 en 11 classes	30

Figure 28: Figure 30 : Matrice de confusion du modèle Resnet-34 en 8 classes, sur l'échantillon Golden Test.....	31
Figure 29: Métriques qualifiant la qualité de la prédiction du modèle Resnet-34 en 8 classes, sur l'échantillon Golden Test	31
Figure 30: Matrice de confusion du modèle Resnet-34 en 11 classes, sur l'échantillon Golden Test ...	32
Figure 31: Métriques qualifiant la qualité de la prédiction du modèle Resnet-34 en 11 classes, sur l'échantillon Golden Test	32
Figure 32: Comparaison d'une cellule mal classifiée avec des cellules de sa classe labélisée et des cellules de sa classe prédite	33
Figure 33: Signature radiale de l'ensemble des images du jeu de données.....	39
Figure 34: Signature radiale de l'ensemble des basophiles	39
Figure 35: Signature radiale de l'ensemble des éosinophiles	40
Figure 36: Signature radiale des érythroblastes.....	40
Figure 37: Signature radiale des granulocytes immatures	41
Figure 38: signature radiale des lymphocytes.....	41
Figure 39: signature radiale des monocytes.....	42
Figure 40: Signature radiale des neutrophiles.....	42
Figure 41: Signature radiale des plaquettes	43
Figure 42: Matrice de confusion de la classification basée sur la signature radiale des cellules : pour chaque cellule, la catégorie d'affectation est celle pour laquelle la distance à la signature radiale de la cellule est la plus courte	43
Figure 43: Spécification et résultats du modèle U-Net en 9 classes.....	44
Figure 44: Quelques cellules mal classées par le modèle Resnet-34 à 11 classes, avec leur catégorie prédite et leur catégorie labélisée.....	45
Figure 45: - Exemple de masques de cellules mal classées par le modèle Resnet-34 à 11 classes, avec mention de leur classe labélisée et de leur classe prédite.....	45

Bibliographie

- ACEVEDO, A., S. ALFÉREZ, A. MERINO, L. PUIGVÍ, AND J. RODELLAR. (2019): “Recognition of peripheral blood cell images using convolutional neural networks,” *Computer methods and programs in biomedicine*, 180, 105020.
- ACEVEDO, A., A. MERINO, S. ALFÉREZ, Á. MOLINA, L. BOLDÚ, AND J. RODELLAR. (2020): “A dataset of microscopic peripheral blood cell images for development of automatic recognition systems,” *Data in brief*, 30.
- BOLDÚ, L., A. MERINO, A. ACEVEDO, A. MOLINA, AND J. RODELLAR. (2021): “A deep learning model (ALNet) for the diagnosis of acute leukaemia lineage using peripheral blood cell images,” *Computer Methods and Programs in Biomedicine*, 202, 105999.
- REZATOFIGHI, S.H., SOLTANIAN-ZADEH, H.: Automatic recognition of five types of white blood cells in peripheral blood. *Computerized Medical Imaging and Graphics* 35 (4) (2011) 333-343.

Annexes

- Signature radiale de chaque catégorie de cellule

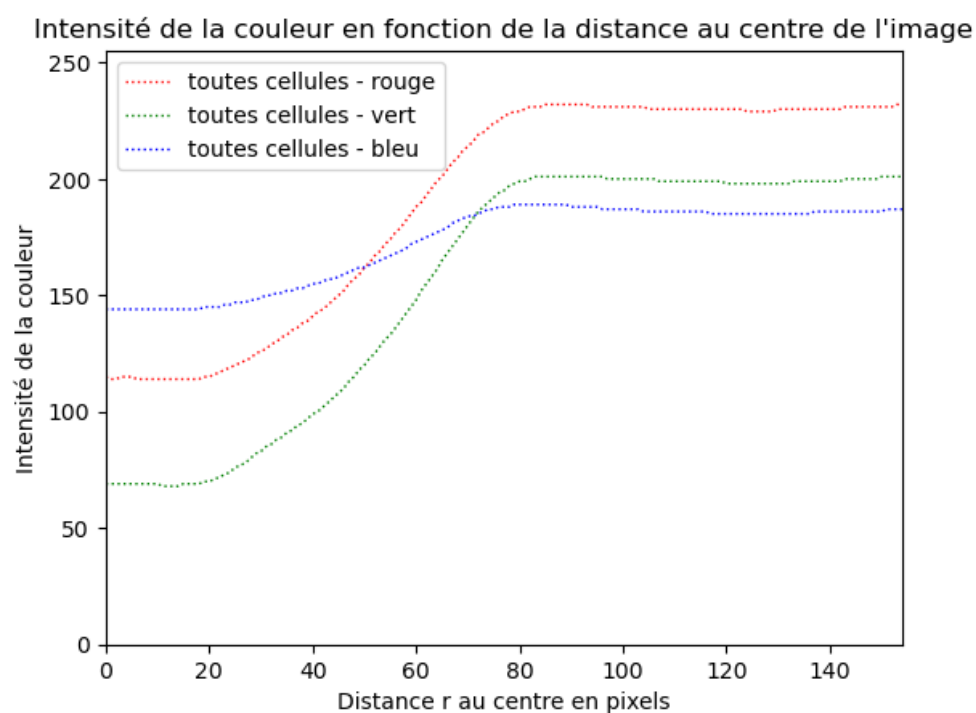


Figure 33: Signature radiale de l'ensemble des images du jeu de données

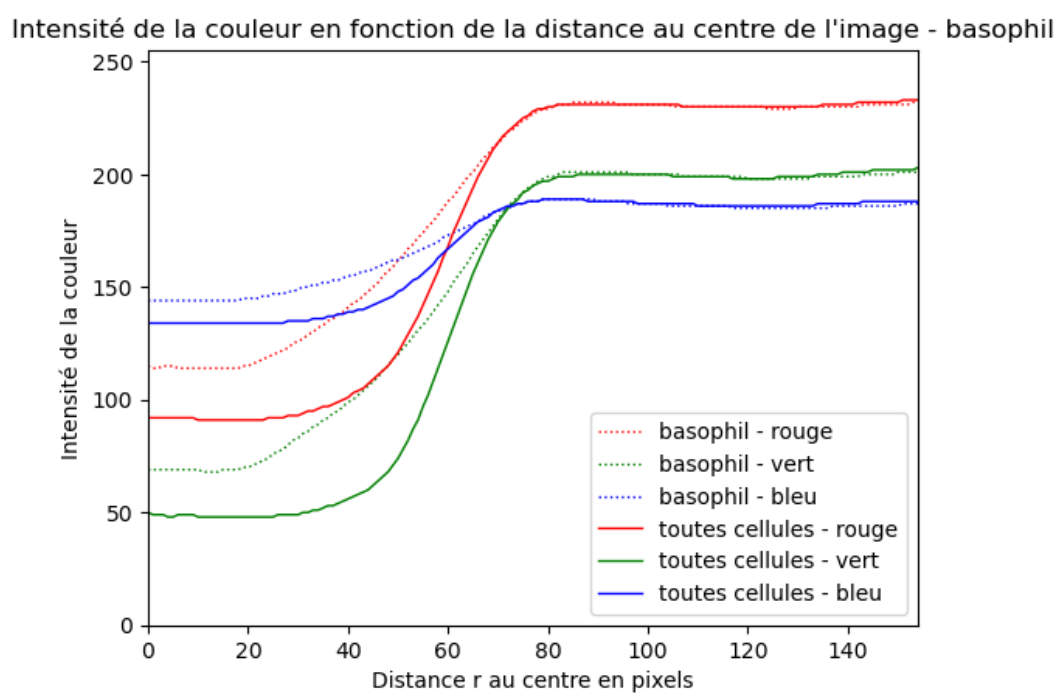


Figure 34: Signature radiale de l'ensemble des basophiles

Intensité de la couleur en fonction de la distance au centre de l'image - eosinophil

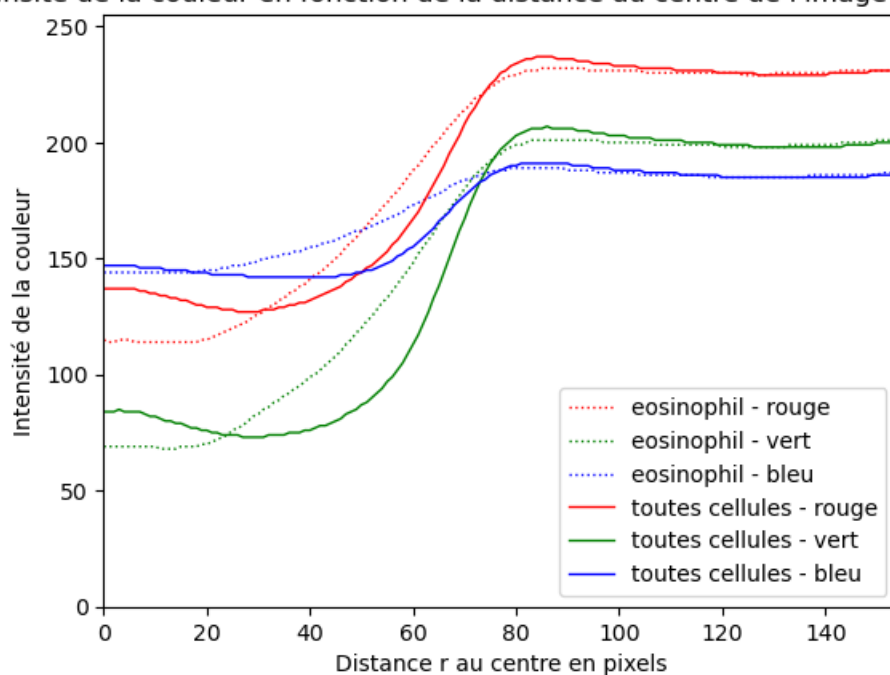


Figure 35: Signature radiale de l'ensemble des éosinophiles

Intensité de la couleur en fonction de la distance au centre de l'image - erythroblast

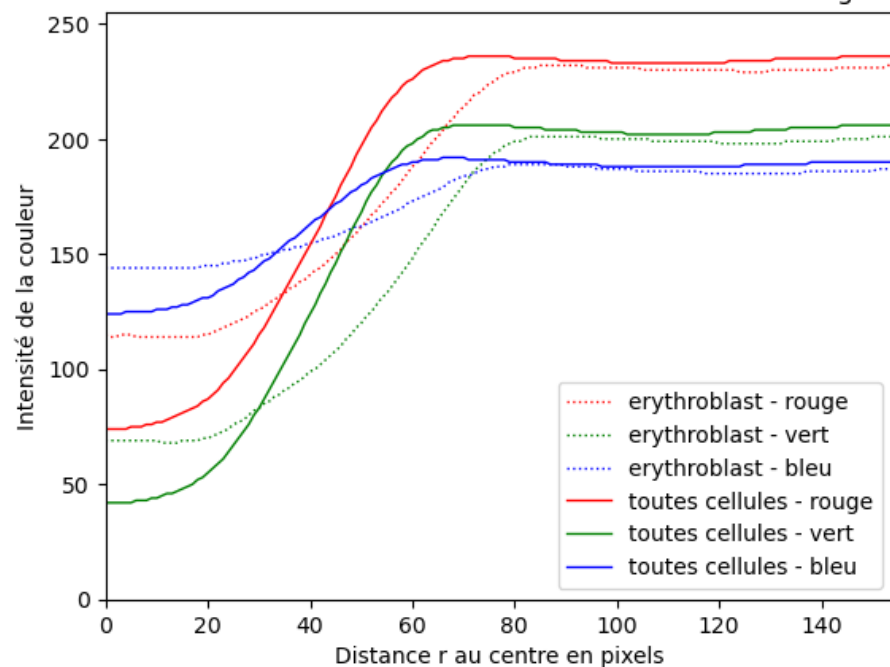


Figure 36: Signature radiale des érythroblastes

Intensité de la couleur en fonction de la distance au centre de l'image - ig

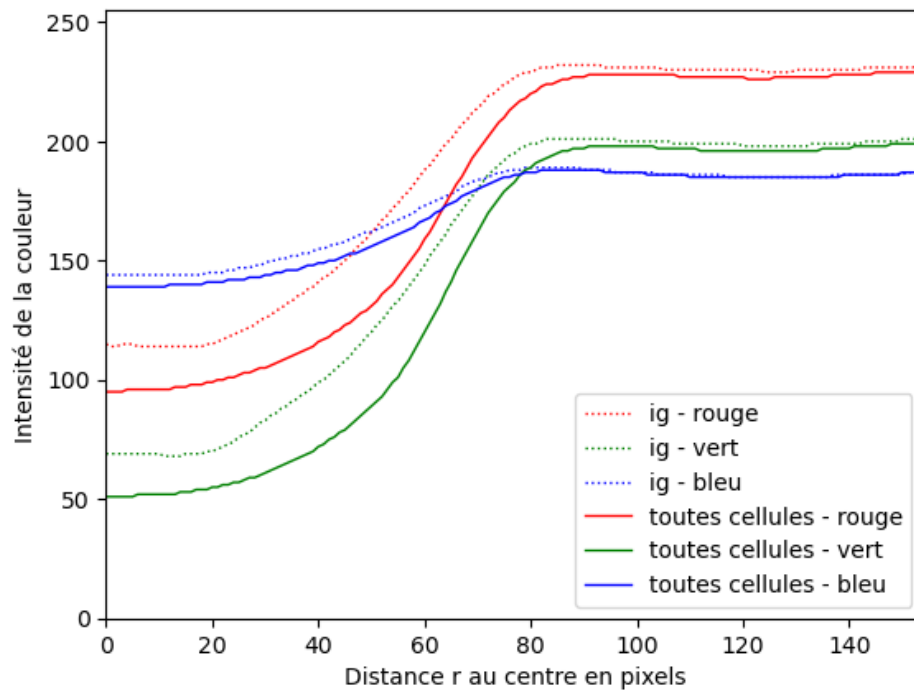


Figure 37: Signature radiale des granulocytes immatures

Intensité de la couleur en fonction de la distance au centre de l'image - lymphocyte

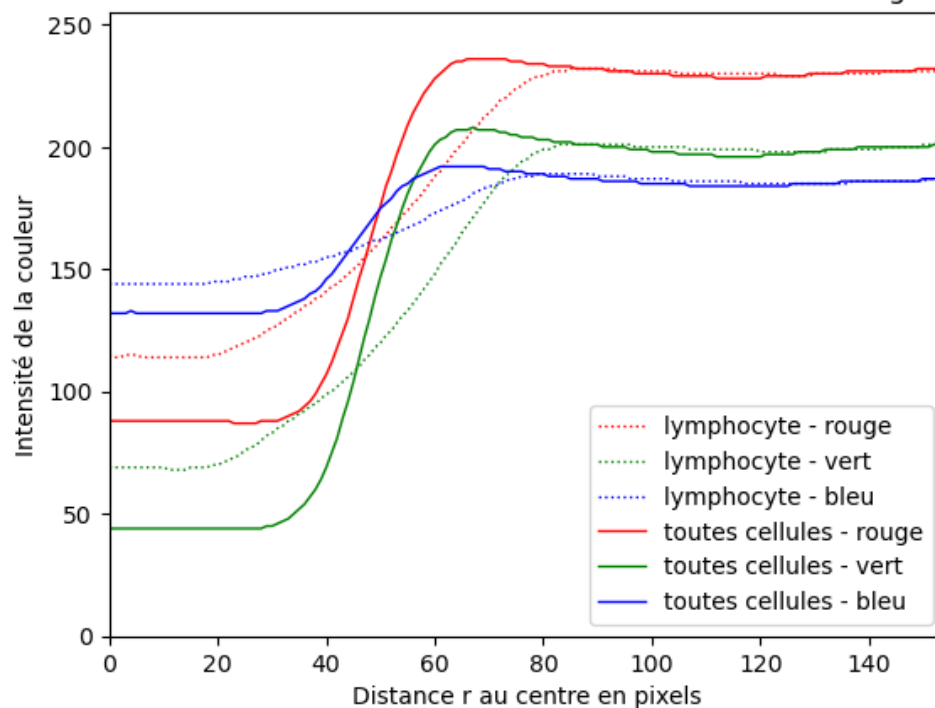


Figure 38: signature radiale des lymphocytes

Intensité de la couleur en fonction de la distance au centre de l'image - monocyte

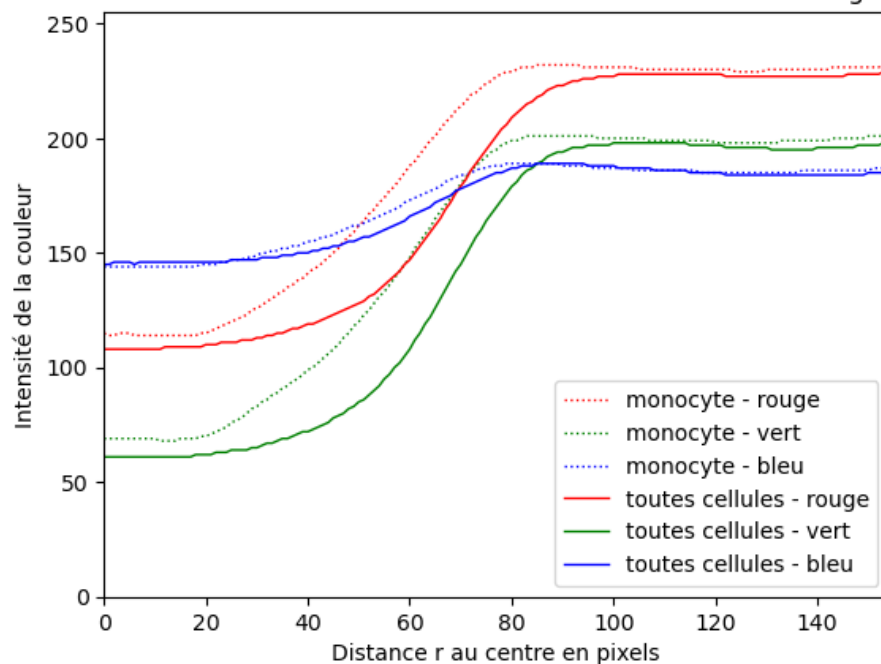


Figure 39: signature radiale des monocytes

Intensité de la couleur en fonction de la distance au centre de l'image - neutrophil

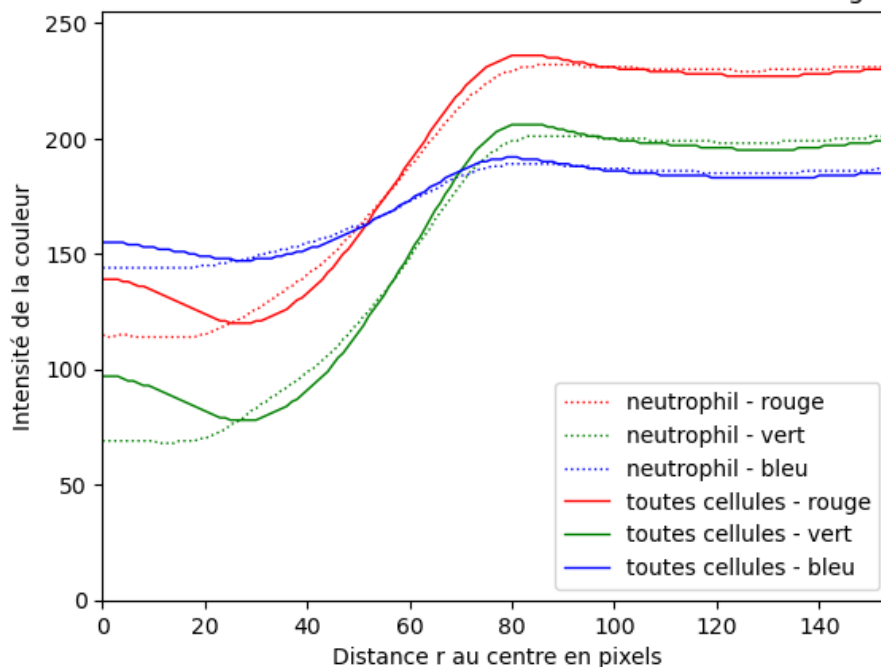


Figure 40: Signature radiale des neutrophiles

Intensité de la couleur en fonction de la distance au centre de l'image - platelet

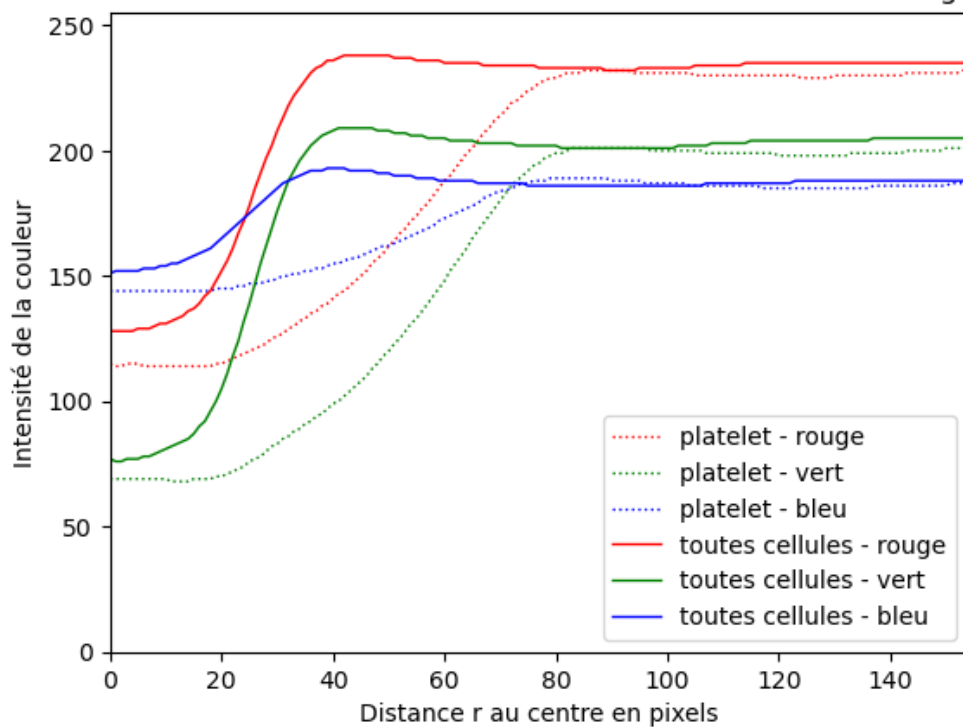


Figure 41: Signature radiale des plaquettes

- Matrice de confusion de la classification basée sur la signature radiale des cellules

Matrice de contingence									Notice
Catégorie prédite	0	1	2	3	4	5	6	7	
Catégorie réelle									
0	702	7	2	153	143	196	13	2	0: basophil
1	213	1861	23	306	59	286	356	13	1: eosinophil
2	109	5	1038	24	236	10	81	48	2: erythroblast
3	713	79	37	779	238	930	118	1	3: <u>ig</u>
4	115	1	135	23	926	2	10	2	4: lymphocyte
5	175	65	0	181	4	964	31	0	5: monocyte
6	36	299	319	299	188	55	2126	7	6: <u>neutrophil</u>
7	0	0	60	0	0	0	2	2286	7: <u>platelet</u>

Figure 42: Matrice de confusion de la classification basée sur la signature radiale des cellules : pour chaque cellule, la catégorie d'affectation est celle pour laquelle la distance à la signature radiale de la cellule est la plus courte

- Résultat de l'estimation de la segmentation U-Net en 9 classes

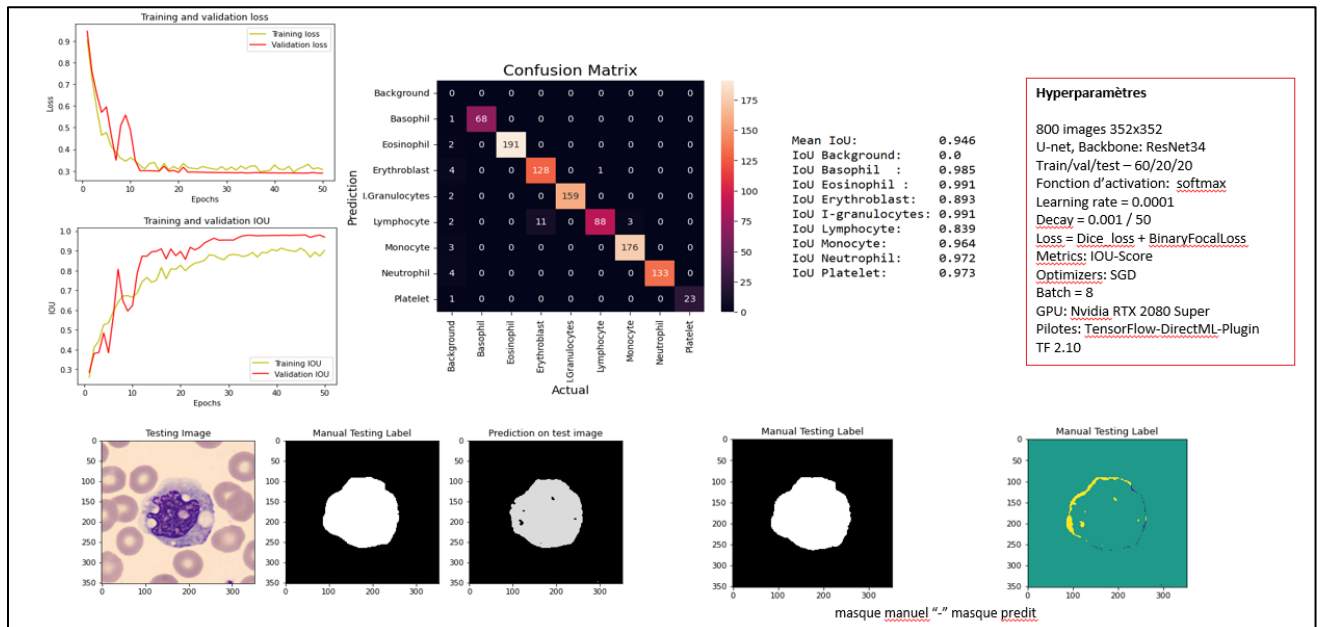
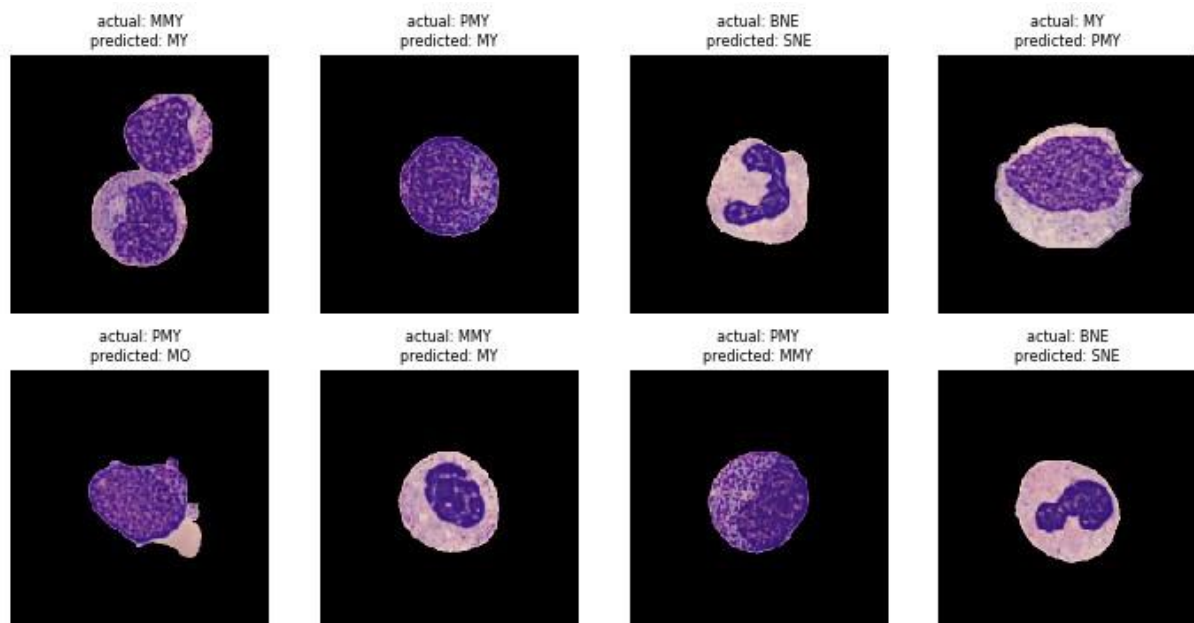


Figure 43: Spécification et résultats du modèle U-Net en 9 classes

- Exemple de cellules mal classées par le modèle Resnet-34 à 11 classes



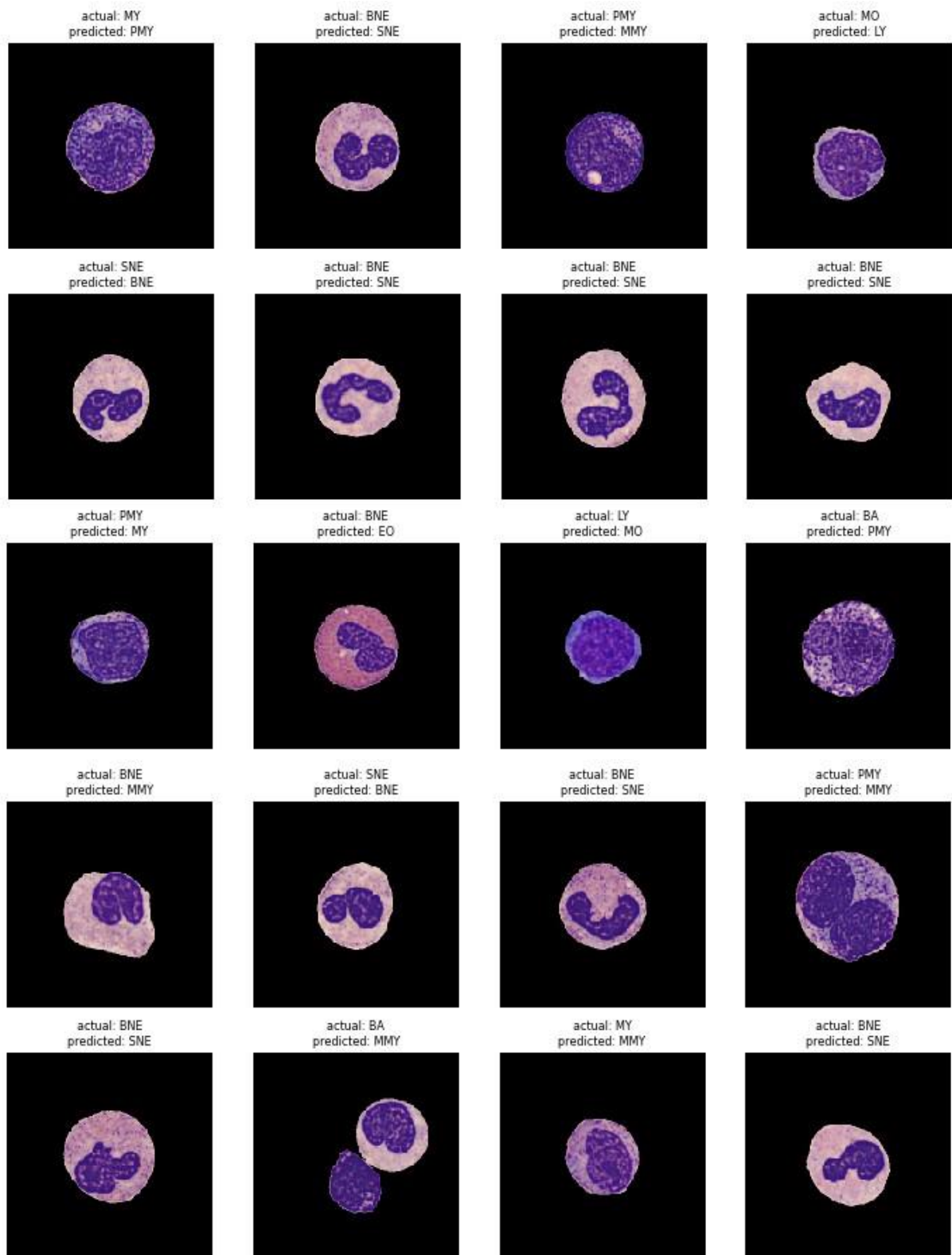


Figure 45: - Exemple de masques de cellules mal classées par le modèle Resnet-34 à 11 classes, avec mention de leur classe labélisée et de leur classe prédite