

RAPPORT DE PROJET

Détection Automatisée des Cellules Cancéreuses dans le Sang



DEPARTMENT INFORMATIQUE
MODULE TRAITEMENT D'IMAGE ET VISION PAR ORDINATEUR

Réalisé par

Imad OISSAFE

Encadré par

Mme. Laila AMIR

MAI 2024

Table des matières

0.1	Introduction	1
0.2	Jeu de données utilisée	1
0.3	Approche Adoptée	2
0.3.1	Diagramme de flux de travail de la méthodologie proposée	2
0.3.2	Fonctionnement du modèle proposé	2
0.4	Résultats expérimentales	8
0.4.1	Résultats générales	8
0.4.2	Modèle KNN	8
0.4.3	Modèle SVM, Arbre de décision Et Forêt aléatoire	9
0.5	Défis rencontrés	9
0.5.1	Sélection de paramètres pour les algorithmes de traitement d'image	9
0.5.2	Déséquilibre des données	10
0.5.3	Temps de calcul	10
0.6	Conclusion et Travaux Futurs	10

Liste des figures

1	Les quatre catégories des images	1
2	Flux de travail	2
3	Comparaison entre segmentation par seuillage et par clustering	2
4	Les étapes de segmentation	3
5	Avant appliquer la segmentation par watershed	5
6	Après appliquer la segmentation par watershed	5
7	Données normalisées	5
8	Matrice de corrélation	6
9	Caractéristiques sélectionnées	6
10	Rapport de classification pour KNN	8
11	Matrice de confusion pour le modèle knn	8
12	Rapport de classification pour les 3 modèles	9
13	Matrice de confusion pour les 3 modèles	9

Liste des tableaux

1	Caractéristiques des cellules regroupées par nature	4
2	Précision des modèles	8

0.1 Introduction

Le cancer du sang, également connu sous le nom de leucémie, constitue l'un des types de cancer les plus répandus à travers le monde, touchant des millions de personnes chaque année. Un diagnostic précoce et précis de cette maladie revêt une importance cruciale pour garantir un traitement efficace et des résultats cliniques optimaux. Dans ce contexte, les récentes avancées dans le domaine de l'intelligence artificielle, notamment en vision par ordinateur et en apprentissage automatique, ouvrent de nouvelles perspectives pour la détection et la caractérisation des cellules cancéreuses dans les échantillons sanguins.

Ce projet a pour objectif de développer un système automatisé permettant la détection et la caractérisation des cellules cancéreuses dans les échantillons sanguins, afin de déterminer le stade du cancer. Cette initiative repose sur l'utilisation de techniques avancées de traitement d'images et d'apprentissage automatique. Elle revêt une importance cruciale pour un diagnostic précoce du cancer, notamment la leucémie lymphoblastique aiguë (ALL), afin d'améliorer les chances de traitement et les résultats cliniques pour les patients.

0.2 Jeu de données utilisée

Nous avons utilisé un ensemble de données disponible sur Kaggle, intitulé "Blood Cells Cancer (ALL)". Les images de ce jeu de données ont été préparées au laboratoire de la moelle osseuse de l'hôpital Taleqani à Téhéran, en Iran. Ce jeu de données se compose de 3242 images de frottis sanguins périphériques (FSP) provenant de 89 patients suspects de leucémie lymphoblastique aiguë. Les échantillons sanguins ont été préparés et colorés par du personnel de laboratoire qualifié. L'ensemble de données se divise en deux classes : bénigne et maligne. La première classe comprend des images hématogènes, tandis que la seconde classe représente le groupe de LLA, avec trois sous-types de lymphoblastes malins : Early Précoce Pre-B, Pre-B et Pro-B. Toutes les images ont été capturées à l'aide d'un appareil photo Zeiss monté sur un microscope avec un grossissement de 100x et ont été enregistrées au format JPG. La détermination définitive des types et sous-types de ces cellules a été effectuée par un spécialiste à l'aide de l'outil de cytométrie en flux.

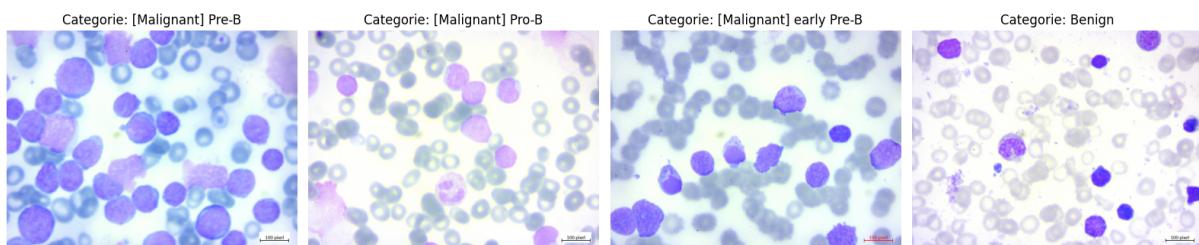


Figure 1: Les quatre catégories des images

0.3 Approche Adoptée

0.3.1 Diagramme de flux de travail de la méthodologie proposée

Pour réaliser une segmentation efficace de nos images, nous avons mis en œuvre le processus suivant :

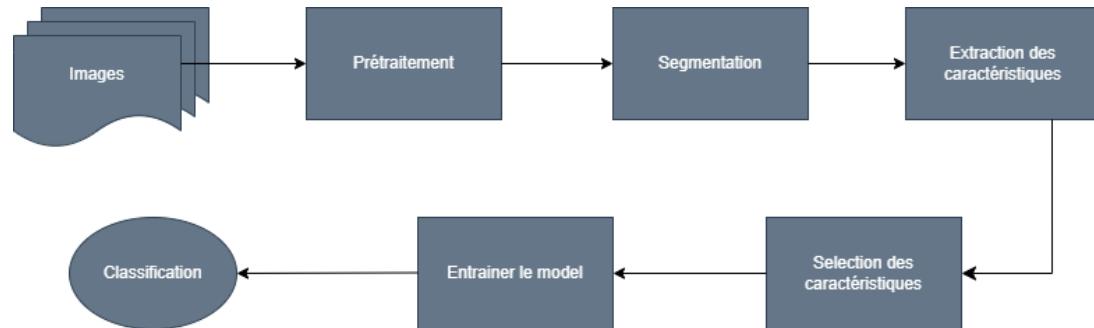


Figure 2: Flux de travail

0.3.2 Fonctionnement du modèle proposé

Prétraitement

Durant cette phase, nous avons redimensionné les images et les avons renommées en fonction des catégories correspondantes.

Pour réduire le temps de calcul, l'image RGB originale (768 par 1024 pixels) a été réduite à 224 par 224 pixels.

Segmentation

La segmentation d'image est le processus d'extraction de la région d'intérêt de l'image, puisque la leucémie affecte les globules blancs. Dans cette recherche, les algorithmes de segmentation K-means et de segmentation par seuillage binaire ont été appliqués pour segmenter avec succès la région d'intérêt.

L'algorithme de clustering K-means a segmenté la région d'intérêt et l'a séparée des autres cellules sanguines. Le choix du K-means s'explique par sa performance par rapport à la méthode de seuillage.

Voici un exemple d'une image segmenté utilisant k-means et seuillage:

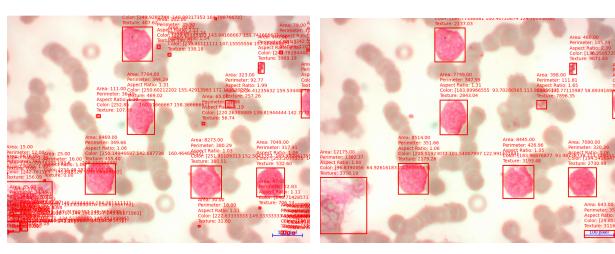


Figure 3: Comparaison entre segmentation par seuillage et par clustering

Nous avons noté que le processus de seuillage est quelque peu sensible au bruit, ce qui peut avoir un impact sur les résultats.

Les opérations morphologiques, telles que le remplissage des trous, ont été employées avec succès pour appliquer l'algorithme k-means. En particulier, le remplissage des trous a été ciblé pour améliorer la précision du système dans la détection de la leucémie.

- a - Convertir l'image RGB en espace couleur LAB.
- b - Extraire le canal A de l'image LAB.
- c - Appliquer le regroupement k-means sur le canal A.
- d - Appliquer un seuillage binaire sur le canal A.
- e - Remplir les trous dans l'image binaire.
- f - Supprimer les petits objets et les petits trous de l'image binaire.

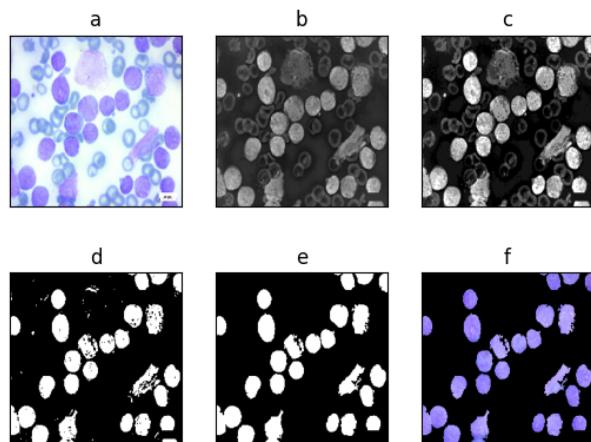


Figure 4: Les étapes de segmentation

Extraction des caractéristiques

Dans cette étape, différentes caractéristiques sont extraites.

Table 1: Caractéristiques des cellules regroupées par nature

Nature	Caractéristiques
Morphologiques	Solidité : Le rapport entre la surface de la cellule et la surface de l'enveloppe convexe de la cellule. Excentricité : L'excentricité de la cellule. Orientation : L'angle entre l'axe majeur de la cellule et l'axe des x. Diamètre équivalent : Le diamètre d'un cercle ayant la même surface que la cellule.
Géométriques	Surface : La surface de la cellule. Périmètre : Le périmètre de la cellule. Longueur de l'axe majeur : La longueur de l'axe majeur de la cellule. Longueur de l'axe mineur : La longueur de l'axe mineur de la cellule. Ratio d'aspect : Le rapport entre la longueur de l'axe majeur et la longueur de l'axe mineur de la cellule. Étendue : Le rapport entre la surface de la cellule et la surface de la boîte englobante de la cellule. Centroïde : Les coordonnées du centroïde de la cellule. Boîte englobante : Les coordonnées de la boîte englobante de la cellule. Surface convexe : La surface de l'enveloppe convexe de la cellule. Circularité : Le rapport entre le périmètre de la cellule et le périmètre d'un cercle ayant la même surface que la cellule.
Texture	Contraste : Le contraste de la cellule. Corrélation : La corrélation de la cellule. Énergie : L'énergie de la cellule.
Intensité	Intensité moyenne : L'intensité moyenne de la cellule. Écart type de l'intensité : L'écart type de l'intensité de la cellule. Intensité médiane : L'intensité médiane de la cellule.
Couleur	Couleur : La couleur en RGB de la cellule.

En complément des caractéristiques déjà extraites, nous avons intégré la moyenne des cellules dans chaque image. Lors de l'analyse visuelle de plusieurs images, nous avons observé que les catégories telles que Pre-B et Early-Pre-B présentent un nombre plus élevé de cellules malignes par rapport aux catégories Benign et Pro-B. C'est pourquoi nous avons jugé pertinent d'inclure cette nouvelle caractéristique.

Dans certains cas, des cellules voisines peuvent être détectées comme une seule cellule en raison de leur proximité ou de leur chevauchement. Pour résoudre ce problème, nous avons tenté d'appliquer un autre algorithme de segmentation; Watershed pour la segmentation des cellules individuelles.

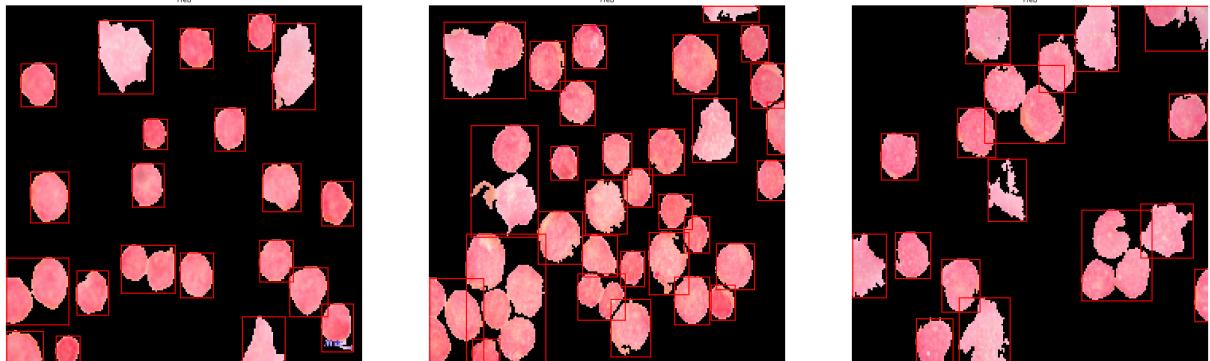


Figure 5: Avant appliquer la segmentation par watershed

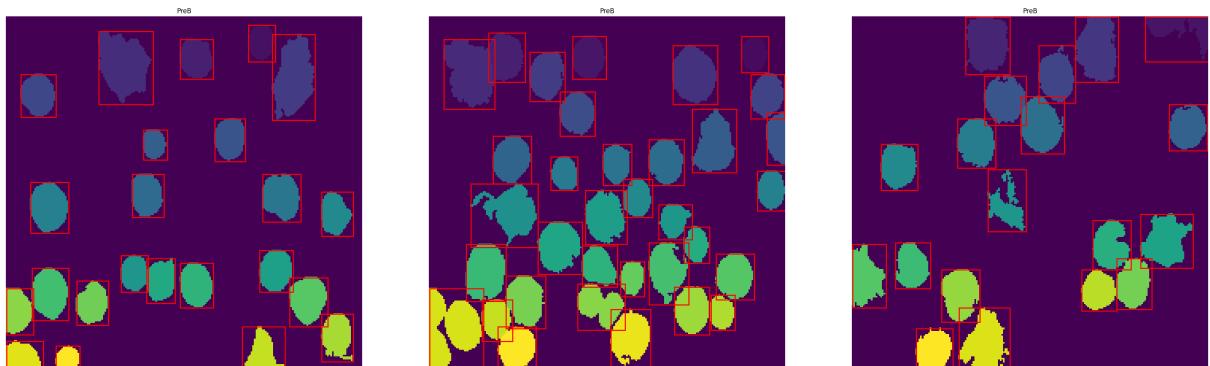


Figure 6: Après appliquer la segmentation par watershed

Une fois les caractéristiques des cellules segmentées extraites, nous les avons stockées dans une dataframe. Après diverses transformations des données, nous avons augmenté le nombre de caractéristiques à 24. Pour garantir une mise à l'échelle correcte des données, nous avons inclus une étape de normalisation afin que les caractéristiques soient toutes sur la même échelle et que le modèle puisse apprendre de manière plus efficace, sans être biaisé par des valeurs disproportionnées.

	solidity	eccentricity	orientation	equivalent_diameter	contrast	mean_intensity	median_intensity	std_intensity	correlation	energy	...	area	perimeter	aspect_ratio	R	G
0.710192	0.616203	-0.073933		-0.826363	-1.098226	0.633933	-0.369388	-1.572539	-0.772115	-0.751116	...	-0.815151	-0.887876	-0.056729	1.593164	0.319226
0.696314	-0.671994	0.119467		-0.014285	-0.043772	0.306969	0.055086	-0.338616	-0.525104	-0.614769	...	-0.121051	-0.378616	-0.056729	0.553944	0.370636
0.807547	-1.798842	0.151161		0.470919	-0.936581	0.346409	-0.182120	-1.398977	-1.014359	-0.839517	...	0.361374	-0.108010	-0.056729	0.912841	0.014458
0.501842	-0.008051	-0.007568		0.380704	-0.352975	-0.182289	-0.344419	-0.333839	-0.578055	-0.573808	...	0.267843	0.011317	-0.056729	0.416381	-0.327739
0.450392	0.596042	0.263748		-0.419732	-0.666970	-0.881415	-0.868768	-0.097066	0.040180	-0.107307	...	-0.485330	-0.525215	-0.056729	0.058573	-1.117599

Figure 7: Données normalisées

Sélection des caractéristiques

Lors de l'analyse de la matrice de confusion, nous avons remarqué une forte corrélation entre certaines caractéristiques, ce qui indique une possible redondance dans les données. Cela peut affecter la performance du modèle et nécessite une attention particulière lors de la sélection des caractéristiques pour éviter le surapprentissage et améliorer la généralisation du modèle. Dans notre processus d'analyse, nous avons cherché à identifier les

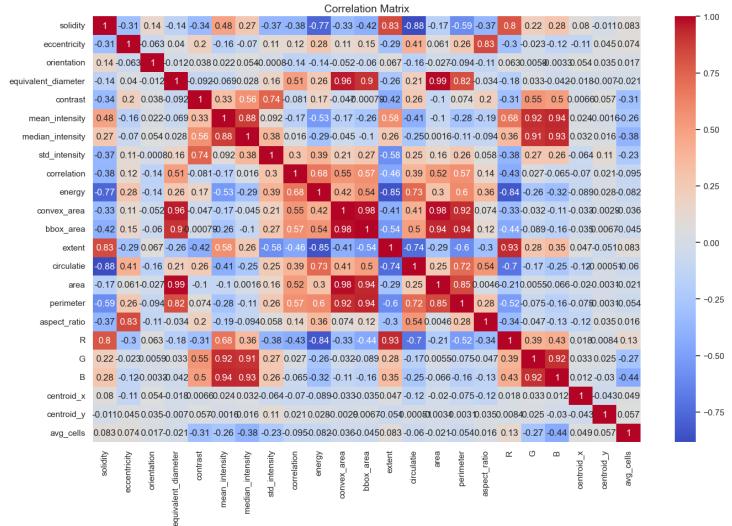


Figure 8: Matrice de corrélation

caractéristiques les plus pertinentes pour notre modèle. Pour ce faire, nous avons utilisé une technique appelée Recursive Feature Elimination (RFE), ou Élimination Récursive de Caractéristiques en français.

L'idée derrière la RFE est de sélectionner un modèle initial (dans notre cas, une régression logistique) et d'itérer sur les différentes combinaisons de caractéristiques en éliminant les moins importantes à chaque étape. À chaque itération, le modèle est entraîné sur le sous-ensemble actuel de caractéristiques et les caractéristiques les moins importantes sont éliminées, jusqu'à ce qu'un nombre prédéfini de caractéristiques soit atteint.

Dans notre approche, nous avons employé la RFE avec une régression logistique comme modèle initial. Cette démarche implique d'utiliser la régression logistique pour évaluer l'importance des diverses caractéristiques et sélectionner celles qui ont le plus d'impact sur la performance du modèle. Nous avons restreint cette sélection aux cinq caractéristiques les plus pertinentes pour la tâche de modélisation que nous avons entreprise.

equivalent_diameter	area	R	B	avg_cells
-0.826363	-0.815151	1.593164	-0.159095	1.71058
-0.014285	-0.121051	0.553944	-0.067298	1.71058
0.470919	0.361374	0.912841	0.008829	1.71058
0.380704	0.267843	0.416381	-0.517054	1.71058
-0.419732	-0.485330	0.058573	-1.192177	1.71058

Figure 9: Caractéristiques sélectionnées

Entraînement du modèle

Pour la phase d'entraînement, nous avons exploré plusieurs algorithmes afin de sélectionner le meilleur modèle pour notre projet. Voici une brève description des algorithmes que nous avons testés :

- ****k-nearest neighbor (knn)**** : Le k-nearest neighbor est un algorithme de classification qui attribue une étiquette à un point de données en se basant sur les étiquettes des points voisins les plus proches. Il fonctionne en mesurant la distance entre les points de données et en choisissant les k voisins les plus proches pour effectuer une prédiction.
- ****SVM (Support Vector Machine)**** : Les machines à vecteurs de support sont des algorithmes de classification qui trouvent l'hyperplan optimal pour séparer les données en deux classes. Ils fonctionnent en trouvant l'hyperplan qui maximise la marge entre les différentes classes, ce qui les rend efficaces pour les tâches de classification avec des données de grande dimension.
- ****Arbre de décision**** : L'arbre de décision est un algorithme de classification utilisé pour séparer les données en sous-groupes homogènes en fonction des caractéristiques. Il fonctionne en partitionnant récursivement l'espace des caractéristiques en sous-ensembles, en choisissant à chaque étape la caractéristique qui permet de diviser au mieux les données. Cette approche permet de construire un arbre où chaque noeud représente une décision basée sur une caractéristique, et chaque feuille représente une classe ou une valeur de sortie.
- ****Forêts aléatoires**** : Les forêts aléatoires sont un ensemble d'arbres de décision utilisés pour la classification ou la régression. Elles fonctionnent en combinant les prédictions de plusieurs arbres de décision entraînés sur des sous-ensembles aléatoires des données d'entraînement. Cela permet de réduire le sur-ajustement et d'améliorer la précision des prédictions.

Après avoir testé ces différents algorithmes, nous avons sélectionné celui qui présentait les meilleures performances pour notre problème spécifique.

0.4 Résultats expérimentales

0.4.1 Résultats générales

La table ci-dessous décrit les résultats des métriques de précision pour chaque modèle.

Table 2: Précision des modèles

Modèle	KNN	SVM	RandomForest	DecisionTree
Précision (%)	96.2	100	100	100

0.4.2 Modèle KNN

	precision	recall	f1-score	support
EarlyPreB	0.94	0.92	0.93	776
PreB	1.00	1.00	1.00	776
ProB	0.99	0.99	0.99	776
benign	0.92	0.94	0.93	776
accuracy			0.96	3104

Figure 10: Rapport de classification pour KNN

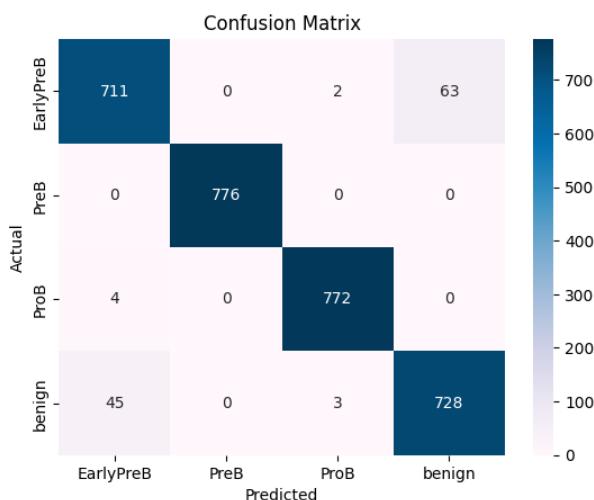


Figure 11: Matrice de confusion pour le modèle knn

Ces résultats révèlent que le modèle rencontre des difficultés à distinguer entre la classe EarlyPreB et la classe Benign.

0.4.3 Modèle SVM, Arbre de décision Et Forêt aléatoire

Pour les modèles SVM, arbre de décision et forêt aléatoire, ils présentent les mêmes résultats en termes de rapport de classification ainsi que de matrice de confusion.

	precision	recall	f1-score	support
EarlyPreB	1.00	1.00	1.00	776
PreB	1.00	1.00	1.00	776
ProB	1.00	1.00	1.00	776
benign	1.00	1.00	1.00	776
accuracy			1.00	3104

Figure 12: Rapport de classification pour les 3 modèles

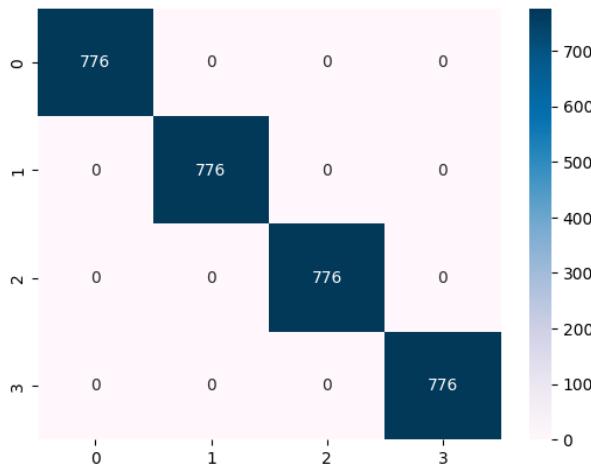


Figure 13: Matrice de confusion pour les 3 modèles

Il semble que les trois modèles aient réussi à classifier les classes avec succès.

0.5 Défis rencontrés

0.5.1 Sélection de paramètres pour les algorithmes de traitement d'image

La sélection de paramètres pour les algorithmes de traitement d'image a été un défi majeur dans notre projet. Trouver les valeurs optimales pour les paramètres tels que le seuil de segmentation, la taille du noyau pour les opérations morphologiques a nécessité une exploration minutieuse et parfois coûteuse en termes de temps de calcul. Avec nos modestes connaissances, nous avons entrepris de tester manuellement diverses techniques et paramètres, en nous guidant par la visualisation des résultats obtenus. Cette démarche expérimentale nous a aidés à identifier les méthodes les plus adaptées à chaque étape du traitement d'image, ce qui a conduit à une amélioration de la qualité de la segmentation et de la classification des cellules sanguines.

0.5.2 Déséquilibre des données

Le déséquilibre des données, où une classe est représentée par un grand nombre d'échantillons par rapport à d'autres classes, a été un défi important. Cela peut entraîner un biais dans les performances du modèle, où il peut être plus enclin à prédire la classe majoritaire. Pour remédier à cela, nous avons tenté de réduire la taille des classes majoritaires pour les aligner avec la taille de la classe minoritaire.

0.5.3 Temps de calcul

Le temps de calcul était également un défi, en particulier lors de l'entraînement de modèles sur de grands ensembles de données ou lors de l'exploration de nombreux paramètres pour trouver les meilleures performances.

0.6 Conclusion et Travaux Futurs

Dans ce projet, nous avons abordé la problématique de la segmentation et de la classification d'images de cellules sanguines, une tâche essentielle dans de nombreux domaines médicaux et de recherche. Notre objectif était de développer des méthodes efficaces pour automatiser ce processus, ce qui pourrait permettre une analyse rapide et précise des échantillons sanguins.

Les résultats obtenus ont montré une segmentation efficace des cellules sanguines, ainsi qu'une classification précise des différents types de cellules. Malgré les défis rencontrés, notre approche a permis d'obtenir des performances satisfaisantes, ouvrant la voie à de futures améliorations et applications dans le domaine médical et de la recherche biomédicale.

Pour les travaux futurs, plusieurs axes de développement sont envisageables. Tout d'abord, nous envisageons d'explorer davantage les techniques d'apprentissage profond (deep learning) pour améliorer la précision de la segmentation et de la classification des cellules. Les réseaux de neurones convolutionnels (CNN) sont particulièrement adaptés à cette tâche et pourraient offrir des performances supérieures aux méthodes traditionnelles.

Enfin, nous envisageons d'explorer des techniques de visualisation avancées pour mieux comprendre les caractéristiques discriminantes des différentes classes de cellules sanguines, ce qui pourrait aider à améliorer la précision de nos modèles et à fournir des informations cliniquement pertinentes aux professionnels de la santé.