

Détection Automatisée des Cellules Cancéreuses dans le Sang

Imad OISSAFE

*Module Traitement d'image et vision par ordinateur
Master IAII*

Programme

1 Introduction

2 Jeu de données

3 Approche Adoptée

4 Résultats expérimentales

5 Défis rencontrés

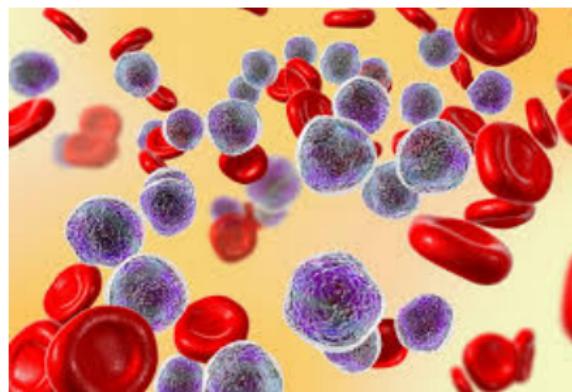
6 Conclusion et Travaux Futurs

Introduction

Problématique

La leucémie est une forme de cancer du sang qui affecte les globules blancs et constitue l'une des principales causes de décès chez les humains. Actuellement, le diagnostic de la leucémie se fait par inspection visuelle des images microscopiques de cellules sanguines, ce qui est long, fastidieux et nécessite des experts humains formés. Par conséquent, l'absence d'un système de détection automatique, précoce et efficace de la leucémie est un grand défi.

Un des types principaux de ce cancer est la leucémie lymphoïde aiguë (LLA). Elle affecte les lymphocytes, un type de globule blanc. Elle progresse rapidement et nécessite un traitement urgent.



Objectif

L'objectif principal de ce travail est de développer un système de détection précoce et de classification automatique pour diagnostiquer la leucémie LLA à partir d'images sanguines en utilisant des algorithmes d'apprentissage automatique et de traitement d'images.

Jeu de données

Jeu de données

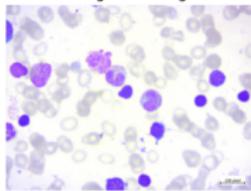
Le jeu de données utilisé pour ce travail est disponible sur Kaggle, intitulé "Blood Cells Cancer (ALL)". Ce jeu de données se compose de 3242 images provenant de 89 patients suspects de leucémie lymphoblastique aiguë. Les échantillons sanguins ont été préparés et colorés par du personnel de laboratoire qualifié. L'ensemble de données se divise en 4 classes :

- Benign : signifie que les cellules sont normales.
 - Early Pre-B : signifie que les cellules sont dans la phase précoce de la différenciation B.
 - Pre-B : signifie que les cellules sont dans la phase intermédiaire de la différenciation B.
 - Pro-B : signifie que les cellules sont dans la phase finale de la différenciation B.

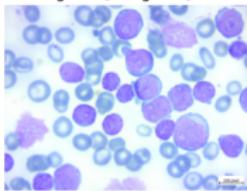
. Toutes les images ont été capturées à l'aide d'un appareil photo Zeiss monté sur un microscope avec un grossissement de 100x et ont été enregistrées au format JPG.

Exemples des classes

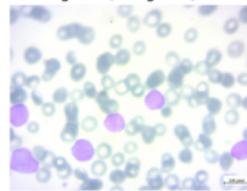
Categorie: Benign



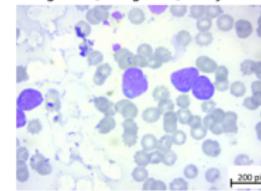
Categorie: [Malignant] Pre-B



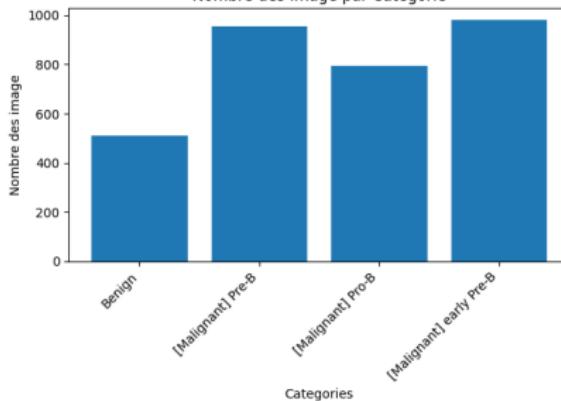
Categorie: [Malignant] Pro-B



Categorie: [Malignant] early Pre-B

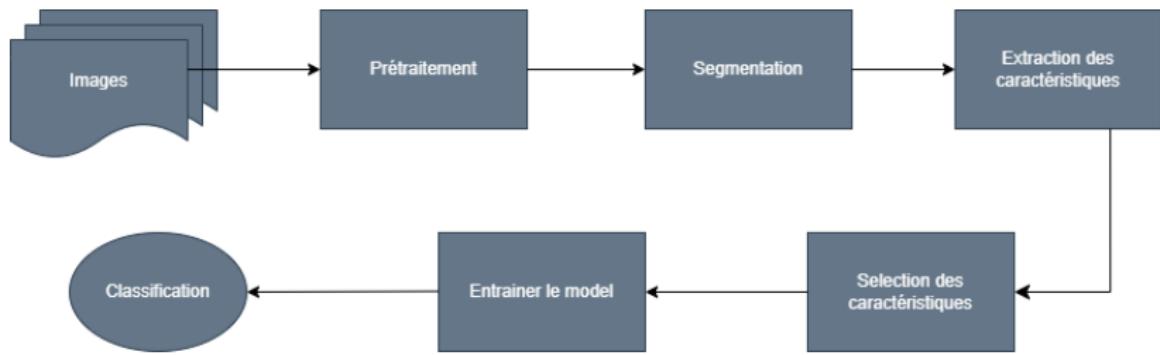


Nombre des image par Categorie



Approche Adoptée

Diagramme de flux



Pré-traitement

Durant cette phase, nous avons :

- Renommé les images en fonction des catégories correspondantes
 - Nom original : [Malignant] early Pre-B / Sap_148(1).jpg
 - Nom modifié : EarlyPreB / [Malignant] early Pre-B2.jpg
- Redimensionné les images RGB originales (768 par 1024 pixels) à 224 par 224 pixels

Segmentation

Étant donné que la leucémie affecte les globules blancs. Dans cette recherche, les algorithmes de segmentation K-means et de seuillage binaire ont été appliqués pour segmenter avec succès la région d'intérêt.

L'algorithme de clustering K-means a permis de segmenter et de séparer la région d'intérêt des autres cellules sanguines. Le choix du K-means est justifié par sa performance supérieure par rapport à la méthode de seuillage.

- K-means : détection de 10 cellules
- Seuillage : détection de 32 cellules

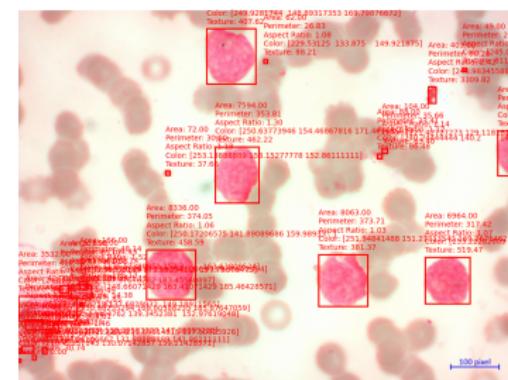
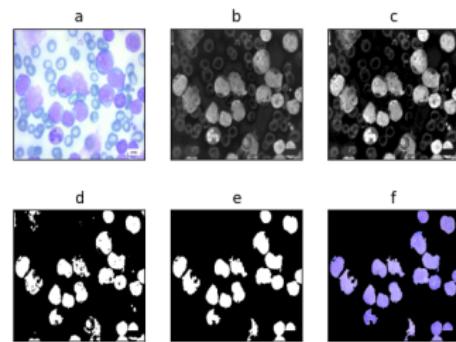


Figura: k-means

Algorithme

Notre algorithme pour la segmentation est le suivant

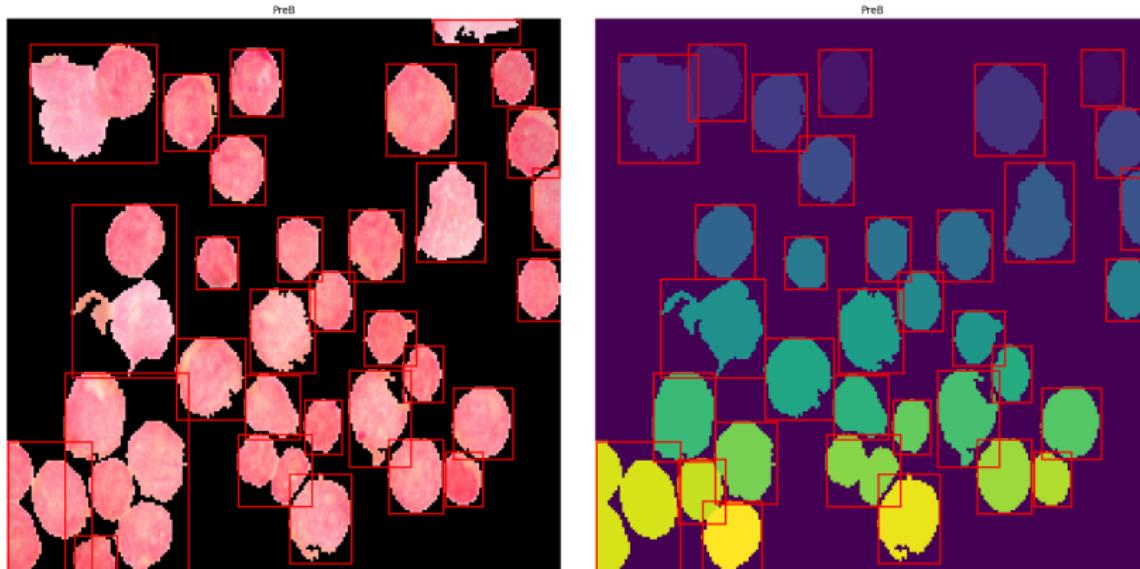
- 1 Convertir l'image RGB en espace couleur LAB.
- 2 Extraire le canal A de l'image LAB.
- 3 Appliquer le regroupement K-means sur le canal A.
- 4 Appliquer un seuillage binaire sur le canal A.
- 5 Remplir les trous dans l'image binaire.
- 6 Supprimer les petits objets et les petits trous de l'image binaire.



Segmentation

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 l'algorithme Watershed: Est une technique de segmentation d'image inspirée de la manière dont l'eau se propage à travers un terrain.

- Efficace pour les scénarios où les objets d'intérêt sont complexes ou touchent les uns aux autres.



Extraction des caractéristiques

caractéristiques

Dans cette étape, différents caractéristiques sont extraites :

Morphologiques

- Solidité
- Excentricité
- Orientation
- Diamètre équivalent
- Surface
- Périmètre
- Ratio d'aspect
- Étendue
- Centroïde
- Boîte englobante
- Surface convexe
- Circularité

Texture

- Contraste
- Corrélation
- Énergie

Couleur

- Couleur (RGB)

Visuelle

- Moyenne de cellules cancéreuses

Intensité

- Intensité moyenne
- Écart type de l'intensité
- Intensité médiane

Extraction des caractéristiques

Transformations des Données

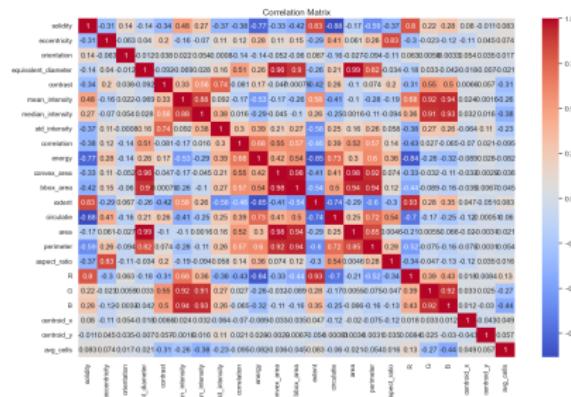
Une fois les caractéristiques des cellules segmentées extraites, nous avons effectué quelques transformations sur les données de type objet :

- Division de la colonne "couleur"en 3 colonnes séparées : R, G, B.
- Division de la colonne "centre"en 2 colonnes : centreX et centreY.

Avec ces transformations, nous avons augmenté le nombre de caractéristiques à 24. De plus, nous avons inclus une étape de normalisation pour mettre toutes les caractéristiques sur la même échelle, permettant ainsi au modèle d'apprendre de manière plus efficace.

Sélection des caractéristiques

Pendant l'examen de la matrice de confusion, nous avons constaté une forte corrélation entre certaines caractéristiques, suggérant une possible redondance dans les données. Cette redondance pourrait influencer la performance du modèle.



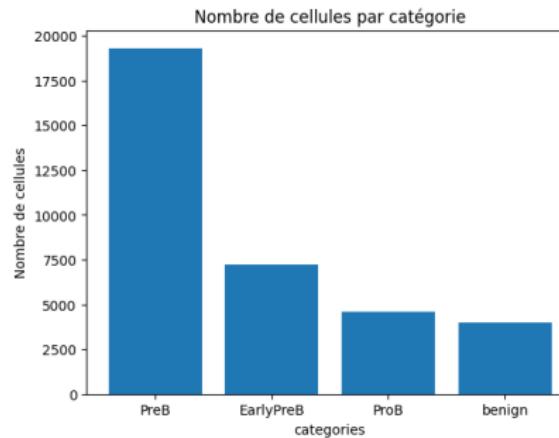
Sélection des caractéristiques

Pour résoudre ce problème, nous avons utilisé la Recursive Feature Elimination (RFE) avec une régression logistique comme modèle initial. Cette approche implique l'utilisation de la régression logistique pour évaluer l'importance des différentes caractéristiques et sélectionner celles qui ont le plus d'impact sur la performance du modèle. Nous avons ensuite restreint cette sélection aux cinq caractéristiques les plus pertinentes pour la tâche de modélisation que nous avons entreprise.

- Diamètre équivalent
- Surface
- Composante R de couleur
- Composante B de couleur
- Moyenne des cellules dans l'image

Sous échantillonage

Le déséquilibre de données entre les classes peut impacter la performance du modèle lors de l'entraînement. Pour remédier à cela, nous avons décidé d'utiliser une méthode de sous-échantillonnage. Cette technique consiste à réduire le nombre d'échantillons de la classe majoritaire afin d'équilibrer les données et d'éviter un biais dans l'apprentissage du modèle. En réduisant la prédominance de la classe majoritaire, nous nous assurons que le modèle est capable de mieux généraliser et de prendre en compte toutes les classes de manière équilibrée.



Entraînement du modèle

Pour la phase d'entraînement, nous avons exploré plusieurs algorithmes afin de sélectionner le meilleur modèle pour notre projet.

- k-nearest neighbor (knn)
- Support Vector Machine (SVM)
- Arbre de Décision
- Forêt Aléatoire

Résultats expérimentales

Résultats expérimentales

Résultats

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

Tabela: Résultats des métriques de précision pour chaque modèle

Modèle	Précision	Rappel	F1-score	Support
KNN	0.96	0.96	0.96	3104
SVM	1	1	1	3104
Arbre de Décision	1	1	1	3104
Forêt Aléatoire	1	1	1	3104

Défis rencontrés

Défis rencontrés

Les principaux défis rencontrés incluent :

- **Le choix des paramètres** : Nous avons dû expérimenter manuellement différentes techniques et paramètres, en nous basant sur notre compréhension limitée du domaine et en évaluant visuellement les résultats obtenus.
- **Le 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

Conclusion et Travaux Futurs

Conclusion et Travaux Futurs

- Objectif : Développer des méthodes automatisées pour l'analyse précise et rapide des échantillons sanguins.
- Approche : Exploration et application de techniques de traitement d'image et de machine learning.
- Résultats : Segmentation efficace des cellules sanguines et classification des différents types de cellules.
- Motivation : Engagement à améliorer les compétences et la compréhension pour optimiser les méthodes.

Perspectives

- Utilisation des réseaux de neurones convolutionnels (CNN) pour améliorer la précision de la segmentation et de la classification des cellules sanguines, dépassant ainsi les méthodes traditionnelles.
- Exploration de données dynamiques pour découvrir de nouveaux aspects et obtenir des résultats plus complets et précis.