الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي و البحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة أبي بكر بلقايد – تلمسان –

Université Aboubakr Belkaïd – Tlemcen –

THESE

Présentée pour l'obtention du **grade** de **DOCTORAT 3ème Cycle**

En : Génie Biomédical

Spécialité : Signaux et Images en Médecine

Par : BECHAR Mohammed El Amine

Sujet

Classification partiellement supervisée des données médicales brutes

Soutenue publiquement, le 09 / 07 /2017, devant le jury composé de :

M Bessaid Abdelhafid	Professeur	Univ. Tlemcen	Président
M Chikh Mohammed Amine	Professeur	Univ. Tlemcen	Directeur de thèse
M Adel Mouloud	Professeur	AIX-MARSEILLE Université	Co- Directeur de thèse
M Abderrahim Mohammed El Amine	MCA	Univ. Tlemcen	Examinateur 1
M Rahmoun Abdellatif	Professeur	ESI SBA	Examinateur 2
M Souier Mehdi	MCA	Ecole Supérieure de Management Tlemcen	Examinateur 3
Mme Settouti Nesma	MCB	Univ. Tlemcen	Invitée 1

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abou Bekr Belkaid
Faculté de Technologie
Département de Génie Biomédical
Laboratoire de Génie Biomédical GBM

THÈSE DE L'UNIVERSITÉ DE TLEMCEN

pour obtenir le grade de

DOCTORAT 3EME CYCLE

Spécialité : **Signaux et Images en Médecine**

présentée et soutenue publiquement
par

BECHAR Mohammed El Amine

le 09 Juillet 2017

Titre:

Classification partiellement supervisée des données médicales brutes

Jury

Président du jury.	Pr. BESSAID Abdelhafid,	UABB Tlemcen
Examinateurs.	Pr. RAHMOUN Abdellatif,	ESI SBA
	. Dr. ABDERRAHIM,	MCA, UABB Tlemcen
	. Mohammed El Amine,	
	. Dr. SOUIER Mehdi,	MCA, ESM Tlemcen
Invitée.	Dr. SETTOUTI Nesma,	MCB, UABB Tlemcen
Directeur de thèse.	Pr. CHIKH Mohamed Amine,	UABB Tlemcen
Co-Directeur de thèse.	Pr. ADEL Mouloud,	AIX-MARSEILLE Université

Je dédie ce travail à :

Mes grands parents,
Mes parents,
Mes frères,

Qu'ils trouvent ici l'expression de toute ma reconnaissance.

Résumé

Les données partiellement supervisées, c'est un effet qui reflète une véritable problématique concernant la difficulté d'étiquetage manuel des données. En classification supervisée des données médicales, l'hypothèse d'apprentissage nécessite une connaissance apriori sur les données où le médecin a apporté l'étiquette nécessaire. Néanmoins, face aux volumes de données disponibles actuellement, la supervision des données médicales est devenue une tâche fastidieuse pour le médecin et parfois même coûteuse dans certaines applications. De ce fait, les données non étiquetées sont plus nombreuses et disponibles par rapport aux données étiquetées. Cependant, sachant que la performance d'un classifieur est liée au nombre de données d'apprentissage, la principale question qui ressort est comment améliorer l'apprentissage d'un classifieur en intégrant des données non étiquetées à l'ensemble d'apprentissage. La technique d'apprentissage issue de la réponse à cette question est appelée l'apprentissage semi-supervisé.

Dans cette thèse, nous détaillons notre problématique majeure à savoir l'étiquetage automatique par apprentissage semi-supervisé en se basant sur le principe « d'auto-apprentissage ». L'auto-apprentissage est un algorithme de référence en classification semi supervisée, son usage est fondamental dans plusieurs applications. Dans l'auto-apprentissage (self-training), nous entrainons un classifieur supervisé avec les données étiquetées. Ensuite ce classifieur est utilisé pour prédire les étiquettes manquantes des données non étiquetées. Les données nouvellement étiquetées avec un haut degré de confiance sont ajoutées à la base étiquetée. Le classifieur est ré-entrainé sur les nouvelles données et cette procédure est répétée jusqu'à satisfaire un critère d'arrêt (convergence). Nous introduisons de manière progressive le concept d'auto-apprentissage dans des applications médicales. Une première partie dans cette thèse a été réservée pour la compréhension du principe d'auto-apprentissage par l'étude de l'algorithme *SNNRCE*. Par la suite, nous détaillons notre contribution proposée au problème d'annotation des données médiales qui est portée sous le nom de *R-COSET*.

Dans la dernière partie de cette thèse, nous nous intéressons plus particulièrement à la segmentation des images médicales utilisant les procédés de classification. La classification super-pixellique est devenue une méthode fréquente et importante dans la segmentation automatique. Une étude expérimentale est proposée dans cette thèse, nous mettons en discussion de manière empirique les considérations requises dans la classification super-pixellique à savoir l'information couleur de l'image et la caractérisation super-pixellique. La classification est effectuée par un apprentissage supervisé et semi-supervisé afin de mettre en évidence l'importance du semi supervisé dans la segmentation des images médicales.

Mots clés

Apprentissage semi-supervisé ; classification ; segmentation ; auto-apprentissage ; SETRED ; SNNRCE ; données médicales.

Abstract

The semi-supervised data is a context that reflects a real problem concerning the difficulty of manual data labeling. In supervised classification of medical data, the learning hypothesis requires the supervised information where the doctor has brought the necessary label. However, in view of the currently available data, the supervision of medical data has become a tedious task for the doctor. Therefore, the unlabeled data is more numerous and available than the labeled data. However, the classifier performance is related to the number of learning data, the main question that emerges is how to improve the learning by including unlabeled data at the learning set. The paradigm using both labeled and unlabeled data is called semi-supervised learning. In recent years, the semi-supervised classification, which used unlabeled data to improve the accuracy of the learned hypothesis, experienced significant growth, this in particular in the Machine Learning community.

In this thesis, we focus on understanding the semi supervised learning, and we detail our main problem about automatic labeling by semi-supervised learning based on the "self-training" principle. Self-training is a reference method in semi supervised classification, its use is fundamental in several applications. The self-training method is an iterative mechanism. First, it trains a supervised classifier on labeled data to predict the labels of unlabeled data. Second, it iteratively enriches the labeled set by adding newly labeled examples with high confident prediction from the unlabeled data (confidence data). We introduce progressively the concept of self-training pcosses in medical applications. A first part of this thesis has been reserved for the understanding of the self-learning principle starting with the study of the *SNNRCE* algorithm. Next, we detail our proposed contribution to the annotation problem of the medial data, which is called *R-COSET*.

In the last part of this thesis, we are particularly interested in the segmentation of medical images using the classification processes. The super-pixel classification has become a frequent and important method for the automatic segmentation. An experimental study is proposed in this thesis. We discuss empirically the considerations required in the super-pixel classification, namely the space color information and the super-pixel characterization. The classification is performed by supervised and semi-supervised learning in order to highlight the importance of semi-supervised in the medical image segmentation.

Keywords

Semi-supervised learning ; classification ; segmentation ; self-training ; SETRED ; SNNRCE ; medical dataset.

Remerciements

Une thèse est un effort collectif et elle est loin d'être un travail solitaire. En effet, je n'aurais jamais pu réaliser ce travail doctoral sans le soutien d'un grand nombre de personnes dont la générosité, la bonne humeur et l'intérêt manifestés à l'égard de ma recherche m'ont permis de progresser dans cette phase délicate. Je tiens donc ici à adresser mes remerciements à toutes les personnes qui ont participé de près ou de loin à cet ouvrage.

Je tiens tout d'abord à adresser mes remerciements les plus sincères au Professeur Mohammed Amine CHIKH pour avoir dirigé cette thèse et m'avoir permis de la réaliser dans les meilleures conditions. Je le remercie pour son soutien depuis mon master jusqu'à la fin de ma thèse à tous les niveaux. Son assistance et ses conseils m'ont chaque fois permis de rebondir dans les moments difficiles. Je le remercie vivement pour l'aide scientifique précieuse et pour tous les conseils donnés pendant la durée de cette thèse.

Mes remerciements s'adressent ensuite au Professeur Mouloud ADEL qui a co-encadré cette thèse. Ses conseils avisés tout au long de mon séjour de finalisation de thèse au sein d'institut FRESNEL, AIX-Marseille Université, m'ont permis d'enrichir ce travail. J'aimerais également lui dire à quel point j'ai apprécié sa grande disponibilité et son respect sans faille des délais serrés de relecture des documents que je lui ai adressés.

Je tiens à remercier vivement monsieur Abdelhafid BESSAID, Professeur et chef d'équipe Imagerie médicale dans le laboratoire de génie biomédical à l'Université de Tlemcen, pour l'honneur qu'il m'a fait en acceptant d'être président de mon jury de thèse. Je tiens à l'assurer de ma profonde reconnaissance pour l'intérêt qu'il porte à ce travail.

J'associe à ces remerciements monsieur Abdellatif RAHMOUN, Professeur à l'école supérieure en informatique de Sidi Bel Abbès, monsieur Mohammed El Amine ABDERRAHIM, Maitre de conférences à l'Université de Tlemcen, mohammed Mehdi SOUIER, Maitre de conférences à l'école Supérieure de Management Tlemcen et madame Nesma SETTOUTI, Maitre de conférences à l'Université de Tlemcen pour avoir accepté d'examiner mon travail. J'imagine le travail que représente la lecture d'une thèse et les en remercie sincèrement.

Les différents membres du laboratoire Génie Biomédical GBM et tout particulièrement de l'équipe CREDOM qui ont également contribué à la réalisation de ces travaux, que ce soit à travers de longues discussions scientifiques ou bien grâce à l'ambiance chaleureuse et au très bon état d'esprit qu'ils entretiennent. J'ai une pensée particulière pour Nesma SETTOUTI, Mostafa El HABIB DAHO, Khalida DOUIBI, et Sara BOUCHIKHI avec qui j'ai contribué à la concrétisation de plusieurs travaux de recherche.

Table des matières

Table des figures

Liste des tableaux

Glossaire

ACP : Analyse à Composante Principale
APVs : Adjusted p-values
Bagging : Bootstrap Aggrigating
CART : Classification and Regression Trees
CIE : Commission Internationale de l'Eclairage
CIF : Conditional Inference Forest
COSTRA : Condence-based Self-training
D_A : Donnée d'apprentissage
DE : Distance Euclidienne
D_E : Données étiquetées
D_{NE} : Données non-étiquetées
D_T : Donnée de test
EM : Expectation-maximisation algorithm
EoC : Ensemble of Classifiers
FAR : Friedman Aligned-Ranks
FCM : Fuzzy C-means
fMRI : Functional Magnetic Resonance Imaging
FS : Feature Selection
GDI : Indice de Gini
KNN : k nearest neighbors
MCS : Multiple Classifier Systems
PPV : Plus Proche Voisin
RCOSET : Reinforced Confidence in Self-training
RF : Random Forest
RGB : Red Green Blue
SETRED : self-training with data editing
SNNRCE : Semi-supervised learning based on nearest neighbor rule and cut edges
SSL : Semi-Supervised Learning
SubBag : Subspaces Bagging
SVM : Support Vector Machine
UCI : University California Irvine
WBC : White Blood Cell

Première partie

I

Introduction générale

Introduction générale

1 Motivations

Données partiellement labélisées, c'est un contexte qui décrit une véritable réalité existante sur scène. Actuellement, l'innovation technologique a donné un usage exponentiel dans différentes disciplines, apportant d'une part des facilités sur des exercices quotidiens (exemple : santé, médical, communication, transport, . . .), et d'autre part une explosion en terme d'acquisition et stockage de données. Exemple, dans un service de radiologie médicale et avec la technologie développée en imagerie médicale, un scanner peut produire quotidiennement un nombre important de séquences d'images sur de nombreux individus pour un faible coût. Toutefois, l'expertise humaine de cette masse d'image devient de plus en plus compliquée, difficile et coûteuse.

Dans la pratique, l'annotation manuelle d'une masse de données est devenue difficile à cause :

- **Du coût :** L'action de l'étiquetage manuel des données exige la présence d'un expert humain. Dans le cas général, les experts ont besoin d'être rémunérés.

- **De la difficulté :** Parfois, les objets doivent être subdivisés en parties cohérentes avant de pouvoir être étiquetés. Par exemple, les signaux et les images doivent être précisément segmentés en objets, respectivement avant que l'étiquetage puisse être effectué.

- **De l'ambiguïté :** Les objets peuvent avoir des étiquettes différentes ou l'étiquetage lui-même peut ne pas être fiable en raison d'un désaccord entre les experts.

- **Du vocabulaire d'étiquetage :** Le réglage d'étiquetage typique consiste à sélectionner une étiquette à partir d'une liste pré-spécifiée, cette liste peut ne pas décrire complètement ou précisément un objet. A titre d'exemple, pour une collection d'images bien spécifiée, habituellement elles sont accompagnées d'un vocabulaire pré-spécifié qui ne peut décrire que les images déjà présentées dans les données d'apprentissage.

- **De l'incertitude au niveau de détail :** Une possibilité que les étiquettes des objets changent avec la granularité à laquelle l'expert examine l'objet. Exemple, une image d'une personne peut être étiquetée comme «personne», ou à un plus grand détail «visage», «yeux», etc.

Suite à ces conséquences, les données sont partiellement labélisées, et en parallèle une disposition faramineuse des données non expertisées qui présente une richesse d'information non exploitée.

La donnée médicale ne cesse d'augmenter avec le développement technologique, et l'effet de la difficulté d'obtenir l'information expertisée a exigé un besoin d'un développement adapté à cette problématique. La recherche scientifique s'appuie notamment sur l'informatique médicale afin de trouver des solutions efficaces, permettant de manipuler les données expertisées et non expertisées dans un système d'aide au diagnostic fiable.

Pour faire face à cette problématique, les techniques d'aide au diagnostic nécessitent des transformations et optimisations stratégiques majeures pour assurer l'exploitation, l'exploration et l'interprétabilité de ces masses de données. Elles nécessitent également une prise de conscience et une modification des pratiques des experts du domaine, pour lesquels ces évolutions constituent un défi pour minimiser les erreurs médicales.

2 Contexte

La qualité des soins, c'est l'objectif qui a pris une priorité majeure dans le domaine de la recherche en santé et en ingénierie biomédicale. Des innovations brillantes ont vu le jour notamment dans l'acquisition et le stockage des données médicales. Par cette innovation, l'acquisition des données est devenue de plus en plus facile, précise et efficace.

La donnée médicale telle que les images, les signaux, les analyses sanguines, les analyses biologiques, . . . sont devenues des éléments indispensables dans la pratique des soins médicaux. De ce fait, un taux inestimable des données est acquis annuellement dans chaque établissement hospitalier. Les médecins se plaignent de la difficulté de prendre en charge toute cette masse de données. D'une part, les médecins ont besoin à d'outils qui réalisent automatiquement cette annotation. D'autre part, ça explique le fait que les données non étiquetées représentent la majorité des cas, surtout en comparaison avec la quantité de données étiquetées. L'utilisation de ces données nécessite un soin particulier, car plusieurs problèmes se posent à l'augmentation de la dimensionnalité et à l'absence d'étiquette.

Les outils issus de la discipline fouille de données, fournissent des méthodologies permettant d'automatiser les tâches requises par les experts pour l'annotation de ces données [2]. L'apprentissage automatique est le domaine qui est largement concerné dans cette résolution. L'apprentissage automatique utilise des outils mathématiques dans la conception d'une fonction $f : x \rightarrow y$, avec x qui décrit la description d'une donnée désignée et y leur étiquette prédite par la fonction f. Traditionnellement, il existe deux modes d'apprentissage automatique. Les méthodes supervisées qui s'adressent généralement à la classification des données sur la base de connaissances préalables acquises par un apprentissage basé sur des données déjà annotées par l'expert, alors que les données dans les méthodes non supervisées (clustering) sont basées uniquement sur la similitude des instances de données sans apprentissage. Cette dernière pourrait être considérée comme avantageuse sur les méthodes supervisées. Cependant, les méthodes non supervisées, en général, nécessitent que le nombre de groupes cibles soit pré-spécifié par l'utilisateur, et les résultats ne soient pas associés à des étiquettes de classe.

Suite à la problématique de la dimensionnalité des données non étiquetées et la difficulté d'obtenir leurs étiquettes, des nouveaux travaux se focalisent sur la résolution de cette problématique. Ces travaux ont apporté une grande valeur au développement de l'apprentissage semi-supervisé. Contrairement à l'apprentissage supervisé, l'apprentissage semi-supervisé vise les problèmes avec relativement peu de données étiquetées et une grande quantité de données non étiquetées.

L'apprentissage semi-supervisé, comme son nom l'indique, est l'apprentissage qui se situe entre l'apprentissage supervisé et non supervisé, est alors une solution envisageable [3]. La qualité de semi-supervision réside dans la production d'une règle de prédiction à partir d'un ensemble de données, avec une particularité : cet ensemble contient à la fois des objets étiquetés, et d'autres qui ne le sont pas (non étiquetés).

Le sujet de cette thèse concerne la manipulation des données étiquetées et comment bénéficier de la richesse des données non étiquetées dans la classification semi-supervisée. Cela dans un objectif d'étudier l'efficacité de cet apprentissage dans la classification des données médicales, et relevant le défi d'améliorer ce mode d'apprentissage augmentant la précision de la classification. Essentiellement, l'application de l'apprentissage semi-supervisé dans la segmentation automatique des images médicales, cette application a pour traiter la complexité et la difficulté du marquage manuel des régions d'intérêt des images médicales par les experts médicaux.

3 Contributions

Les références fondamentales de cette thèse se résument dans [4], [5], [6], [7], [8], [9], [10], [11] et [12]. Les références citées montrent clairement l'intérêt d'appliquer l'apprentissage semi-supervisé pour résoudre les problèmes de classification automatique avec peu de données expertisées. Les principales contributions dans cette thèse sont menées pour étudier attentivement l'apprentissage semi-supervisé, et l'appliquer sur des données médicales dans le but d'aider les médecins dans la routine clinique. Nos contributions dans cette thèse sont résumées dans les points suivants :

1. Proposition d'une étude détaillée sur la définition de similarité en apprentissage semi supervisé, dans le but d'étudier et comprendre l'influence de cette mesure sur les performances de la classification.
2. Proposition d'une nouvelle approche de classification des données médicales avec la présence de peu de données expertisées en utilisant un apprentissage semi-supervisé, dans l'objectif d'apporter des outils d'aide au diagnostic.
3. Proposer l'application des techniques semi supervisées dans une méthode d'annotation des images médicales au niveau super-pixellique pour la segmentation des régions d'intérêt de l'image.

4 Organisation du manuscrit

Le reste de ce manuscrit est structuré en deux parties principales réparties comme suit :

Partie 1 : Classification des données médicales par apprentissage semi-supervisé : auto-apprentissage.

– Dans le chapitre II.1, Une introduction à la classification semi supervisée est présentée dans ce chapitre, nous exposons une synthèse non exhaustive des travaux de la littérature qui décrivent cette thématique de recherche, et nous tentons d'y montrer les avantages et les inconvénients de ces méthodes afin de situer notre contribution proposée par rapport à celles-ci.

– Dans le chapitre II.2, une première contribution concernant la définition de la similarité en semi supervisé est présentée dans ce chapitre, sous forme d'une étude comparative entre différentes mesures de similarité appliquées dans un algorithme de classification semi supervisée [46], et discutant l'influence de cette mesure sur la performance de classification.

– Dans le chapitre II.3, nous présentons une nouvelle méthode qui a été développée dans le cadre de cette thèse (R-COSET [47]), cette proposition réside dans un nouvel algorithme qui permet d'améliorer les performances de l'apprentissage supervisé et semi-supervisé, et ainsi d'augmenter la précision de la classification.

Partie 2 : Segmentation supervisée et semi-supervisée des cellules sanguines par classification super-pixellique.

– Dans le chapitre III.1, la classification super-pixellique pour la segmentation nécessite une étape primordiale, ça concerne la caractérisation ou l'extraction des caractéristiques à partir de chaque super-pixel de l'image. En effet, cette étape a été étudiée attentivement dans ce chapitre dans l'objectif de produire une bonne caractérisation super-pixellique permettant par la suite une meilleure classification et segmentation.

– Dans le chapitre III.2, nous discutons les performances de la classification super-pixellique en utilisant l'approche semi-supervisé comme technique d'apprentissage. Une comparaison est menée par rapport à l'apprentissage supervisé afin de montrer l'utilité et l'efficacité de la classification semi-supervisée.

Conclusions et Perspectives

– Dans le chapitre IV.1, nous présentons une conclusion générale qui résume une synthèse des contributions apportées ainsi que les chemins définissant les perspectives possibles pour de futurs travaux.

Deuxième partie

II

Classification des données médicales par apprentissage semi-supervisé : auto-apprentissage

CHAPITRE 1

Introduction à la classification semi supervisée

L'apprentissage semi supervisé est un régime d'apprentissage qui s'inspire de la façon dont les systèmes naturels comme les êtres humains apprennent en présence de données étiquetées et non étiquetées. Généralement, l'apprentissage automatique a été étudié soit dans un contexte non supervisé (le regroupement) où toutes les données d'apprentissage sont non étiquetées, ou dans un contexte supervisé (par exemple, classification, régression) où toutes les données d'apprentissage sont étiquetées. L'apprentissage semi-supervisé présente un grand intérêt dans l'exploration de données, car il peut utiliser des données non étiquetées qui sont facilement disponibles sur scène pour améliorer les tâches d'apprentissage supervisé, particulièrement lorsque les données étiquetées sont rares ou coûteuses à obtenir. Dans ce chapitre, nous présentons quelques modèles connus d'apprentissage semi supervisé, y compris l'auto-apprentissage, les modèles génératifs, le co-training/multiview, les méthodes à base des graphes et le SVM en version semi supervisé (S3VM).

1 Apprentissage à partir de données étiquetées et non étiquetées

Comme son nom l'indique, l'apprentissage semi-supervisé se situe entre l'apprentissage supervisé et non supervisé. En effet, la plupart des stratégies d'apprentissage semi-supervisées sont basées sur l'extension de l'apprentissage supervisé ou non supervisé pour inclure des informations supplémentaires typiques de l'autre régime d'apprentissage. Plus précisément, l'apprentissage semi-supervisé englobe plusieurs contextes différents, notamment :

Classification semi-supervisée. Également connu sous le nom de classification avec des données étiquetées et non étiquetées (ou des données partiellement étiquetées), il s'agit d'une extension du problème de classification supervisée. Les données d'apprentissage sont constituées de l'information étiquetée et non étiquetée, avec e instances étiquetées $D_E = \{(x_i, y_i)\}_{i=1}^{e}$, et ne instances non étiquetées $D_{NE} = \{x_j\}_{j=e+1}^{e+ne}$. Cette hypothèse suppose que les données non étiquetées sont beaucoup plus nombreuses que celles étiquetées, c'est-à-dire $ne \gg e$. L'objectif de la classification semi-supervisée est de faire apprendre une hypothèse h à partir des données étiquetées et non étiquetées, de sorte qu'il soit préférable que l'hypothèse utilisée soit supervisée et en intégrant les données non étiquetées dans le processus d'apprentissage.

Regroupement semi-supervisé. Il s'agit d'une extension au clustering non supervisé. L'ensemble d'apprentissage se compose de données non étiquetées $D_{NE} = \{x_i\}_{j=1}^{ne}$, ainsi fournissant une certaine information supervisée pour chaque cluster à partir du peu de données étiquetées qui existe. L'objectif ici est d'adapter les algorithmes de regroupement afin de pouvoir prendre en charge les contraintes d'étiquetage, permettant par la suite de regrouper les données non étiquetées en clusters.

L'étude de l'apprentissage semi-supervisé est motivée par deux facteurs : sa valeur pratique dans la construction de meilleurs algorithmes informatiques, et sa valeur théorique dans la compréhension de l'apprentissage dans les machines et les êtres humains. L'apprentissage semi-supervisé a une valeur pratique considérable dans la proposition des solutions efficace pour l'annotation automatique en présence de peu de données supervisées. Dans de nombreuses tâches, les étiquettes y sont devenues difficiles à obtenir car elles nécessitent des experts humains, des dispositifs spéciaux ou des expériences coûteuses et lentes. Par exemple :

– Dans la reconnaissance de la parole, une instance x est un énoncé de la parole, et l'étiquette y est la transcription correspondante. Voici un exemple d'une transcription phonétique détaillée des mots tels qu'ils sont énoncés :

  film⇒ f ih_n uh_gl_n m
  be all⇒ bcl b iy iy_tr ao_tr ao l_dl

  Une transcription précise par des annotateurs experts est une action lente et couteuse, une expérience déjà envisagée a nécessité jusqu'à 400 heures pour transcrire une heure de discours issu des conversations téléphoniques [13] (Enregistrements de participants appariés au hasard qui discutent de divers sujets tels que les questions sociales, économiques, politiques et environnementales).

– Dans la vidéo de surveillance, une instance x est une séquence de vidéo et l'étiquette y est l'identité de l'objet dans la vidéo. L'étiquetage manuel des objets dans un grand nombre de séquences vidéo est une opération fastidieuse.

– Dans la prédiction de la structure 3D de protéine, une instance x est une séquence d'ADN, et l'étiquette y est la structure de pliage de protéine 3D. Des mois de travail sont nécessaires dans des laboratoires avec des experts du domaine pour identifier la structure 3D d'une seule protéine.

De ce fait, les données étiquetées $D_E = \{(x_i, y_i)\}_{i=1}^{e}$ sont souvent difficiles à obtenir dans plusieurs domaines, les données non étiquetées $D_{NE} = \{x_j\}_{j=e+1}^{e+ne}$ sont disponibles en grande quantité et faciles à collecter. Exemple :

– Les énoncés de la parole peuvent être enregistrés à partir d'émissions de radio,
– Les phrases de texte peuvent être explorées à partir du Web,
– Accès aux courriels électroniques à partir des serveurs de messagerie,
– Caméras de surveillance fonctionnent 24/24 heures par jour,
– Des séquences d'ADN de protéines sont facilement disponibles à partir de bases de données de gènes,
– Les services radiologiques acquis des milliers de séquences d'imageries médicales sur plusieurs patients.

Cependant, les méthodes traditionnelles d'apprentissage supervisé ne peuvent pas intégrer les données non étiquetées dans l'apprentissage du classifieur.

L'apprentissage semi-supervisé est la solution envisagée pour résoudre la problématique de la difficulté d'annotation manuelle, car il peut utiliser à la fois des données étiquetées et non étiquetées pour obtenir de meilleures performances que l'apprentissage supervisé.

D'un point de vue différent, l'apprentissage semi-supervisé peut atteindre le même niveau de performance que l'apprentissage supervisé, mais avec moins d'exemples étiquetés.

La figure 1 illustre un exemple simple d'apprentissage semi-supervisé. Soit chaque instance représentée par une entité uni-dimensionnelle $x \in \mathbb{R}$. Il existe deux classes : positive $(+)$ et négative $(-)$. Considérant les deux scénarios suivants :

1. Dans l'apprentissage supervisé, on considère seulement deux cas étiqueté pour l'apprentissage $(x_1, y_1) = (-1, -)$ et $(x_2, y_2) = (+1, +)$, cela est représenté dans la figure 1. La meilleure estimation de la limite de décision est évidemment $x = 0$: toutes les occurrences avec $x < 0$ doivent être classées comme $y = -$, alors que celles avec $x \geqslant 0$ comme $y = +$.

2. En outre, nous avons également intégré un grand nombre d'exemples non étiquetés, représentés comme des points dans la figure 1. Cependant, nous observons qu'ils forment deux groupes cohérents. Cela veut dire que les données non étiquetées peuvent enrichir la base de connaissances. Les techniques semi-supervisées sont capables de générer une décision plus fiable que le supervisé. La semi-supervision dans cet exemple a estimé la limite de décision à $x \approx 0.5$.

Si notre hypothèse est vraie, l'utilisation de données à la fois étiquetées et non étiquetées nous donne une estimation plus fiable de la limite de décision. Intuitivement, la distribution de données non étiquetées aide à identifier les régions de même label, et les données étiquetées fournissent alors les étiquettes réelles. Dans ce chapitre, nous présenterons quelques autres hypothèses d'apprentissages semi-supervisées couramment utilisées.

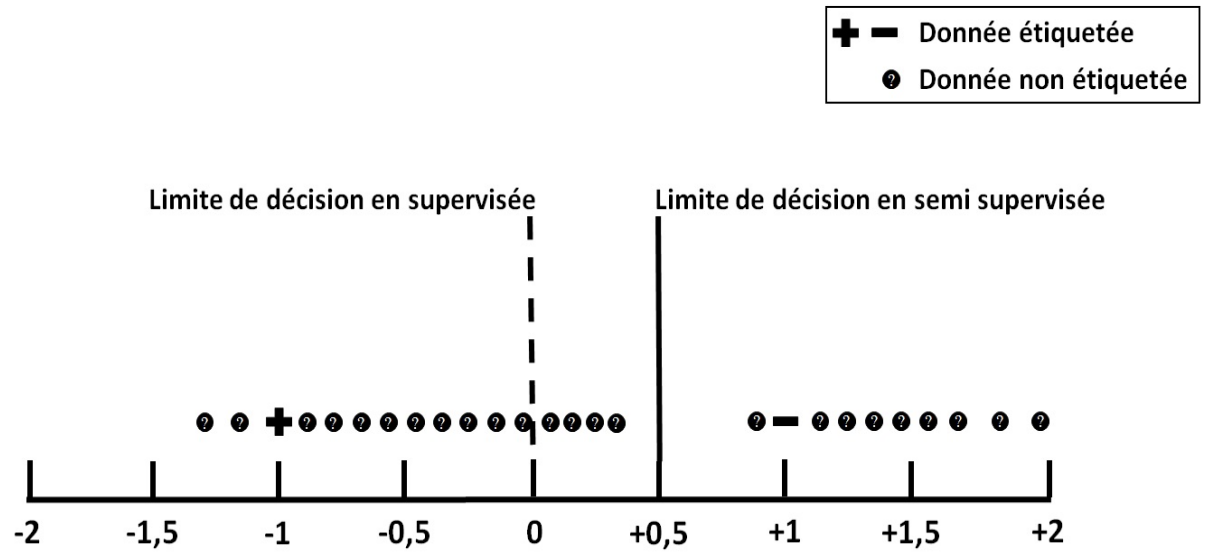


Figure 1 – Un exemple simple pour démontrer comment l'apprentissage semi-supervisé est possible.

2 Apprentissage semi supervisé transductif / inductif

Il existe en effet deux réglages d'apprentissage semi-supervisés légèrement différents, à savoir l'apprentissage semi-supervisé transductif et inductif. Dans la classification supervisée, les échantillons d'apprentissage sont entièrement étiquetés, de sorte que l'on est toujours intéressé par la performance de classification sur les futures données de test. Dans

la classification semi supervisée, cependant, les données d'apprentissage contiennent aussi des données non étiquetées. On distingue deux objectifs de cet effet, le premier consiste à prédire les futurs données de test qui est la phase inductive. Le deuxième objectif est de prédire les étiquettes des instances non étiquetées dans l'échantillon d'apprentissage qui est la phase transductive.

Définition 1. Apprentissage tranductif semi supervisé. Compte tenu d'un échantillon d'apprentissage $D_E = \{(x_i, y_i)\}_{i=1}^{e}$, $D_{NE} = \{x_i\}_{j=e+1}^{e+ne}$, l'apprentissage transductif forme une fonction $f : X^{e+ne} \mapsto Y^{e+ne}$ de sorte que f devrait être un bon prédicteur sur les données non étiquetées $D_{NE} = \{x_i\}_{j=e+1}^{e+ne}$.

Définition 2. Apprentissage inductif semi supervisé. Compte tenu d'un échantillon d'apprentissage $D_E = \{(x_i, y_i)\}_{i=1}^{e}$, $D_{NE} = \{x_j\}_{j=e+1}^{e+ne}$, l'apprentissage inductif forme une fonction $f : X \mapsto Y$ de sorte que f devrait être un bon prédicteur sur les données de test $D_T = \{x_i\}_{i=1}^{t}$, qui n'ont pas été utilisées lors d'apprentissage.

Plusieurs sortes de familles ont été développées pour réaliser l'apprentissage semi-supervisé. Ce chapitre résume les méthodes populaires de ce régime d'apprentissage. Ceci a pour but de mettre en évidence la variété des hypothèses du modèle, ainsi que de préparer le terrain pour notre nouveau travail décrit dans les chapitres suivants. Nous commençons par le modèle d'auto-apprentissage (self-training).

3 Auto-apprentissage

L'auto-apprentissage est caractérisé par le fait que le processus semi-supervisé utilise les propres prédictions d'une seule fonction de prédiction f pour enrichir la base étiquetée. L'auto-apprentissage peut être inductif ou transductif, selon la nature du prédicteur f.

Algorithm 1 Algorithme d'auto-apprentissage

1: **Entrée**
   D_E : Données étiquetées
   D_{NE} : Données non-étiquetées
2: **Répéter :**
3: Faire apprendre une hypothèse supervisée h à partir de D_E
4: Appliquer h sur D_{NE} pour la prédiction des étiquettes
5: Supprimer le sous-ensemble S de D_{NE}
6: Ajouter $\{(x, h(x)) | \in S\}$ à D_E

L'idée principale de l'auto-apprentissage est de former tout d'abord une hypothèse de prédiction supervisée h sur les données étiquetées D_E. L'hypothèse h est ensuite utilisée pour prédire les étiquettes des données non étiquetées D_{NE}. Un sous-ensemble S à partir des données non étiquetées, ainsi que leurs étiquettes prédites, sont ensuite identifiées, sélectionnées et ajoutées à la base étiquetée, afin d'enrichir cette dernière. Typiquement, S se compose des quelques exemples non étiquetés avec les prédictions de h les plus confiants. Une fois la base étiquetée est enrichie, l'hypothèse h subit un réapprentissage sur ces nouvelles données, et cette procédure se répète jusqu'à la convergence de l'hypothèse. L'une des premières applications fructueuses de l'auto-apprentissage a été sur la désambiguïsation de mots [14], aussi il a eu des succès dans l'analyse du langage naturel [15].

Les principaux avantages de l'auto-apprentissage résident dans sa simplicité d'implémentation algorithmique, et le fait qu'il s'agit d'une méthode de type « wrapper ». En effet, le choix de l'hypothèse h d'apprentissage est considéré comme un choix personnel de l'utilisateur (étape 3 de l'algorithme 1). Par exemple, l'apprentissage peut être effectué par

un simple algorithme KNN, comme aussi il peut être un classificateur compliqué. La propriété « wrapper » de l'auto-apprentissage a pour objectif de générer un sous-ensemble bien adapté à un certain critère de sélection, cette sélection permettra d'identifier les données de confiance qui sont bien classées.

Néanmoins, il est concevable qu'une erreur précoce commise par h, suivi par une mauvaise sélection de sous-ensemble S avec des étiquettes erronées, peut conduire à une dégradation d'apprentissage et diverger l'auto-apprentissage à des résultats incorrects. Diverses heuristiques ont été proposées pour atténuer ce problème, présentées dans [5–7].

4 Co-Training

Le premier à avoir vu le jour dans la catégorie des techniques d'apprentissage ensemblistes en semi-supervisé est l'algorithme *co-Training*, proposé par Blum et Mitchell [16] pour la classification semi-supervisée des pages web.

Le *co-Training*, suppose que les variables sont naturellement partitionnées en deux ensembles $x = (x_1; x_2)$. Par exemple, pour les pages Web on considère l'ensemble des liens hypertextes et l'ensemble du contenu, sous les hypothèses suivantes :

1. Chaque composant est suffisant pour la classification,
2. Les composants sont indépendants conditionnellement à la classe,

Blum et Mitchell (1998) démontrent des garanties de type Probably Approximately Correct (PAC) (Valiant [17]) sur l'apprentissage en présence de données étiquetées et non étiquetées (Algorithme 2).

Algorithm 2 *co-Training* pseudo code pour la classification de documents

1: **Entrée :** une collection initiale de documents étiquetés et non étiquetés
2: **Sortie :** Deux classifieurs, $C1$ et $C2$, qui prédisent l'étiquette des nouveaux documents.
3: **Répéter jusqu'à ce qu'il n'y est plus de document sans étiquette.**
- Construire le classifieur $C1$ en utilisant la partie x_1 de chaque document
- Construire le classifieur $C2$ en employant la partie x_2 de chaque document.
- Pour chaque classe k, ajouter à la collection de documents étiquetés le document non-étiqueté classé dans la classe k par le classifeur $C1$ avec la plus forte probabilité.
- Pour chaque classe k, ajouter à la collection de documents étiquetés le document non-étiqueté classé dans la classe k par le classifeur $C2$ avec la plus forte probabilité.
4: Ces prédictions peuvent ensuite être combinées.

Les auteurs ont également montré que l'indépendance des deux sous-ensembles d'attributs est une condition nécessaire pour que cet algorithme du *co-Training* améliore la prédiction. En pratique, cependant, il n'est pas toujours possible d'obtenir deux sous-ensembles d'attributs indépendants par rapport à l'étiquette, ce qui rend le *co-Training* difficilement généralisable.

Pour contourner cette difficulté, Goldman and Zhou [18] proposent d'adapter cette stratégie à un ensemble de classifieurs hétérogènes, appelé *statistical co-learning*. Leur méthode suit la procédure mise en place par Blum and Mitchell [16], même s'ils sont obligés de contrôler la qualité des exemples nouvellement étiquetés avant de les ajouter définitivement à l'apprentissage.

Grâce à l'aide de trois classifieurs au lieu de deux, Zhou et al. [19] ont proposé l'algorithme d'apprentissage *Tri-Training*, qui nécessite des sous-ensembles d'attributs ni suffisants ni redondants et pas d'algorithmes d'apprentissage supervisé spéciaux qui pourraient diviser l'espace d'exemple dans un ensemble de classes d'équivalence. L'algorithme *Tri-Training* réalise la prédiction par un vote majoritaire plutôt que par un classifieur combiné ou un stacking.

D'autre part, une amélioration de l'algorithme *co-Training* a vu le jour sous le nom de *co-Forest*, qui étend le paradigme du *co-Training* en appliquant *Random Forest* [9]. Il a été introduit par Li et Zhou [20] dans l'application à la détection de micro calcifications pour le diagnostic du cancer du sein. *co-Forest* utilise $N \geq 3$ classifieurs au lieu des 3 dans *Tri-training*. Les $N-1$ classifieurs sont employés pour la détermination des exemples de confiance, appelés Ensemble de concomitance $= H_i = H_{N-1}$. La confiance d'un exemple non étiqueté peut être simplement estimée par le degré de confiance sur l'étiquetage, à savoir le nombre de classifieurs qui sont d'accords sur l'étiquette assignée par H_i.

5 Méthode générative

Dans les approches génératives, on suppose que les données étiquetées et non étiquetées proviennent d'un même modèle paramétrique où les probabilités $p(y)$ et $p(x|y)$ sont connues et correctes. Une fois les paramètres du modèle prêts et l'apprentissage effectué, les données non étiquetées sont classées en utilisant les composants du modèle associés à chaque classe. Les méthodes de cette catégorie, comme dans [21, 22], traitent généralement les étiquettes de classe des données non étiquetées D_{NE} comme des étiquettes manquantes et utilisent l'algorithme EM (Expectation-Maximization) [23] pour effectuer l'estimation de maximum de vraisemblance du modèle. EM commence par un modèle initial formé sur les exemples étiquetés D_E. Il utilise alors itérativement le modèle pour estimer les probabilités de classe de tous les exemples non étiquetés et maximise ensuite la probabilité des paramètres sur tous les exemples étiquetés jusqu'à ce qu'il converge.

Différents concepts ont été utilisés pour l'ajustement du modèle, par exemple, la distribution gaussienne est exploitée pour la classification semi supervisée des images [24], la distribution Naive Bayes [21] est utilisée pour la catégorisation du texte et Hidden Markov Models (HMM) [25] est utilisé pour la reconnaissance de la parole. Les méthodes de cette catégorie souffrent d'un problème sérieux. Autrement dit, lorsque l'hypothèse du modèle est incorrecte, l'ajustement du modèle à l'aide d'une grande quantité de données non étiquetées entraînera une dégradation des performances [26]. Et pour réduire le risque de cette dégradation [27], il faut construire attentivement le modèle génératif.

6 Méthode à base des graphes

L'intégration de la théorie des graphes dans le semi supervisé a été initialisée par Blum et Chawla [28]. Ils ont construit un graphe dont les sommets représentent les données d'apprentissage étiquetées et non étiquetées et les arêtes entre les sommets sont pondérées en fonction de la similarité entre les exemples correspondants. Avec la théorie des graphes, on cherche à trouver la coupe minimale du graphe de telle sorte que les sommets de chaque composante connectée ont la même étiquette. Dans [28, 29], une fonction prédictive discrète est utilisée qui assigne une étiquette possible à chaque donnée non étiquetée. Zhu et al. [30] ont introduit une fonction de prédiction continue pour la classification semi-supervisée. Ils ont modélisé la répartition de la fonction de prédiction sur le graphe avec des champs aléatoires gaussiens, et ils ont démontré que la fonction de prédiction ayant la plus faible énergie devrait avoir la propriété harmonique. Ils ont conçu par la suite une stratégie de propagation de l'étiquette sur le graphe, en utilisant une telle propriété harmonique où les étiquettes se propagent des sommets étiquetés aux non

étiquetés. L'apprentissage semi-supervisé à base des graphes suppose que les sommets possédant une forte connexion tendent à avoir le même label de classe et vice versa [30].

7 SVM semi supervisé (S3VM)

Les SVMs (support vector machine) sont des algorithmes de classification supervisée, donc sont initialement fondés sur un apprentissage totalement supervisé. Dans le contexte semi-supervisé, plusieurs initiations de SVM en version semi-supervisé ont été proposées ces dernières années [31, 32], les versions proposées ont été développées pour des problèmes de classification binaire. Bennett et Demiriz [31] proposent l'idée de construire le meilleur séparateur pour les données étiquetées tout en minimisant le nombre des données non étiquetées dans la marge (entre les deux hyperplans).

Le but de l'algorithme d'apprentissage S3VM est d'exploiter les données non étiquetées $D_{NE} = \{x_i\}_{j=e+1}^{e+ne}$ pour ajuster la limite de décision, initialement construite à partir d'une petite quantité de données étiquetées $D_E = \{(x_i, y_i)\}_{i=1}^{e}, y_i = \pm 1$. Un problème est suggéré concernant l'optimisation du séparateur, cela est résolu en fonction des paramètres de frontière de décision (w, b) et par des étiquettes prédites des données non étiquetées $\hat{y}_{D_{NE}} = (\hat{y}_1, ..., \hat{y}_{ne})^T \in \{-1, +1\}$.

Depuis sa première implémentation par Joachims [33], le problème associé à S3VM motive le développement d'un certain nombre de techniques d'optimisation, par exemple, la recherche combinatoire locale [34, 35], la descente en gradient [36], les techniques de poursuite [37], programmation semi-définie [38] et optimisation de l'algorithme génétique [39]. Exemple, dans [33], un premier classifieur SVM est d'abord construit en utilisant les exemples étiquetés disponibles, et ensuite le séparateur est utilisé pour la prédiction des données non étiquetées. Les auteurs proposent d'examiner la marge entre les deux classes de données (les données étiquetées et nouvellement étiquetées). Donc la limite de décision optimale est celle qui présente une erreur d'apprentissage minimale sur les données étiquetées et non étiquetées.

8 Conclusion

Dans ce chapitre, nous avons introduit le concept d'apprentissage semi-supervisé et nous avons présenté une revue de quelques travaux de la littérature. L'apprentissage semi-supervisé est un domaine vaste avec une littérature étendue. Il est impossible de couvrir tous les aspects de l'apprentissage semi-supervisé dans ce chapitre. De ce fait, nous avons présenté les méthodes souvent utilisées dans ce contexte. Cependant, la plupart des algorithmes de classification semi-supervisée sont des extensions de méthodes populaires d'apprentissage supervisé, et qu'ils sont déjà fréquemment utilisés dans de nombreuses applications.

Dans cette thèse, nous partons de l'idée que lorsque différentes méthodes peuvent être utilisées pour exploiter une quantité minime de données étiquetées et faire prédire une large quantité de données non étiquetées, pourquoi ne pas se focaliser sur un modèle basé sur la simplicité, la rapidité et l'efficacité qui est bien connu par l'auto-apprentissage, notre objectif est de répondre à la question suivante « Le modèle d'auto-apprentissage est-il pertinent ? ». Pour cela, la première partie de cette thèse est consacré à l'approfondissement du principe de l'auto-apprentissage appliqué sur des données médicales. Amplifiant la complexité et la difficulté du problème dans la deuxième partie de cette thèse, ceci réside dans la segmentation des images médicales. Dans cette deuxième partie, nous discutons les limites de l'auto-apprentissage et comment ces limites peuvent être soulevées en intégrant les techniques ensemblistes pour la classification.

CHAPITRE 2

L'impact de la mesure de similarité en auto-apprentissage

1 Objectifs

Le besoin croissant d'automatiser les tâches requérant une expertise humaine et la complexité ainsi que le coût de l'étiquetage manuel, ont conduit à l'utilisation de méthodes semi-supervisées pour lesquelles une faible quantité étiquetée d'informations est fournie. En effet, plusieurs approches de classification semi-supervisée ont vu le jour ces derniers années. Dans ce travail nous nous sommes intéressés plus particulièrement à l'auto-apprentissage. Ces techniques d'auto-apprentissage utilisent le même principe que celles supervisées mais avec une mesure de confiance qui permet de sélectionner uniquement les données classées avec un grand degré de confiance. Pour le calcul de confiance, la majorité des travaux se sont basés sur la mesure de similarité. Dans ce chapitre, nous nous penchons sur l'étude de la pertinence des différentes mesures de similarités sur les performances de classification en auto-apprentissage.

Dans ce chapitre, nous nous intéressons aux approches de classification semi-supervisée (auto-apprentissage). En auto-apprentissage un classifieur supervisé est mis en application pour prédire les données non étiquetées. Ensuite une mesure de confiance est effectuée afin de sélectionner uniquement les données qui sont classées avec un grand degré de confiance. Cette opération sera répétée jusqu'à satisfaction d'un critère d'arrêt.

Dans la littérature il existe plusieurs algorithmes d'auto-apprentissage tel que :

– SETRED (self training with editing) [6]
– Approche de classification semi supervisée basée sur le plus proche voisin [48]
– COSTRA (Confidence-based Self-Training) [49]
– Auto-apprentissage utilise le classifieur KNN renforcé par SVM [50]
– SNNRCE (Semi-supervised learning based on nearest neighbor rule and cut edges) [7].

Celui qui a retenu notre attention est le SNNRCE et cela est dû à son principe de fonctionnement. L'algorithme de SNNRCE se base sur le principe de la règle du plus proche voisin et la génération du graphe de voisinage pour faire la classification. De ce fait, la mesure de similarité est effectuée entre les données d'apprentissage à chaque étape de l'algorithme. D'après notre hypothèse, cette mesure a un impact réel sur les performances de classification en apprentissage semi supervisé. L'idée étant de faire le bon choix de métrique de similarité pour améliorer les résultats.

Dans ce chapitre, nous étudions l'influence des différentes mesures de similarité dans l'algorithme d'auto-apprentissage SNNRCE. Pour cela nous avons réparti le reste de ce chapitre comme suit : dans la section 2 nous exposons l'état de l'art du domaine. Ensuite dans la section 3 viendra la partie théorique de SNNRCE. Dans la section résultats nous mettons en valeur les différentes mesures appliquées et montrons clairement leurs importances sur les performances de cet algorithme. A la fin, nous clôturons le présent chapitre par une conclusion qui présente une synthèse de l'étude réalisée.

2 État de l'art du domaine

2.1 Classification semi-supervisée

Dans cette section nous introduisons la problématique de la classification semi-supervisée et les notations utilisées. Nous considérons les données non étiquetées D_{NE} de taille $|D_{NE}|$, les données étiquetées D_E de m classes et de taille $|D_E|$, avec x_i les instances et y_i les étiquettes, soit $D_{NE} \gg D_E$. L'ensemble $D_E \cup D_{NE}$ forme les données d'apprentissage notées D_A. Le but de la classification semi supervisée est de construire une hypothèse robuste en utilisant les données D_E afin de classer les données de D_{NE} avec une grande précision. Cette hypothèse peut être utilisée dans deux types d'apprentissage : L'apprentissage transductif et L'apprentissage inductif.

Dans l'apprentissage transductif on utilise les données D_E pour prédire les étiquettes manquantes de D_{NE}. La phase d'apprentissage inductif consiste à classer correctement les données de test D_T en employant l'hypothèse apprise dans la phase transductive.

2.2 Auto-apprentissage

Dans l'auto-apprentissage (self training) [50] nous entrainons un classifieur supervisé avec les données de D_E. Ensuite ce classifieur est utilisé pour prédire les étiquettes manquantes de D_{NE}. Les données étiquetées avec un haut degré de confiance sont ajoutées à D_E. Le classifieur est ré-entrainé sur les nouvelles données de D_E et cette procédure est répétée jusqu'à satisfaire un critère d'arrêt (convergence).

Dans [10], Isaac Triguero et al. ont présenté un état de l'art de tous les algorithmes semi-supervisés existants, ils ont classé les algorithmes d'auto-apprentissage en trois catégories : *incremental*, *batch* et *amending*.

Mode incrémental : Les algorithmes semi supervisés de catégorie incrémentale commencent par traiter les données de D_{NE} instance par instance si l'une d'elle vérifie un certain critère d'addition, elle sera ajoutée à D_E. Généralement dans ce type d'algorithme le degré de confiance est calculé par la probabilité d'appartenance à chaque classe. L'un des points les plus importants est le nombre d'instances ajouté à chaque itération. Ce nombre peut être défini comme une constante ou bien il peut être choisi comme un nombre proportionnel au nombre d'instances de chaque classe. L'avantage de cette catégorie est la rapidité du calcul au cours de la phase d'apprentissage. Par contre, cette catégorie peut ajouter des instances erronées lors d'apprentissage. On note que l'algorithme standard d'auto-apprentissage [50] appartient à cette catégorie.

Mode batch : Dans cette catégorie le test se fait sur toutes les données de D_{NE} et les instances qui vérifient les critères d'addition sont ajoutées à la fois à D_E. Cette catégorie souffre de la lenteur de calcul comme l'APSSC [51].

Mode amending : Cette catégorie est apparue comme une solution à l'inconvénient de la catégorie incrémentale. Les algorithmes de cette catégorie sont itératifs et peuvent ajouter ou supprimer une instance de D_{NE} qui vérifie un certain critère spécifique et même

permettent de corriger quelques étiquettes qui sont déjà ajoutées à D_E selon d'autre critère. Cette catégorie évite l'ajout des instances bruitées à D_E. SETRED [6] et SNNRCE [7] sont les algorithmes de référence de cette catégorie.

Les travaux actuels [7], [52], [48], [49] et [10] se sont focalisés d'avantage sur la conception d'une méthode d'auto-apprentissage fiable que sur la recherche de métriques pertinentes pour la mesure de confiance, et présentent des mesures simples souffrants de cas dégénérés, et d'un manque de pouvoir discriminant. Dans ce contexte, l'objectif de ce travail est de proposer une évaluation des différentes métriques existantes et relever la plus pertinente. Ce chapitre introduit principalement une étude approfondie et une comparaison des différentes métriques existantes adressées au problème de mesure de similarité, dans cette étude, nous discutons l'influence de la métrique de distance dans la classification semi supervisé en utilisant l'algorithme d'auto-apprentissage SNNRCE.

L'objectif majeur est de comparer les mesures de similarités pour un type de données. Pour cela, il faut choisir sur quel point de vue on souhaite comparer les mesures. Plusieurs méthodologies sont possibles. Par exemple, on peut comparer les mesures d'un point de vue applicatif comme l'ont fait Penney et al. [53] pour des images médicales ou Chang et al. [54] dans un algorithme de clustering appliqué à des données issues d'électrocardiogrammes ou bien encore Weken et al. [55] et [56] pour comparer des images modélisées par des ensembles flous. Nous avons opté pour le premier angle de comparaison qui considère une comparaison "quantitative" des mesures de similarité.

3 Matériels et méthodes

3.1 Principe de l'algorithme SNNRCE

A partir de l'étude bibliographique, nous avons choisi l'algorithme SNNRCE qui a été proposé par Wang et al. en 2010 [7]. Cet algorithme a un potentiel et une manière différente de faire la classification par rapport aux autres algorithmes.

C'est un algorithme de classification semi supervisé de type auto-apprentissage. Il est divisé en quatre parties où le niveau de confiance de classification dans la première partie est bien élevé que celle de la deuxième partie. La troisième partie prend en charge les corrections nécessaires et les données restantes seront classées à la fin. Cet algorithme fait appel à deux techniques :

Règle du plus proche voisin

La règle du plus proche voisin (ppv) a été proposée par Fix et Hodges [57], c'est une méthode non paramétrique, où la règle de classification est obtenue en posant que la classe d'une donnée non étiquetée est celle de la plus proche parmi les étiquettes de ses voisins dans les échantillons d'apprentissage. La détermination de leur similarité est basée sur des mesures de distance.

Par la suite, cette règle a été développée en $k-ppv$, elle est basée sur la définition d'un nombre k de voisinage et l'étiquette d'une donnée non-classée est celle qui est majoritaire parmi les étiquettes de ses k plus proches voisins.

Graphe de voisinage relatif

Le graphe de voisinage est un outil issu de la géométrie computationnelle qui a été exploité dans de nombreuses applications d'apprentissage automatique. Par définition un graphe de voisinage $G = (S, E)$ [58] associé à un ensemble de données dont les sommets « S » composant l'ensemble des arêtes « E ».

Chaque donnée dans un graphe de voisinage est représentée par des sommets, il existe des arêtes entre les sommets x_i et x_j si la distance entre ces derniers satisfait l'équation 2.1.

$$(x_i, x_j) \in E \Leftrightarrow dist(x_i, x_j) \leq max\ (dist(x_i, x_k), dist(x_j, x_k)), \forall\ x_k \in \mathbb{TR}, k \neq i, j. \tag{2.1}$$

Avec : $dist(x_i, x_j)$: la distance entre x_i et x_j.

Cette définition signifie qu'il n'existe pas de sommet à l'intérieur de l'intersection de deux cercles de centre x_i et x_j ainsi de rayon $dist(x_i, x_j)$.

D'après la définition ci-dessus, on peut construire un graphe de voisinage relatif dans lequel le voisinage de chaque donnée est un ensemble d'échantillons connectés avec des arêtes. Muhlenbach et al. [4] ont exploité l'information des arêtes pour calculer un poids statistique afin de couper les arêtes des données connexes de différente classe comme dans le cas d'algorithme de filtrage des données bruités [4] ou comme la classification et les mesures de confiance suivant le principe de SNNRCE.

La première partie de l'algorithme SNNRCE est basée sur une construction d'un graphe de voisinage. Pour ce faire, un calcul d'une simple distance est effectuée entre chaque donnée non annotée par rapport à l'ensemble annotée, si une donnée vérifie l'équation 2.1 une connexion sera reliée entre (x_i, x_j). A la fin, un calcul d'un rapport du poids R_i entre les données connexes est fait selon les équations (2.1, 2.2,2.3,2.4 et 2.5) si le rapport est égal à « 1 » cela signifie que toutes les données connexes sont de même classe, donc chaque donnée non annotée a un rapport $R_i = 1$ sera classée comme ses voisins du graphe.

$$R_i = \frac{J_i}{I_i} \tag{2.2}$$

Avec :

$$I_i = \sum_{(j \in Neighborhood(x_i))} w_{ij} \tag{2.3}$$

$$J_i = \sum_{(j \in Neighborhood(x_i), y_j \neq y_i)} w_{ij} \tag{2.4}$$

$$w_{ij} = \frac{1}{((1 + dist(x_i, x_j)))} \tag{2.5}$$

Dans la deuxième partie, un degré de confiance est calculé à partir de l'équation 2.6, cette mesure est fondée sur la distance du plus proche voisin, le résultat est le rapport entre la distance minimale appartenant à la classe $+$ et l'autre de la classe $-$.

$$CL(x_i) = \frac{exp(-dist(x_i, x^+_{1NN(x_i)}))}{exp(-dist(x_i, x^-_{1NN(x_i)})} \tag{2.6}$$

Si la quantité de $CL(x_i)$ est relativement importante la classe de x_i est de classe « $+$ » et vice versa, à chaque itération, deux données seront classées possédant la valeur maximale et la valeur minimale de CL 2.6.

La troisième partie de cet algorithme prend en charge les corrections nécessaires en utilisant le même principe de la première partie (les mêmes calculs), si le rapport R_i est inférieur à un certain seuil la classe de x_i sera corrigée par la classe opposée. Le reste des données sera classé dans la quatrième partie en appliquant la règle du plus proche voisin, l'algorithme 3 résume toutes les étapes de SNNRCE.

Algorithm 3 : Algorithme de *SNNRCE*

1: **Entrée** D_E : Données étiquetées avec deux classes D_E^+, D_E^-
D_{NE} : Données non-étiquetées
Critère d'arrêt : $N_{max}^+ = |D_E^+|/|D_E||D_{NE}|$, $N_{max}^- = |D_E^-|/|D_E|.|D_{NE}|$.
k : nombre de voisins
Seuil
2: **Étape 1**
- *Pour* chaque données D_{NE}
- Calculer le graphe de voisinage Eq.2.1
- Calculer le rapport R_i Eq.2.2
- Classer les données qui ont $R_i = 1$
- Ajouter les données classées à la base étiquetée
- *Fin pour*
3: **Étape 2**
- *Tant que* $|D_E^+| < N_{max}^+ et |D_E^-| < N_{max}^-$
  - *Pour* chaque données D_{NE}
  - Calculer CL_i Eq.2.6
  - *Fin pour*
- Classer les données qui ont CL_{max} et CL_{min}
- Ajouter les données classées à la base étiquetée
- *Fin tant que*
4: **Étape 3**
- Calculer le graphe de voisinage Eq.2.1
- Calculer le rapport R_i
- Corriger les données qui ont $R_i < seuil$
5: **Étape 4** Classer les données restantes en utilisant la règle du plus proche voisin
6: **Sortie :** les classes prédites de D_{NE}.

3.2 Les mesures de similarité

Dans l'apprentissage supervisé et semi-supervisé l'exploitation des similarités entre les échantillons est une opération commune. Les métriques de similarité varient d'un domaine d'application à un autre. Selon le domaine d'application, on peut trouver différentes relations de similarité entre des exemples tels que la distance euclidienne pour la distance droite entre deux points, la similarité cosinus pour les vecteurs binaires dispersés à haute dimension, cardinalité de l'intersection de deux ensembles, etc. D'autre part, il n'est pas toujours possible de traiter uniformément tous les attributs d'une instance. Dans de telles situations, des mesures de distance composites peuvent être nécessaires.

En pratique, la similarité est évaluée, en général par une mesure de similarité ou une distance (l'ordre donné est alors inversé). La multitude de mesures de similarité existantes dans la littérature (des nouvelles mesures sont toujours proposées aujourd'hui) est à mettre en rapport avec la multitude de méthodes et de domaines où une mesure de similarité intervient. Ce constat nous motive à proposer des moyens pour comparer des mesures de similarité afin de mieux appréhender leur comportement et de choisir la bonne mesure de similarité.

Dans ce qui suit, nous présentons succinctement les différentes métriques existantes et les plus connues utilisées pour la mesure de similarité. Nous donnons les définitions suivantes pour deux vecteurs T et U, chacun ayant la dimension n.

Similarité basée sur la distance euclidienne

La distance euclidienne entre deux vecteurs T et U de dimension n est :

$$DEu(T,U) = \sqrt{(T_1 - U_1)^2 + (T_2 - U_2)^2 + ... + (T_n - U_n)^2}$$

Sachez que la Distance Euclidienne (DEu) est calculée à partir des données brutes de n dimension, et non des données centrées-réduites. C'est la distance la plus populaire utilisée dans les concepts de classification. Toutefois, cette distance peut être largement affectée par les différences d'unités de mesure des dimensions pour lesquelles ces distances sont calculées. Exemple : le passage de dimension du centimètre au millimètre peut influencer la mesure de distance, et par conséquent, les résultats de la classification pourront être très différents.

Mesure de similarité par Minkowsky :

La distance de Minkowsky est une métrique normalisée dans l'espace vectoriel qui peut être considérée comme une généralisation des deux distances Euclidienne et Manhattan.

$$DMin(T,U) = \sqrt[p]{\sum_{i=1}^{n} |T_i - U_i|^p}$$

Mesure de similarité de Manhattan :

Pour $p = 1$ de la distance Minkowsky, on obtient la distance de Manhattan (aussi appelée distance "city-block" ou métrique absolue) :

$$DMan(T,U) = |(T_1 - U_1) + (T_2 - U_2) + ... + (T_n - U_n)|$$

C'est une distance est considérée comme la somme des différences entre les dimensions. Dans la plupart des cas, cette mesure de distance produit des résultats proches de ceux obtenus par la distance euclidienne simple.

Mesure de similarité par Chebychev :

La distance de Chebychev (aussi appelée distance « Queen-wise »ou encore métrique maximum) :

$$DChe(T,U) = \max\Big((T_1 - U_1), \ldots, (T_n - U_n)\Big)$$

La mesure de similarité cosinus

La mesure cosinus permet de calculer la similarité entre deux vecteurs à n dimensions en déterminant le cosinus de l'angle entre eux. Cette métrique est fréquemment utilisée en fouille de textes.

Soit deux vecteurs T et U, l'angle $\cos\theta$ s'obtient par le produit scalaire et la norme des vecteurs :

$$\cos\theta = \frac{T \cdot U}{\|T\| \cdot \|U\|}.$$

Comme la valeur $\cos\theta$ est comprise dans l'intervalle [-1,1], la valeur -1 indiquera des vecteurs résolument opposés, 0 des vecteurs indépendants (orthogonaux) et 1 des vecteurs similaires (colinéaires de coefficient positif). Les valeurs intermédiaires permettent d'évaluer le degré de similarité.

Coefficient de corrélation de rang de Spearman R_s

C'est le premier test statistique réalisé à partir des rangs. Cette statistique est appelée R_s. Elle mesure l'association entre deux variables mesurées au moins dans une échelle ordinale.

Son principe commence par classer séparément les valeurs de la variable T et celles de la variable U. Supposons qu'il existe une liaison linéaire positive entre T et U, le calcul des rangs permettra de se rendre compte que les sujets qui ont les plus petites valeurs de T ont également les plus petites valeurs de U, et inversement les sujets qui ont les plus grandes valeurs de T ont également les plus grandes valeurs de U. En revanche, s'il n'existe aucune liaison linéaire entre T et U, les sujets qui ont les plus petites valeurs de T ont des valeurs de U dispersés dans leur classement [59].

$$R_s = 1 - \frac{6 \sum di^2}{n(n^2 - 1)}$$

di représente la différence des rangs au niveau de l'observation i.

Mesure de corrélation

Les coefficients de corrélation permettent de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est monotone. Le coefficient de corrélation de Pearson permet d'analyser les relations linéaires et le coefficient de corrélation de Spearman les relations non-linéaires monotones [59]. La mesure de corrélation se calcule ainsi :

$$R(T, U) = \frac{Cov(T, U)}{\sigma_T . \sigma_U}$$

La distance de Hamming

La distance de Hamming est une distance au sens mathématique du terme, elle calcule le nombre d'éléments différents dans les vecteurs binaires :

$\forall a, b \in F : d(a, b) = d(b, a)$ (symétrie)

$\forall a, b \in F : d(a, b) = 0 \Leftrightarrow a = b$ (séparation)

$\forall a, b, c \in F : d(a, c) \leq d(a, b) + d(b, c)$ (inégalité triangulaire)

Coefficient de Jaccard

Le coefficient de Jaccard est défini comme étant le quotient du cardinal de l'intersection par celui de l'union [60]. Les coefficients sont calculés par la formule suivante :

$$simJaccG(T, U) = (T.U)/(|T|^2 + |U|^2 - (T.U))$$

4 Résultats et Interprétations

Dans cette section, nous présentons des résultats empiriques sur plusieurs données de référence et des données réelles de petites et moyennes dimensions. Huit ensembles de données principalement choisis dans le répertoire de Machine Learning UCI [61], et de l'ASU feature selection Repository [62]. Ces derniers sont utilisés pour étudier l'impact des différentes mesures d'importance sur la performance de *SNNRCE*, leurs caractéristiques sont décrites dans le tableau 1.

Pour chaque ensemble de données, une validation croisée égale à cinq (5 cross validation) est effectuée pour l'évaluation. Les données d'apprentissage sont aléatoirement divisées en deux ensembles : L labellisé et U non labellisé fixés par un taux μ, qui est calculé par la taille de U sur la taille de $L \cup U$. Quatre différents taux de « labellisation » $\mu =$ 20%, 40%, 60% et 80%, sont étudiés. Aussi nous prenons note que les distributions de classe pour L et U sont maintenues similaires à celle de l'ensemble original.

Bases	**#instances**	**#variables**	**#Classes**
Colon	64	2000	2
Heberman	306	3	2
Heart	270	13	2
Hepatitis	155	20	2
Leukemia	73	7129	2
Ovarian	54	1536	2
Pancreatic	119	6771	2
Prostate	102	12533	2

Table 1 – Description des bases d'expérimentation

Dans les expérimentations, les résultats obtenus ont été réalisés avec les paramètres suivants :$k = 5$ pour le nombre de voisinage, le seuil de confiance θ est fixé à 0,75 et $p = 2$ pour la distance de Minkowsky.

Afin d'étudier l'impact de la mesure de similarité en auto-apprentissage, plusieurs mesures de similarité on été utilisées dans chaque étape de mesure de *SNNRCE*. Les tests ont été réalisés sur des ensembles de données médicales à dimension variée. Les performances de *SNNRCE* pour les différents taux de μ sont rapportées dans les tableaux 2, 3, 4 et 5.

Techniques/Bases	*Colon*	*Heberman*	*Heart*	*Hepatitis*	*Leukemia*	*Ovarian*	*Pancreatic*	*Prostate*
euclidean	59,3103	70,3560	58,8889	74,9367	76,8730	74,8148	50,2174	66,9231
cityblock	46,8966	71,3930	62,7778	76,4557	75,7619	73,3333	49,1304	**69,6154**
minkowski	59,3103	70,3560	58,8889	74,9367	76,8730	74,8148	50,2174	66,9231
chebychev	60,6897	67,4546	58,7500	74,6835	64,9524	52,5926	49,1304	56,1538
cosine	55,8621	**71,6923**	61,2500	73,4177	80,1905	71,1111	**51,9565**	67,6923
correlation	60,6897	64,1720	60,4167	73,4177	81,3333	70,3704	**51,9565**	67,3077
spearman	58,6207	69,6888	65,6944	77,4684	**85,3492**	69,6296	51,7391	62,6923
hamming	**65,5172**	63,3553	74,5833	**80,7595**	65,5238	69,6296	49,1304	57,3077
jaccard	**65,5172**	57,6835	**78,4722**	75,1899	67,1905	68,8889	49,1304	59,6154

Table 2 – Performances moyennes de SNNRCE avec différentes mesures de similarités sur l'ensemble des données avec un taux de labellisation $\mu = 20\%$

Techniques/Bases	*Colon*	*Heberman*	*Heart*	*Hepatitis*	*Leukemia*	*Ovarian*	*Pancreatic*	*Prostate*
euclidean	50,4762	72,9390	60,0000	72,8814	87,5692	76,8421	47,0588	76,3158
cityblock	48,5714	**73,2825**	65,5556	74,2373	87,5385	82,1053	51,1765	70,5263
minkowski	50,4762	72,9390	60,0000	72,8814	87,5692	76,8421	47,0588	**76,3158**
chebychev	60,9524	72,5812	60,7407	72,2034	76,7692	48,4211	52,3529	69,4737
cosine	60,9524	72,5812	60,7407	72,2034	76,7692	48,4211	52,3529	69,4737
correlation	61,9048	67,5019	63,8889	74,2373	86,8000	73,6842	**56,7647**	74,7368
spearman	**67,6190**	69,6996	72,9630	79,3220	**91,4769**	72,6316	55,8824	73,1579
hamming	65,7143	72,1976	**78,3333**	79,6610	65,9385	**84,2105**	50,2941	58,9474
jaccard	65,7143	70,0915	77,9630	**83,3898**	63,6308	83,1579	50,2941	65,2632

Table 3 – Performances moyennes de SNNRCE avec différentes mesures de similarités sur l'ensemble des données avec un taux de labellisation $\mu = 40\%$

Techniques/Bases	*Colon*	*Heberman*	*Heart*	*Hepatitis*	*Leukemia*	*Ovarian*	*Pancreatic*	*Prostate*
euclidean	56,9231	72,2802	63,6111	72,3077	92,5000	80,0000	51,3636	78,3333
cityblock	56,9231	72,5469	68,0556	75,8974	88,7500	76,9231	51,8182	**82,5000**
minkowski	56,9231	72,2802	63,6111	72,3077	92,5000	**80,0000**	51,3636	78,3333
chebychev	**69,2308**	**74,2077**	63,8889	73,8462	86,2500	46,1538	53,6364	66,6667
cosine	58,4615	71,0779	63,8889	76,9231	90,0000	69,2308	49,5455	78,3333
correlation	58,4615	64,2681	61,6667	76,9231	90,0000	69,2308	48,1818	78,3333
spearman	**69,2308**	72,5456	70,5556	82,0513	**92,5000**	72,3077	52,7273	71,6667
hamming	66,1538	72,5456	**78,3333**	**82,5641**	76,2500	76,9231	**55,9091**	58,3333
jaccard	66,1538	72,2753	76,9444	81,0256	76,2500	76,9231	**55,9091**	60,8333

Table 4 – Performances moyennes de SNNRCE avec différentes mesures de similarités sur l'ensemble des données avec un taux de labellisation $\mu = 60\%$

Techniques/Bases	*Colon*	*Heberman*	*Heart*	*Hepatitis*	*Leukemia*	*Ovarian*	*Pancreatic*	*Prostate*
euclidean	56,0000	73,2551	63,1111	75,4717	88,4615	83,3333	66,0000	80,0000
cityblock	64,0000	74,3361	68,4444	75,8491	91,5385	82,2222	48,0000	78,0000
minkowski	56,0000	73,2551	63,1111	75,4717	88,4615	**83,3333**	**66,0000**	**80,0000**
chebychev	64,0000	**74,4315**	61,1111	75,0943	82,3077	51,1111	60,0000	64,0000
cosine	64,0000	68,0959	62,4444	76,6038	90,7692	72,2222	46,0000	74,0000
correlation	68,0000	68,0849	63,3333	76,9811	91,5385	71,1111	50,0000	74,0000
spearman	**80,0000**	73,2551	72,6667	**85,2830**	**94,6154**	73,3333	57,0000	72,0000
hamming	**80,0000**	72,6668	75,3333	82,6415	67,6923	82,2222	58,0000	64,0000
jaccard	**80,0000**	72,6668	**80,8889**	81,5094	67,6923	**83,3333**	58,0000	58,0000

Table 5 – Performances moyennes de SNNRCE avec différentes mesures de similarités sur l'ensemble des données avec un taux de labellisation $\mu = 80\%$

A partir des tableaux 2, 3, 4 et 5 nous constatons que les mesures de similarité autres que la mesure classique *euclidienne* sont en mesure d'améliorer les performances de *SNNRCE* à partir de données non étiquetées, cela concerne les différents taux de non labellisation. Les performances moyennes sur les quatre taux μ des différents tableaux ont été remarquablement améliorées respectivement.

Nous noterons également par les mesures comme *Hamming* et *Jaccard*, *SNNRCE* a des performances nettement supérieures à celles de la métrique *euclidienne* sous les différents taux de μ.

Ceci peut être expliqué par le fait que, la mesure de similarité de *Hamming* est une mesure qui prend en compte les 4 caractéristiques associées à un couple d'objets, à savoir, leur intersection, leur différence ensembliste et l'intersection de leurs complémentaires. C'est sur ce point où réside la différence entre la mesure de *Hamming* et celle de *Jaccard* qui quant à elle se base sur les cardinalités. Ceci confirme une expérience antérieure qui a réalisé une amélioration remarquable pour les bases de données à volume important comme celle de *Colon*, *Leukemia*, *Prostate*.

Dans un autre cas de figure, les mesures de *Hamming*, *Spearman* et *Jaccard* réalisent des performances considérables même dans les conditions les plus extrêmes avec un taux de labellisation μ=80% ; ce qui nous ramène en citant la base *Colon* par exemple, avec seulement 13 individus labellisés a pu réaliser une amélioration de 6.2% sur la performance moyenne.

5 Conclusion

Par ce travail, nous espérons avoir contribué à faire avancer la question de l'évaluation de comparaison des méthodes en vigueur dans le domaine de l'apprentissage semi-supervisé, au sens large, tout en montrant que le choix d'une mesure de similarité est souvent jugé crucial. Nous avons proposé un cadre formel pour pouvoir analyser et comparer différentes mesures de similarité afin de permettre un choix faisant coïncider autant que possible le but de poursuivre le comportement de la mesure.

Il serait nécessaire de prolonger cette recherche par des comparaisons par rapport à d'autres mesures de similarité, comme l'écart réduit ou ceux issus des théories de l'information.

CHAPITRE 3

Renforcement de la confiance en auto-apprentissage

1 Objectifs

La complexité ainsi que le coût de l'étiquetage manuel des données posent un besoin croissant d'automatiser les tâches requérant une expertise humaine essentiellement dans des applications médicales. Cependant, la mise en œuvre du semi-supervisé est devenue cruciale, son avantage réside sur le fait qu'il requiert une faible quantité étiquetée d'informations. En effet, plusieurs approches de classification semi-supervisée ont vu le jour ces derniers temps. Dans ce chapitre nous nous sommes intéressés plus particulièrement à l'auto-apprentissage. Ces techniques d'auto-apprentissage utilisent le même principe que celles du supervisées mais avec une mesure de confiance qui permet de sélectionner uniquement les données classées avec un grand degré de confiance.

Dans le contexte médical, les méthodes modernes d'acquisition des données médicales permettent d'obtenir de nombreuses variables sur de nombreux patients avec un faible coût. Toutefois, l'annotation des variables d'intérêt est souvent plus difficile à obtenir, car cette difficulté est due à plusieurs causes (l'expertise humaine, manque des experts compétents, temps, coûts ...). Ceci est particulièrement réel dans les problèmes de la classification supervisée. Et généralement, les données non-expertisées sont plus nombreuses que les données expertisées.

Dans ce chapitre, un nouvel algorithme d'auto-apprentissage est proposé, nommé *R-COSET* (Reinforced confidence in self-training). *R-COSET* introduit le graphe de voisinage relatif sur deux niveaux de construction, et la technique de *CEWS* [4] dans la mesure du niveau de confiance, dans le but d'améliorer les performances de classification.

Les avantages de cette proposition portent sur la simplicité et l'efficacité de sélection des échantillons de confiance dans l'apprentissage semi-supervisé. Concernant cette sélection, une nouvelle formule de confiance est proposée, elle se base sur le poids statistique de graphe de voisinage, cette mesure a permis d'améliorer la classification semi-supervisée.

Ce chapitre introduit l'algorithme *R-COSET*, ce qui nous mène à organiser ce chapitre comme suit : Une revue de quelques méthodes d'auto-apprentissage dans le domaine semi-supervisé est effectuée. Nous présentons en détail l'évolution de ces dernières ainsi que leurs avantages et leurs limites. Nous exposons ensuite dans la section 3, le processus général de notre approche proposée et ses différentes étapes. Nous validons notre algorithme et les choix que nous avons réalisés par une phase d'expérimentation. Nous montrons la capacité de notre méthode dans l'amélioration des performances de classification, ceci par une comparaison avec les méthodes représentatives de la littérature.

Finalement, nous terminerons par une conclusion présentant une synthèse des contributions apportées ainsi que les pistes définissant des perspectives possibles pour de futurs travaux.

2 Le semi supervisé par le principe d'auto-apprentissage

L'idée principale de la classification semi-supervisée est d'exploiter un petit nombre de données étiquetées D_E pour construire une hypothèse robuste qui peut prédire correctement les classes des données non étiquetées D_{NE}.

La classification semi-supervisée peut être appliquée dans deux concepts légèrement différents : apprentissage transductif et apprentissage inductif. L'apprentissage transductif a pour but de classer les instances non marquées D_{NE} de l'ensemble d'apprentissage D_A, où $D_A = D_E \cup D_{NE}$. Ceci est en contraste avec l'apprentissage inductif, qui s'intéresse à construire une fonction de prédiction finale utilisant les données issues de la phase transductive. L'application de cette fonction consiste à classer correctement les instances de l'ensemble de test D_T, ce qui n'a pas été utilisé lors de l'apprentissage semi supervisé.

Le premier développement concernant l'utilisation des données étiquetées et non étiquetées est l'algorithme d'auto-apprentissage défini par D. Yarowsky [14], ce dernier est appliqué pour la détection de désambiguïsation des textes anglais en utilisant un classifieur appris sur les données étiquetées et non étiquetées.

Le processus d'auto-apprentissage "Self-Training", comme c'est décrit dans la figure 2. Dans l'auto-apprentissage, un classifieur est d'abord généré avec la petite quantité de données étiquetées. Le classifieur est ensuite utilisé pour classer les données non étiquetées. Les points non labellisés les plus confiants avec leurs étiquettes prédites sont ajoutés à l'ensemble d'apprentissage. Le classifieur est reconstruit sur l'ensemble de ces nouvelles données. A noter que le classifieur utilise ses propres prédictions pour le réapprentissage. Dans l'algorithme standard d'auto-apprentissage, la règle de probabilité à posteriori est appliquée pour mesurer le niveau de confiance sur les exemples non étiquetés [14].

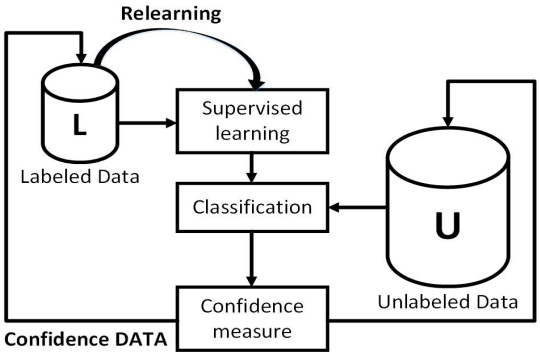


Figure 2 – Processus d'auto-apprentissage.

Le processus standard d'auto-apprentissage peut enrichir la base d'apprentissage par les données non étiquetées nouvellement labélisées de façon incrémentale, les exemples mal classés introduits à l'ensemble d'apprentissage à chaque itération peuvent conduire à des faibles taux de performances, en utilisant des données erronées dans le réapprentissage du processus. Bien que certaines techniques, telles que l'addition en mode *Batch* [6,7], a été employée pour atténuer ce problème. Triguero et al. [10] ont regroupé les mécanismes d'addition en trois catégories *Incremental*, *Batch* et *Amending* (Voir chapitre 1).

L'auto-apprentissage est une technique couramment utilisée en apprentissage semi-supervisé grace à sa simplicité et son efficacité. Plusieurs travaux ont été proposés dans la littérature pour améliorer ce processus. Dans ce contexte, nous citons :

Self-training with Editing (*SETRED*) [6] est un algorithme d'auto-apprentissage amélioré. La mesure de confiance de cet algorithme est basée sur le principe de filtrage des données bruitées. Cet algorithme appartient à la catégorie *Amending*, donc à chaque itération la méthode de filtrage *CEWS* [4] est appliquée, cela pour supprimer les données mal classées et identifier les données bien classées avec un certain degré de confiance.

Wang et al. ont développé une autre méthode d'apprentissage semi-supervisé (*SNNRCE*) [7], basée sur la règle du plus proche voisin et le principe de filtrage *CEWS* [4]. Le processus de cet algorithme est divisé en quatre étapes : la première étape exploite tous les échantillons étiquetés dans la construction d'un graphe de voisinage relatif pour chaque donnée non étiquetée, et si une donnée non étiquetée dans le graphe est reliée à des données de la même classe, cela veut dire que cette donnée est classée avec confiance. Les échantillons nouvellement classés sont ensuite ajoutés à la base d'apprentissage comme données étiquetées. Dans la deuxième étape, les auteurs ont mis au point une formule mathématique Eq.3.1 en fonction de la règle de plus proche voisin pour prévoir la confiance de la classification en calculant la quantité *CL*, cette étape est répétée jusqu'à satisfaire un critère d'arrêt.

$$CL(x_i) = \frac{exp(-dist(x_i, x^{+}_{1NN(x_i)}))}{exp(-dist(x_i, x^{-}_{1NN(x_i)})} \tag{3.1}$$

La troisième étape de cet algorithme est réservée pour effectuer des corrections si nécessaire, pour ce faire, il faut réappliquer l'algorithme *CEWS*. En dernière étape, les données non étiquetées restantes sont classées en utilisant la norme de la règle du plus proche voisin.

Un nouvel algorithme a été proposé par Liu et al. [48], à savoir ordinal semi-supervisé *k*-NN, pour traiter les cas avec quelques instances étiquetées. Cet algorithme se compose en deux parties : instance de classement et une partie semi-apprentissage supervisé (*SSL*). Utilisation de *SSL*, la performance de *k*-NN avec de petits ensembles de formation labellisés peut être améliorée, par le fait d'enrichir la base d'apprentissage par des données de confiance.

Dans une première phase, un classement d'instance est effectué, en se basant sur un calcul de la moyenne pondérée pour chaque donnée non étiquetée par rapport à chaque classe. Cette mesure donne la priorité aux cas non marqués qui sont les plus proches des frontières de décision, et qu'ils sont susceptibles d'être correctement prédit (éléments de confiance). La deuxième partie concerne l'application du mécanisme *SSL*, les prioritaires du classement effectué précédemment seront exploités par la règle de *k*-NN en auto-apprentissage, afin d'enrichir la base d'apprentissage par des données correctement classées.

Récemment, Chen et al. [49] ont développé un algorithme d'auto-apprentissage noté *COSTRA*, qui adopte le principe manifold assumption au processus d'auto-apprentissage. *COSTRA* est une méthode basée sur le graphe de voisinage avec la tolérance au bruit, cette dernière est utilisée pour aider à générer des prévisions fiables sur les données non marquées. En outre, afin d'éviter l'introduction de bruit de classification indésirable, un certain mécanisme est adopté pour augmenter séquentiellement l'ensemble d'apprentissage.

Wang et al. [63] ont abordé une approche d'auto-apprentissage pour la classification de la subjectivité de phrase, mettant en évidence la mesure de confiance en auto-apprentissage

comme un paramètre à discuter. Les auteurs ont mentionné que le classifieur du type Naive Bayes (NB) est souvent utilisé en raison de ces capacités prédictives, ainsi que les estimations de probabilité d'appartenance à une classe ont une bonne performance de classification. La contribution de ce travail est menée sur l'adaptation d'un modèle de VDM (Value Difference Metric) [64], comme un modèle de mesure de confiance en auto-apprentissage qui ne dépend pas de la probabilité d'appartenance des classes. L'idée de VDM est d'évaluer la distance entre les instances à partir des différences entre les estimations de probabilité conditionnelle des caractéristiques. Compte tenu de deux instances x et y, la distance VDM est définie par :

$$VDM(x,y) = \sum_{i=1}^{C} \sum_{j=1}^{N} |P(c_i|a_j(x)) - P(c_i|a_j(y)|\,, \tag{3.2}$$

Où C est le nombre d'étiquettes de classe, N est le nombre de variables dans les instances, et $P(c_i|a_j(x))$ est la probabilité conditionnelle caractéristique de la classe i donnée par la variable a_j de l'instance x.

L'utilisation de VDM en tant que mesure de confiance s'est révélée efficace dans la production de confiance et l'amélioration de l'auto-apprentissage.

Nikos Fazakis et al. [65] proposent un algorithme d'auto-apprentissage à base de Logistic Model Trees (LMT), combinant les caractéristiques de Logistic Trees dans un scénario sélectif des données mal classées. Le LMT est un arbre de décision qui a des modèles de régression linéaire à ses feuilles dans l'objectif de fournir un modèle de régression linéaire [66]. Pour ce faire, les auteurs ont fait appel à l'algorithme de LogitBoost [67] afin de produire un modèle de régression linéaire pour chaque nœud de l'arbre. Les arbres de décision peuvent générer des estimations probabilistes définissant une appartenance à une classe désirée : La probabilité pour une classe particulière est donnée par la fraction de l'instance étiquetée concernant la classe désirée. La mesure de confiance dans cette proposition est basée sur l'estimation de probabilité. Une donnée dite confiante si la probabilité surpasse un certain seuil, donc la donnée avec l'étiquette associée sera déplacée à la base étiquetée. Les résultats expérimentaux réalisés par les auteurs ont montré qu'une bonne option pour le paramètre de seuil est la valeur 0.9, qui a donné des résultats décents indépendamment de l'ensemble de données.

Analysant les travaux de la littérature [6, 7, 48, 49, 63] . . . , les algorithmes proposés dans cette discipline visent à améliorer les performances de classification et d'augmenter la précision de confiance lors de l'apprentissage semi-supervisé. En effet, comme c'est bien noté l'auto-apprentissage dépend essentiellement de l'efficacité de la mesure de confiance. Une mesure de confiance inexacte conduit à ajouter des exemples avec des étiquettes erronées à l'ensemble étiqueté, ce qui implique une dégradation de l'auto-apprentissage ainsi abîmant les performances de classification. Dans le contexte de confiance, plusieurs techniques ont été exploitées pour accroître l'estimation de confiance en auto-apprentissage. Nous remarquons une mesure souvent répétée dans plusieurs algorithmes [6, 7, 48, 49] qui est bien la règle de plus proche voisin (*NNR*), *NNR* a été initialement introduit par Fix et Hodges [68] pour résoudre les problèmes de discrimination. La théorie derrière le *NNR* est qu'une instance non marquée dans l'espace d'entrée est susceptible d'avoir la même classe que ses instances proches en terme de distance. Un autre point commun, le graphe de voisinage mentionné comme étant la base du calcul de confiance dans les algorithmes de référence en auto-apprentissage comme dans [6, 7, 48, 49]. Cette procédure se base elle-même sur la règle de *NNR* pour la construction du graphe de voisinage, donc l'attention du graphe de voisinage avec la règle de *NNR* reflètent les qualités de ces derniers dans l'amélioration des performances.

Dans notre travail, nous nous concentrons sur la mesure de confiance pour améliorer l'étape de l'étiquetage des données dans un processus d'auto-apprentissage, en particulier il est nécessaire de maximiser la confiance dans les premières itérations du processus.

Prenant en compte la mesure du graphe de voisinage et la règle *NNR* pour proposer une approche de classification semi-supervisée, en exploitant le maximum d'informations utiles dans un graphe de voisinage, les informations extraites seront utilisées pour faire apprendre un classifieur robuste qui peut enrichir l'ensemble d'apprentissage par des échantillons correctement classés.

3 Notre approche proposée «L'algorithme R-COSET»

L'algorithme présenté est une amélioration du principe standard d'auto-apprentissage pour la classification semi-supervisée. Nous nous basons dans notre proposition sur le graphe de voisinage pour la mesure de confiance, et ainsi apporter quelques corrections aux limites observées dans ce dernier ; mais avant d'aborder les étapes de notre approche, nous présentons d'abord le principe de graphe de voisinage.

3.1 Graphe de voisinage relatif

Pour exprimer la proximité entre des exemples dans l'espace de représentation, nous utilisons le principe de graphe de voisinage, ce dernier est un graphe dans l'espace de caractéristique de d-dimension où une métrique de distance peut être adoptée, comme il est présenté dans le chapitre 1, et chaque échantillon du graphe est représenté par un sommet. Dans le graphe de voisinage relatif, il existe une arête entre les sommets x_i et x_j, si la distance entre les deux sommets satisfaits à l'équation 2.1 (chapitre 2 section 3.1).

La définition de confiance de SETRED [6], SNNRCE [7] et COSTRA [49] est estimé par rapport aux arêtes des données non étiquetées, si les arêtes sont de classe unique, ils correspondent aussi à la classe prédite par le classifieur, donc cette donnée est prise comme élément de confiance. Ou bien, les arêtes sont comparées par rapport à un seuil de confiance sélectionné en préalable, la comparaison est menée en fonction du rapport R_i (Eq. 2.2) (chapitre 2 section 3.1).

En effet, si le rapport $R_i = 1$ cela signifie que toutes les arêtes sont de classe unique, cependant, cette mesure ne prend pas en considération le nombre des arêtes, ce qui peut être jugé comme une limite en calcul de confiance. Pour mieux saisir cela, nous proposons un exemple dans la figure 3. Nous considérons deux données non étiquetées X_1 et X_2, et à l'aide de l'équation 2.2 nous obtenons les rapports $R_1 = 1$ et $R_2 = 1$. Selon la mesure de confiance de SETRED [6], SNNRCE [7] et COSTRA [49], X_1 et X_2 sont considérées comme des éléments de confiance, logiquement, la donnée X_2 est susceptible d'être plus confiante en raison d'avoir quatre arêtes de même classe en comparaison avec X_1.

Dans les premières itérations d'auto-apprentissage, il est important d'enrichir la base étiquetée par des données classées avec un haut degré de confiance, dans le but de renforcer l'hypothèse de réapprentissage, comme dans l'exemple précédant, c'est la donnée X_2 qui doit être ajoutée en premier lieux. Dans notre proposition, nous avons considéré cette observation pour améliorer le calcul de confiance, cette considération sera détaillée dans la section suivante.

3.2 L'algorithme R-COSET

Notre proposition suit les étapes standard de l'algorithme d'auto-apprentissage, en premier lieu, une hypothèse supervisée sera entrainée sur la disponibilité de données étiquetées. Par la suite, l'hypothèse apprise est mise en exécution pour la classification des données non étiquetées. Ainsi, l'activation du mécanisme de confiance, suivant un calcul donné, nous permet de définir les éléments de confiance à partir des données nouvellement classées, cette sélection sera ajoutée à la base étiquetée. Suite à une procédure itérative, l'hypothèse initiale sera réapprise sur la base enrichie, refaire la classification

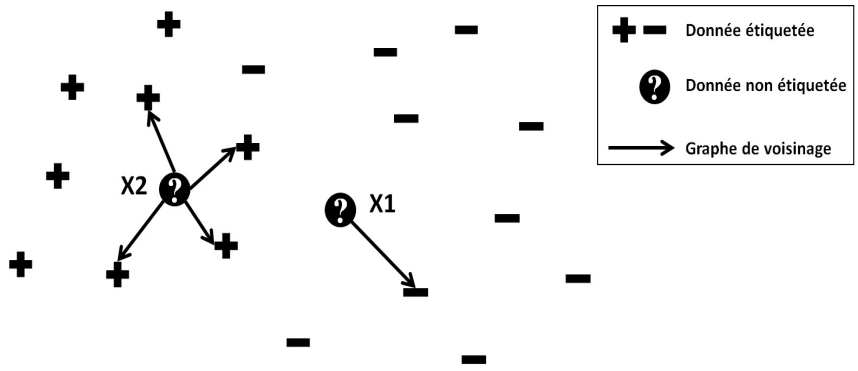


Figure 3 – Example du graphe de voisinage.

et la mesure de confiance jusqu'à la convergence de l'hypothèse.

Dans notre proposition, nous avons cherché à améliorer la mesure de confiance dans l'auto-apprentissage, on se basant sur les algorithmes de référence SETRED [6], SNNRCE [7] et COSTRA [49] dans l'utilisation du graphe de voisinage et de la technique CEWS (Cut Edge Weight Statistic) [4]. Une fois l'hypothèse initiale préparée, et les classes des données non étiquetées affectées à $D_{E'}$, nous proposons une procédure du calcul de confiance comme suit :

Au début, nous procédons à une construction d'un graphe de voisinage relatif [69]. Mettant en exécution l'équation 2.1 pour identifier les arêtes pour chaque donnée de $D_{E'}$. Le graphe construit exprime une certaine information géométrique de $D_{E'}$ en terme de similarité par rapport à D_E.

Dans le R-COSET, nous cherchons à améliorer l'aspect du graphe de voisinage pour une confiance élevée. Au lieu d'exploiter uniquement les informations du premier niveau du graphe, nous avons proposé de renforcer cela par le deuxième niveau du graphe de voisinage, c'est-à-dire, nous cherchons pour une deuxième exécution les arêtes du deuxième niveau à partir des sommets du premier niveau du graphe de voisinage. Cette proposition permettra de prendre une vue plus exhaustive en fonction de la géométrie du graphe.

Une décision sera prise si la donnée est considérée confiante, est évalué par la technique CEWS [4]. Cette dernière est mentionnée comme une technique de filtrage des données bruitées [5]. Dans CEWS, une arête qui relie deux sommets de classe différente est désignée comme une arête coupée en raison de la différence de classe. Ensuite, une procédure statistique est appliquée pour étiqueter les arêtes coupées comme exemples bruités. Il s'agit de la méthode de filtrage utilisée dans la mesure de confiance dans SETRED [6] et SNNRCE [7].

La statistique utilisée est basée sur le rapport R_i, en utilisant l'équation 2.2 pour mesurer le R_i de chaque donnée x_i, le résultat de cette mesure se situe entre l'intervalle $[0-1]$. Le tableau 6 représente une démonstration du rapport R_i.

R_i	**Démonstration**
$R_i = 1$	Tous les sommets ont la même classe par rapport à x_i
$R_i = 0$	Tous les sommets sont de classe opposée par rapport à x_i
$0 < R_i < 1$	Les sommets ont des classes différentes

Table 6 – La démonstration de différent rapport R_i.

Une fois les deux niveaux du graphe de voisinage construits, et le rapport R_i calculé pour chaque donnée x_i nouvellement classée, nous procédons à la mesure de confiance

MC. Dans R-COSET, *MC* est définie en tenant compte du nombre des arêtes du premier niveau du graphe de voisinage (nbr_{arte_i}) et le second niveau (nbr arête$_{ij}$), et également le R_i de chaque échantillon.

La mesure de confiance *MC* indique la probabilité qu'une donnée non marquée appartient à telle classe en fonction du graphe de voisinage et le rapport des arêtes. La mesure est effectuée par l'équation 3.3, une mesure élevée de *MC* implique une forte probabilité que la donnée non étiquetée provient de la classe prédite par le classifieur, tandis qu'une petite mesure de *MC* implique le cas contraire.

$$MC_i = exp - \left(\frac{nbr_{arete_i}.R_i}{sum(nbr_{arete_{ij}}.R_{ij})} \right) \tag{3.3}$$

La figure 4 représente un exemple illustrant la proposition de mesure de confiance en R-COSET, deux ensemble de donnée étiquetée sont disponible (décrit en symbole '+' et '−'), x_1 et x_2 deux données non étiquetées impliquées dans l'espace de représentation.

Dans le premier niveau du graphe de voisinage, les arêtes de x_1 et x_2 sont connectées par des éléments de classe unique, nous considérons que le classifieur a estimé la classe '−' pour x_1 et la classe '+' pour x_2. Ainsi, les rapports statistiques donnent : $R_1 = 1$ et $R_2 = 1$. Notre contribution est bien placée pour renforcer cette confiance en proposant une vue plus exhaustive de l'espace de données, cela par une construction d'un second niveau du graphe de voisinage. Dans la figure 4, le second niveau exprime un supplément d'information qui peut confirmer ou infirmer la confiance initiale. Dans le second niveau, x_2 a donné un supplément, certains sommets sont connectés avec des données de la classe opposée, contrairement à x_1, tous les sommets connexes sont de classe unique, correspondent aussi à la classe prédite par le classifieur. Cette procédure a permis de renforcer la confiance d'étiquetage, l'exemple de l'échantillon x_1 est considéré plus confiant à être classe '−' par rapport à x_2. Les détails de notre proposition sont résumés dans l'algorithme 4.

Algorithm 4 : Algorithme de *R-COSET*

Entrées :
- D_E : Donnée étiquetée
- D_{NE} : Donnée non étiquetée
- Hypothèse : Algorithme d'apprentissage supervisé
- k : Nombre de données de confiance sélectionné à chaque itération
- $Seuil_{confiance}$

Sorties :
- h : Hypothèse finale

Processus :
- $h' \leftarrow$ Faire apprendre l'hypothèse sur (D_E)
- **Répéter jusqu'à** ce qu'il n'y ai aucun changement dans D_E
  - Utiliser h' pour prédire les classes de D_{NE}
  - **Pour** chaque $x_i \in D_{NE}$
    - Construire le premier niveau du graphe $G_{premier}$ de x_i avec D_E
    - Calculer le rapport R_i
    - **Si** $R_i > Seuil_{confiance}$
      - **Pour** chaque sommet x_j de $G_{premier}$
        - Construire le second niveau du graphe G_{second} de x_j
        - Calculer le rapport R_j
        - Calculer la confiance *MC* par Eq. 3.3
        - Identifier les k données de confiance et les ajouter à D_E
        - $h' \leftarrow$ Réapprentissage de l'hypothèse sur nouvelle D_E

Fin
- $h \leftarrow h'$

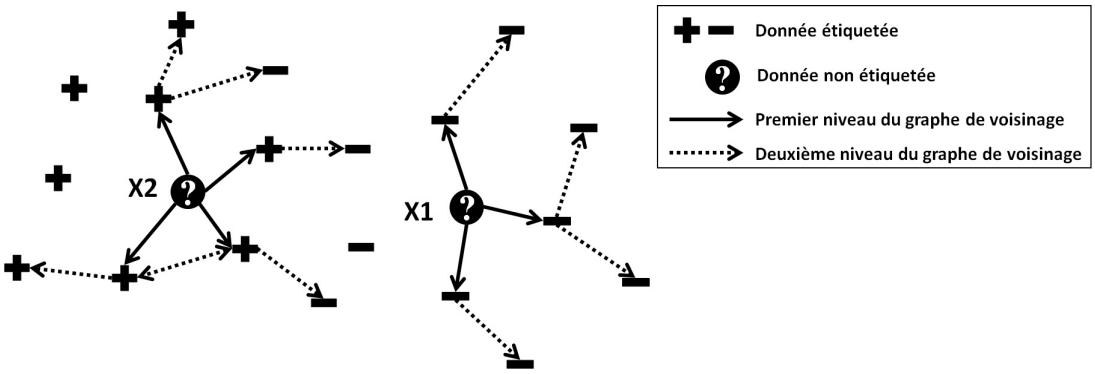


Figure 4 – Construction du graphe de voisinage proposé pour R-COSET.

4 Expérimentations et résultats

Dans cette section, nous réalisons l'étude expérimentale pour vérifier l'efficacité de l'algorithme proposé *R-COSET* sur des problèmes de classification binaire. Treize bases de données médicales et biologiques provenant principalement de référence UCI [61] sont utilisées. Le tableau 7 résume les propriétés des ensembles de données sélectionnées, avec le nombre d'exemples (#instances.) et le nombre d'attributs (#variables).

Bases	**#instances**	**#variables**	**#Classe**
Arcene	900	10000	2
Breast-cancer-wisconsin	699	9	2
Bupa	345	6	2
Colon	62	2000	2
Haberman	306	3	2
Heart	270	13	2
Hepatitis	155	19	2
Leukemia	72	7129	2
Ovarian	54	1536	2
Pancreatic	119	6771	2
Pima	768	8	2
Promoters	106	57	2
Prostate	102	12533	2

Table 7 – Description des bases d'expérimentation.

Les ensembles de données utilisées ont été partagés en cinq partitions en utilisant la procédure de validation croisée, ces partitions sont programmées pour la phase d'apprentissage et de test. Chaque partition d'apprentissage est divisée en deux parties : des données étiquetées et non étiquetées. Afin d'étudier l'impact de la quantité de données étiquetées en apprentissage semi-supervisé, nous simulons différents rapports lors de la division de l'ensemble d'apprentissage. Dans nos expériences, quatre rapports sont utilisés : 20 %, 40 %, 60 % et 80 %. Par exemple, pour les 1000 échantillons dans un ensemble d'apprentissage, lorsque le taux de labélisation est de 20 %, donc 200 échantillons sont placés dans les données étiquetées D_E avec leurs étiquettes alors que les 800 autres échantillons sont considérés comme données non étiquetées D_{NE}.

Dans notre étude expérimentale, nous nous sommes appuyés sur les travaux d'Isaac et al. [10] dans le choix des algorithmes d'auto-apprentissage qui sont les plus pertinent dans la littérature. Nous nous sommes intéressé à améliorer le calcul de confiance dans notre proposition, en se basant sur la théorie de graphe. On compare *R-COSET* à des algorithmes similaires qui utilisent le même principe de graphe de voisinage comme SETRED [6] et SNNRCE [7].

Pour chaque ensemble de données, nous comparons *R-COSET* avec trois autres algorithmes d'apprentissage semi-supervisé : *SETRED* [6], *SNNRCE* [7] et l'auto-apprentissage standard (Basic Self-training). *SETRED* est un algorithme d'auto-apprentissage amélioré employant la technique de filtrage CEWS pour supprimer les données mal classifiées dans l'ensemble d'apprentissage pendant le processus d'auto-apprentissage, et *SNNRCE* est une approche d'auto-apprentissage basée sur la règle du plus proche voisin et de la technique de CEWS. Nous avons utilisé le classifieur de k plus proche voisin (k-NN) comme hypothèse de base dans l'auto-apprentissage standard, *SETRED* et *R-COSET* (avec $k = 5$ et distance euclidienne). Pour une comparaison équitable, le critère d'arrêt utilisé par l'auto-apprentissage standard est similaire à celui utilisé par *SETRED* (40 itérations). Le niveau de confiance utilisé dans *SETRED* est 0.1 et dans *SNNRCE* est 0.5. En outre, le nombre de k utilisé dans *R-COSET* est fixé à 2 et $Seuil_{confiance}$ est fixé à 0.75.

Il existe un certain nombre de critères qui peuvent être utilisés pour comparer la robustesse de chaque algorithme de classification semi-supervisée. Ceux-ci incluent le taux de classification en phase transductive et inductive, en fonction du taux de labélisation. L'apprentissage transductif est la première étape de chaque processus semi-supervisé, il est destiné à prédire la véritable classe des données non étiquetées utilisées pour l'apprentissage. Dans le cadre de l'apprentissage transductif, nous comparons la performance de notre proposition par rapport aux algorithmes de comparaison (Basic Self-training, SETRED et SNNRCE. Contrairement à l'apprentissage transductif, l'apprentissage inductif destiné à vérifier les performances de l'hypothèse finale apprise pendant la phase transductive, cette hypothèse est utilisée pour classer les données de test ainsi nous comparons les résultats inductifs. Pour mieux discuter les résultats de classification en particulier dans le domaine semi-supervisé, il est fortement recommandé d'utiliser des tests non paramétriques et des tests d'addition de bruit. Ceci peut clairement évaluer l'efficacité de l'approche proposée, en comparant statiquement les résultats de chaque technique de classification, et en testant la robustesse contre le bruit rencontré pendant l'apprentissage.

Les tableaux 8 et 9 représentent les résultats de classification et le classement de chaque algorithme d'expérimentation obtenu dans la phase transductive et inductive. Plus précisément, ils montrent les taux de reconnaissances sur les treize ensembles de données utilisées avec 20, 40, 60 et 80% de taux de labélisation. Notez que les améliorations obtenues ont été soulignées en caractères gras.

Commençant par l'analyse transductive, notre proposition surpasse les trois autres algorithmes, sauf que la performance de *R-COSET* a été dégradée dans la base d'Hepatitis avec 20 % du taux de labélisation (table 8). Ceci peut être expliqué par un mauvais apprentissage de l'hypothèse initiale et par l'ajout de données mal classées dans l'ensemble d'apprentissage. Cependant, le *R-COSET* a bénéficié de données non étiquetées puisque les performances sont évidemment améliorées sur l'ensemble de données en comparaison par rapport aux autres algorithmes.

Lorsque le taux de labélisation est à 40 % et 60 % (Tableau 8 et 9), nous observons une nette amélioration en matière de taux de classification pour notre proposition *R-COSET*. Il existe une certaine stabilité de performance lorsqu'on utilise un nombre élevé de données étiquetées (80% de taux de labélisation), par exemple, dans les bases de données Leukemia et Ovarian, notre algorithme a donné des résultats similaires par rapport à *SETRED*, cela peut être expliqué par le fait d'utiliser presque le même principe dans le calcul de confiance.

La comparaison entre l'algorithme *SETRED* et *R-COSET* montre que *R-COSET* a donné de meilleurs résultats de classification, ce qui confirme l'efficacité de sélection des données de confiance lors de la phase d'auto-apprentissage.

		Phase transductive			
Bases	**Taux de labélisation**	**R-COSET**	**Basic Self-training**	**SETRED**	**SNNRCE**
Arcene	20%	**78,51(1)**	54,67(4)	76,39(2)	66,01(3)
	40%	**87,84(1)**	57,37(4)	87,8(2)	72,46(3)
	60%	**93,02(1)**	68,95(4)	91,57(2)	77,64(3)
	80%	**93,83(1)**	73,55(4)	90,51(2)	73,91(3)
Breast-cancer	20%	**97,8(1)**	52,58(4)	96,35(2)	92,58(3)
	40%	**96,82(1)**	83,14(4)	96,76(2)	93,66(3)
	60%	**97,46(1)**	88,87(4)	96,72(2)	93,74(3)
	80%	**99,16(1)**	92,18(4)	97,98(2)	95,61(3)
Bupa	20%	**66,81(1)**	41,12(4)	63,78(2)	62,16(3)
	40%	**67,95(1)**	46,34(4)	65,43(2)	58,6(3)
	60%	**66,49(1)**	55,02(4)	64,23(2)	58,03(3)
	80%	**76,96(1)**	62,57(3)	75,9(2)	61,7(4)
Colon	20%	**70,69(1)**	68,96(2)	68,13(3)	63,44(4)
	40%	**77,33(1)**	72,59(2)	65,92(3)	63,80(4)
	60%	**92,5(1)**	73,24(3)	90,64(2)	67,69(4)
	80%	**90(1)**	75,33(3)	83,33(2)	72(4)
Haberman	20%	**74,85(1)**	72,90(3)	73,73(2)	67,60(4)
	40%	**81,65(1)**	72,67(3)	76,27(2)	67,45(4)
	60%	**78,85(1)**	73,11(3)	76,17(2)	70(4)
	60%	**79,46(1)**	78,36(3)	78,53(2)	68(4)
Heart	20%	**69,87(1)**	56,42(4)	67,97(2)	66,38(3)
	40%	**64,47(1)**	57,4(3)	64,2(2)	55,55(4)
	60%	**73,8(1)**	64,35(3)	69,01(2)	60(4)
	80%	**73,51(1)**	66,5(3)	71,9(2)	61,66(4)
Hepatitis	20%	76,35(3)	**79,74(1)**	78,3(2)	66,83(4)
	40%	**82,26(1)**	79,52(3)	81,91(2)	69,49(4)
	60%	**80,49(1)**	79,26(3)	79,79(2)	73,84(4)
	80%	**86,16(1)**	81,37(3)	83,72(2)	72,94(4)
Leukemia	20%	**88,55(1)**	68,06(4)	85,27(2)	74,79(3)
	40%	**93,37(1)**	70,11(4)	89,64(2)	79,72(3)
	60%	**100(1)**	76,27(4)	98,33(2)	78,58(3)
	80%	100(1,5)	80,33(3)	100(1,5)	78(4)
Ovarian	20%	**92,33(1)**	45,49(4)	83,05(2)	79,25(3)
	40%	**92,64(1)**	54,31(4)	84,73(2)	80(3)
	60%	**93(1)**	66,71(4)	82,26(3)	83,07(2)
	80%	85(1,5)	76,66(3)	85(1,5)	56(4)
Pancreatic	20%	**54,58(1)**	44,1(4)	53,97(3)	54,34(2)
	40%	**56,07(1)**	45,91(4)	53,35(3)	55,58(2)
	60%	**61,96(1)**	46,66(4)	56,88(2)	54,09(3)
	80%	**67,83(1)**	44,54(4)	58,2(2)	57(3)
Pima	20%	**76,56(1)**	65,95(4)	74,62(2)	69,4(3)
	40%	**77,98(1)**	69,82(3)	74,05(2)	68,28(4)
	60%	**80,38(1)**	72,44(3)	79,17(2)	68,51(4)
	80%	**81,44(1)**	74,39(3)	79,69(2)	66,6(4)
Promoters	20%	**83,14(1)**	50,57(4)	72,79(2)	71,85(3)
	40%	**95,45(1)**	52,67(4)	87,93(2)	82(3)
	60%	**83,33(1)**	61,45(4)	83(2)	74,61(3)
	80%	**90(1)**	79,52(3)	88,33(2)	56,66(4)
Prostate	20%	**78,13(1)**	52,1(4)	75,17(2)	65,76(3)
	40%	**81,84(1)**	61,09(4)	78,99(2)	68,94(3)
	60%	**80(1)**	62,61(4)	79,49(2)	66,66(3)
	80%	**89,16(1)**	68,02(3)	86,3(2)	66(4)
Average Rank	20%	**1,15**	3,53	2,15	3,15
	40%	**1**	3,53	2,15	3,30
	60%	**1**	3,61	2,07	3,30
	80%	**1,07**	3,23	1,92	3,76

Table 8 – Taux de classification et Ranking des résultats transductifs.

Par conséquent, la mesure de confiance peut affecter des faux labels à la base étiquetée, ceci est dû au fait que le classifieur peut affecter des classes erronées à certains exemples non étiquetés, à cet effet la capacité de généralisation de l'hypothèse finale sera affectée par l'accumulation de ce bruit dans chaque itération du processus d'apprentissage. Dans ce cas, une faible précision est obtenue à la fin du processus.

L'introduction du deuxième niveau du graphe de voisinage dans le processus d'auto-apprentissage améliore l'exactitude de la classification. Nous remarquons donc que l'algorithme *R-COSET* donne des bons résultats de classification en comparaison avec d'autres types d'auto-apprentissage (test transductif et test inductif).

Phase inductive					
Bases	**Taux de labélisation**	**R-COSET**	**basic Self-training**	**SETRED**	**SNNRCE**
Arcene	20%	**62,35(1)**	55,29(4)	61,47(2)	60,58(3)
	40%	**74,11(1)**	67,05(3)	71,76(2)	59,41(4)
	60%	**74,7(1)**	72,82(3)	72,94(2)	62,64(4)
	80%	**80,88(1)**	79,11(3)	80(2)	64,11(4)
Breast-cancer	20%	**96,15(1)**	71,45(4)	95,98(2)	94,61(3)
	40%	**96,75(1)**	92,73(4)	96,58(2)	95,81(3)
	60%	95,99(2)	95,29(4)	95,89(3)	**96,06(1)**
	80%	**97,09(1)**	96,55(4)	97(2)	96,75(3)
Bupa	20%	**61,03(1)**	44,48(4)	58,27(2)	54,65(3)
	40%	**64,82(1)**	52,24(4)	63,96(2)	58,79(3)
	60%	**69,82(1)**	65,68(3)	67,93(2)	62,41(4)
	80%	**66,03(1)**	65,31(3)	65,51(2)	58,44(4)
Colon	20%	**65,45(1)**	62,72(2)	60,9(3)	49,09(4)
	40%	**60(1)**	59,09(2)	55,45(3)	48,18(4)
	60%	**66,36(1)**	65,45(3)	65,6(2)	63,63(4)
	80%	57,27(1,5)	57(3)	57,27(1,5)	52,72(4)
Haberman	20%	**74,70(1)**	73,33(3)	74,31(2)	70,39(4)
	40%	**74,11(1)**	73,33(4)	73,72(2)	73,7(3)
	60%	**74,7(1)**	73,92(4)	74,21(2)	74,11(3)
	80%	**75,49(1)**	74,9(3)	75,29(2)	72,74(4)
Heart	20%	**57,77(1)**	56,22(4)	56,44(3)	56,66(2)
	40%	**62,66(1)**	59,33(4)	61,11(2)	60,66(3)
	60%	**66,88(1)**	64,44(3)	65,55(2)	60,22(4)
	80%	67,11(1,5)	66,22(3)	67,11(1,5)	64,22(4)
Hepatitis	20%	**79,24(1)**	76,22(2)	74,71(3)	70,94(4)
	40%	**78,11(1)**	76,22(2)	72,45(3)	70,18(4)
	60%	72,07(3)	**78,11(1)**	71,32(4)	75,47(2)
	80%	73,58(3)	74,33(1,5)	74,33(1,5)	73,2(4)
Leukemia	20%	75,38(1,5)	67,69(4)	75,38(1,5)	70(3)
	40%	**88,46(1)**	76,15(3)	83,07(2)	69,23(4)
	60%	**84,61(1)**	81,53(3)	82,30(2)	64,61(4)
	80%	**93,07(1)**	91,53(3)	92,3(2)	59,23(4)
Ovarian	20%	**76,66(1)**	45,55(4)	71,11(3)	72,22(2)
	40%	**87,77(1)**	63,33(4)	77,77(2)	65,55(3)
	60%	**85,55(1)**	74,44(3)	83,33(2)	73,33(4)
	80%	81,11(1,5)	80(3)	81,11(1,5)	48,88(4)
Pancreatic	20%	**53,87(1)**	45,48(4)	53,54(2)	50(3)
	40%	**58,06(1)**	49,03(4)	57,41(2)	55,16(3)
	60%	**52,25(1)**	51,29(3)	50,32(4)	51,61(2)
	80%	**55,48(1)**	54,63(3)	54,83(2)	54,51(4)
Pima	20%	**69,80(1)**	65,99(4)	69,64(2)	66,38(3)
	40%	**72,52(1)**	68,79(3)	70,03(2)	68,48(4)
	60%	**70,89(1)**	70,19(3)	70,66(2)	68,09(4)
	80%	72,16(2)	**72,29(1)**	72,06(3)	65,29(4)
Promoters	20%	**71,66(1)**	50(4)	65(3)	66,11(2)
	40%	**74,44(1)**	56,66(4)	68,88(2)	65(3)
	60%	**81,11(1)**	71,11(3)	80,55(2)	67,77(4)
	80%	**78,33(1)**	76,11(2,5)	76,11(2,5)	50(4)
Prostate	20%	**74,85(1)**	51,42(4)	73,71(2)	64(3)
	40%	**73,14(1)**	63,42(3)	71,42(2)	61,14(4)
	60%	**76,57(1)**	72,57(3)	76(2)	64(4)
	80%	**81,71(1)**	80,57(3)	81,14(2)	54,28(4)
Average Rank	20%	**1,03**	3,61	2,34	3
	40%	**1**	3,38	2,15	3,46
	60%	**1,23**	3	2,38	3,38
	80%	**1,34**	2,76	1,96	3,92

Table 9 – Taux de classification et Ranking des résultats inductifs.

Dans le domaine de l'apprentissage automatique, il est fortement recommandé d'utiliser l'analyse non-paramétrique pour comparer et analyser statistiquement les résultats de classification. L'objectif principal de ce type d'analyse est d'identifier les différences de performance les plus pertinentes trouvées entre les outils expérimentaux [70]. Donc, il est préférable d'utiliser les tests non-paramétriques comme il est démontré dans [71]. Nous nous concentrerons sur l'utilisation du test de Aligned-Ranks (FAR) [72], comme un outil de contraste du comportement de chaque proposition. Son application nous permettra de mettre en évidence l'existence de différences significatives entre les performances de chaque méthode de comparaison.

Taux de labélisation	Algorithme	FAR	i	$z = (R_0 - R_i)/SE$	unadjusted p
Phase transductive					
20%	**R-COSET**	**13.0769**	-	-	-
	SETRED	18.7692	1	0.957626	0.338251
	SNNRCE	31.3846	2	3.079933	0.00207
	Basic Self-training	42.7692	3	4.995186	0.000001
40%	**R-COSET**	**10.3077**	-	-	-
	SETRED	18.7692	1	1.423499	0.154592
	SNNRCE	35.3077	2	4.205791	0.000026
	Basic Self-training	41.6154	3	5.266945	0
60%	**R-COSET**	**10.6154**	-	-	-
	SETRED	16.6154	1	1.00939	0.312788
	SNNRCE	37	2	4.438727	0.000009
	Basic Self-training	41.7692	3	5.241063	0
80%	**R-COSET**	**10.3077**	-	-	-
	SETRED	16.8462	1	1.099976	0.271343
	Basic Self-training	36.5385	2	4.412845	0.00001
	SNNRCE	42.3077	3	5.383413	0
Phase inductive					
20%	**R-COSET**	**13.8846**	-	-	-
	SETRED	19.9615	1	1.022331	0.306624
	SNNRCE	32.3846	2	3.112285	0.001856
	Basic Self-training	39.7692	3	4.354611	0.000013
40%	**R-COSET**	**11**	-	-	-
	SETRED	19.8846	1	1.494673	0.135
	Basic Self-training	36.8462	2	4.348141	0.000014
	SNNRCE	38.2692	3	4.587548	0.000004
60%	**R-COSET**	**15.3462**	-	-	-
	SETRED	20.9615	1	0.944685	0.34482
	Basic Self-training	29.2692	2	2.342302	0.019165
	SNNRCE	40.4231	3	4.218732	0.000025
80%	**R-COSET**	**18.6538**	-	-	-
	SETRED	19.8077	1	0.194113	0.846087
	Basic Self-training	21.6538	2	0.504695	0.613773
	SNNRCE	45.8846	3	4.581077	0.000005

Table 10 – Analyse non-paramétrique des résultats transductifs et inductifs

Après avoir calculé le test Friedman Aligned-Ranks (FAR) (tableau 10), nous pouvons constater l'analyse suivante : Avec 20 % du taux de labélisation, le paramètre FAR prouve que notre proposition est la plus performante en matière d'apprentissage transductif et inductif. *R-COSET* est capable de surpasser significativement les techniques de comparaison. L'algorithme *SETRED* est classé en deuxième position, et il est mis en évidence comme un bon processus d'auto-apprentissage.

Pour un taux de labélisation égal à 40 %, on observe une nette amélioration en termes de taux de classification pour toutes les méthodes. Encore une fois, notre proposition obtient le meilleur classement, et l'algorithme *SETRED* est souligné comme deuxième méthode performante.

R-COSET est l'algorithme le plus robuste en matière de capacités à générer une l'hypothèse efficace, cette proposition a permis de produire des différences significatives en comparaison avec *SETRED*, auto-apprentissage standard et *SNNRCE*. Cela est justifié par la procédure Finner et le paramètre FAR (tableau10).

L'effet qui peut dégrader la performance d'un classifieur semi-supervisé est le bruit qui peut être ajouté en plein phase d'auto-apprentissage. De ce fait, nous étudions la robustesse de notre proposition contre l'effet du bruit, nous générons un bruit avec un degré de 5 % selon le principe du système par paires [73], étant donné, un problème binaire de classe (x, y) et un degré de bruit «p», ce degré signifie qu'une instance avec son étiquette «x» a une chance d'être mal classée comme «y», le degré de bruit indique le pourcen-

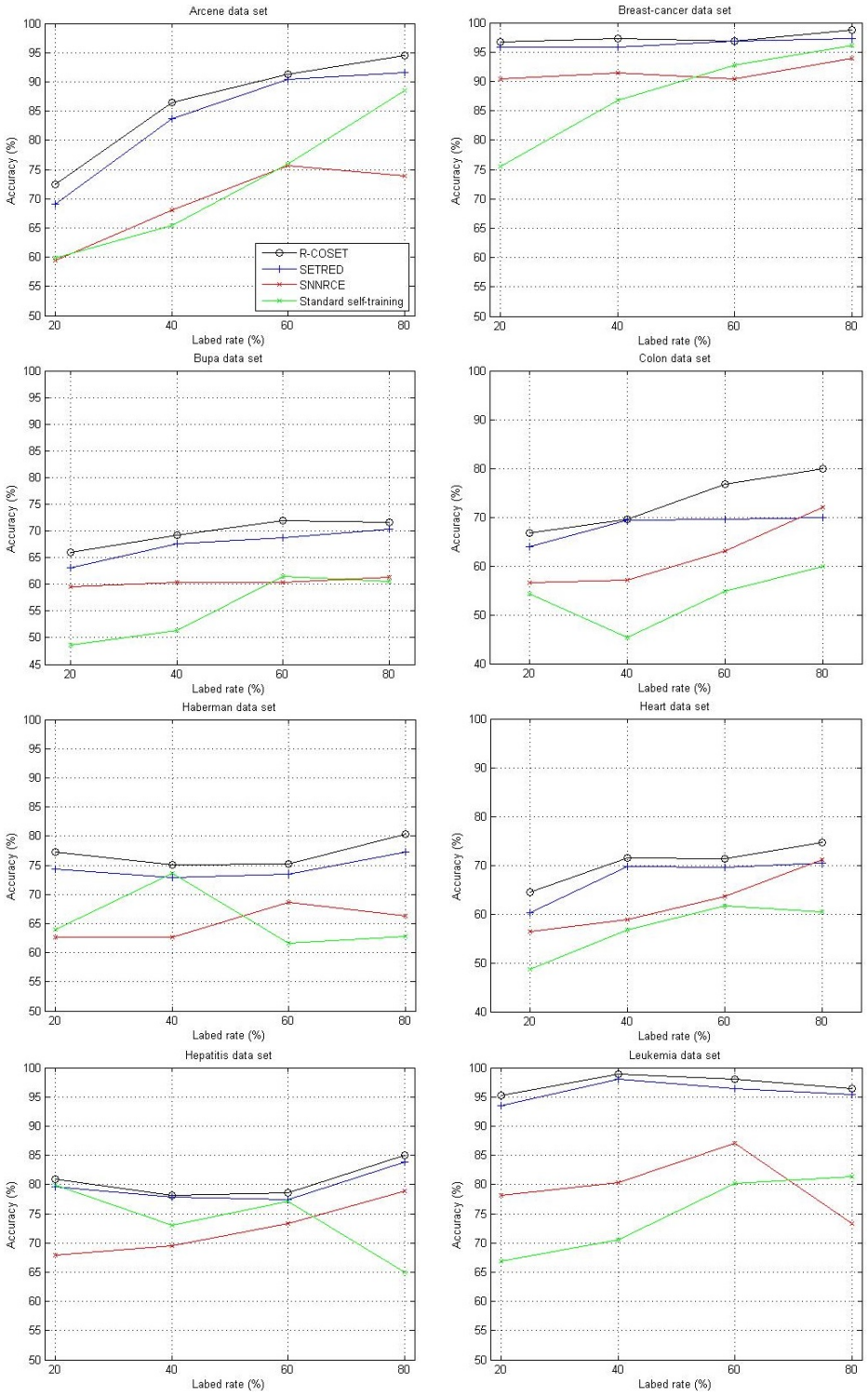


Figure 5 – Résultats de classification considérant l'effet du bruit avec un degré de 5%.

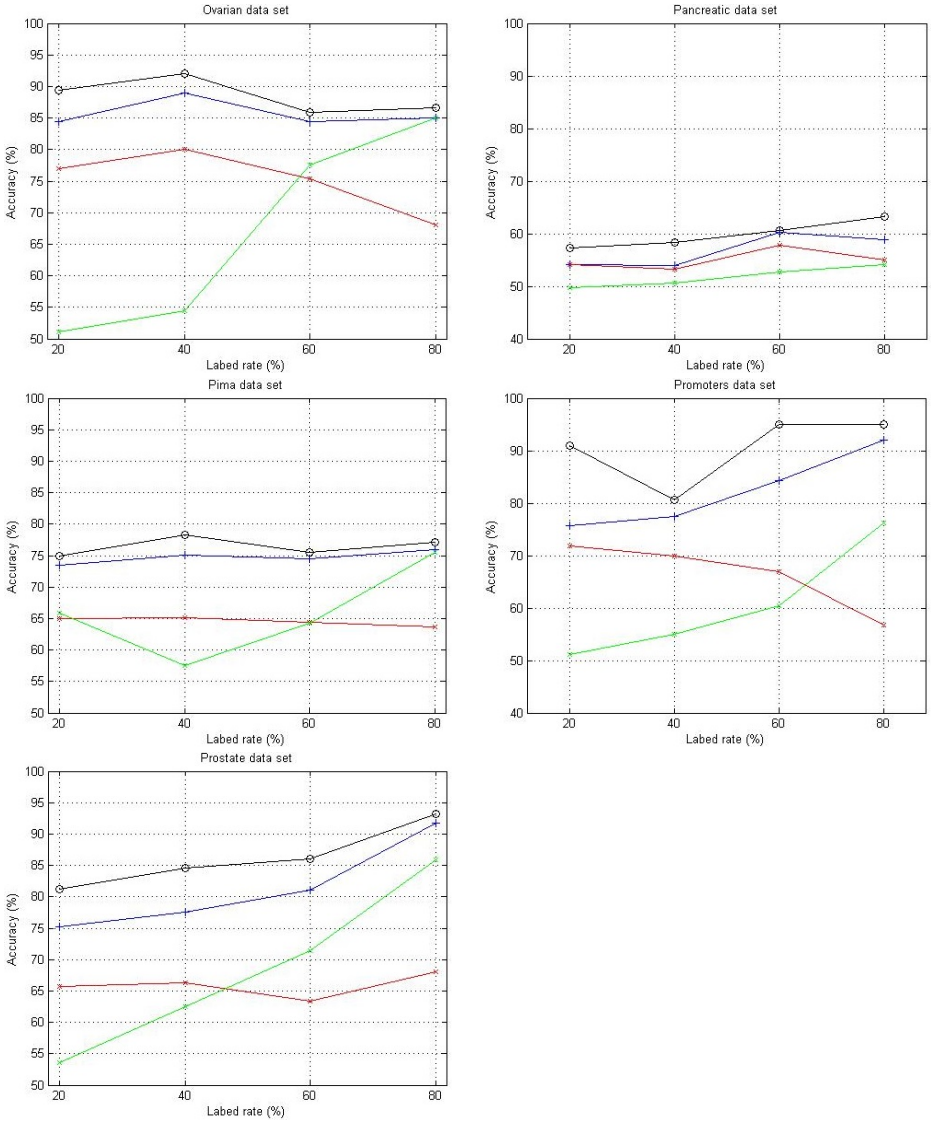


Figure 6 – Résultats de classification considérant l'effet du bruit avec un degré de 5%. (Suite)

tage de données inexactement étiquetées dans la base initiale des données étiquetées.

Nous générons des données bruitées pour tous les ensembles de données expérimentales, et les résultats de classification sont représentés dans les figures 5 et 6. Les résultats obtenus prouvent la robustesse de *R-COSET* contre le bruit généré en phase d'auto-apprentissage. Ainsi, la procédure de confiance proposée affirme l'efficacité de générer une hypothèse de classification robuste. L'auto-apprentissage standard et le *SNNRCE* sont plus affectés par l'effet de bruit particulièrement dans les cas de 20 % et 40 % de taux de labélisation, cela est observé sur les jeux de données Arcene, Leukemia et Prostate. Il est clairement démontré que notre proposition est robuste, et ainsi elle peut améliorer considérablement les résultats de la classification, en particulier avec l'utilisation d'une quantité minimale de données étiquetées.

5 Conclusion

Le succès des approches d'auto-apprentissage réside dans la manière du calcul de confiance et ainsi dans le choix des prédictions avec une confiance élevée, celle-ci joue un rôle fondamental dans l'enrichissement de la base étiquetée d'apprentissage. Dans ce chapitre, l'algorithme *R-COSET* est proposé, il a la faculté de tirer profit des échantillons non étiquetés pour améliorer les performances de l'hypothèse formée à partir des échantillons étiquetés. *R-COSET* est un algorithme d'auto-apprentissage dans lequel les données non étiquetées sont ajoutées à la base étiquetée de manière itérative.

Le principe de chaque itération de *R-COSET* est d'utiliser le graphe de voisinage relatif sur deux niveaux de construction et les informations statistiques CEWS pour décrire la similarité entre la contiguïté des sommets, cette procédure a pu renforcer la confiance de prédiction et enrichir la base d'apprentissage par des échantillons de confiance.

Les résultats expérimentaux sur treize bases de données de référence démontrent que la méthode proposée surpasse les algorithmes de comparaison (auto-apprentissage standard, *SNNRCE* et *SETRED*). Les tests réalisés prouvent la robustesse de notre approche contre les effets du bruit, cet aspect devient plus important lorsque l'ensemble d'apprentissage devient plus petit, et nos résultats ont montré la résistance de notre proposition dans ce cas.

Troisième partie

III

Segmentation supervisée et semi-supervisée des cellules sanguines par classification super-pixellique

Introduction

L'imagerie médicale joue un rôle essentiel dans le diagnostic médical, image à l'appui, pourrait donc refléter des informations cliniques aux médecins traitants. Différentes modalités d'imagerie médicale ont été développées de telle sorte à mettre en image les différentes régions ou les différents organes de l'organisme, permettant une visualisation avec une très bonne résolution spatiale, comme le cœur, les reins, le foie et l'histologie sanguine.

L'expertise des frottis du sang représente la pierre angulaire du diagnostic hématologique. Il est évident que cet examen est un indicateur important dans la détection de certaines anomalies sanguines. La morphologie sanguine se compose de trois éléments : les cellules comme les globules rouges (*érythrocytes*) et les globules blancs (*leucocytes*) ainsi que les plaquettes sanguines (non considérés comme des cellules). L'expression de la forme et le nombre de globules blancs *(WBC : White Blood Cell)* comportent de nombreux indices quantitatifs et informatifs. Par exemple, l'augmentation ou la diminution de leucocytes est très critique et peut susciter une attention médicale.

Les dernières avancées en matière de caméra numérique combinée à des logiciels informatiques ont permis d'acquérir et de stocker un nombre important d'images microscopiques. Toutefois, le traitement manuel de cette masse de données par le médecin hématologue est souvent difficile, coûteux et fastidieux.

L'un des domaines mis en évidence d'aide au diagnostic en hématologie est le problème de la détermination et l'analyse automatique des globules blancs (*WBC*). Grâce à la technologie d'analyse d'images médicales qui a connu une croissance rapide au cours des dernières années, des améliorations considérables ont été apportées aux examens cliniques qui permettent d'apporter des facilités dans le diagnostic médical.

Dans le cadre des techniques d'analyse d'images médicales, la segmentation des globules blancs est un problème clés auquel nous ferons référence dans cette partie de la thèse. La segmentation des images microscopiques utilise l'information issue de l'image (couleur, niveau de gris et spatial) pour délimiter les différentes structures anatomiques, notamment, les globules blancs (*WBC*) qui se compose de : *noyau* et *cytoplasme*.

Plusieurs travaux de recherche ont été menés sur l'utilisation du traitement des images couleur dans la segmentation et la classification de certaines maladies et tumeurs. Irshad et al. [74] présentent un article qui regroupe les techniques les plus intéressantes de segmentation et de classification de noyaux en imagerie microscopique, notamment, les images de type *Hematoxylin-Eosin (H&E)* et *ImmunoHistoChemical (IHC)*. Cette étude met en évidence les principales tendances de segmentation, de caractérisation et de classification de noyaux en imagerie histopathologie, cela en se basant sur une vue exhaustive des différentes techniques existantes dans la littérature.

Suite aux travaux de [74] et [75], la tendance actuelle est vers l'application de la classification pixellique basée sur l'intelligence artificielle. Cette dernière présente un potentiel important dans la segmentation des images couleurs, notamment, l'intégration des techniques de superpixel dans le processus de segmentation. Zhao et al. [76] ont recommandé l'utilisation de superpixel, ce dernier est une technique de clustering qui permet de subdiviser l'image en k cluster homogène permettant d'accélérer et d'améliorer la qualité de segmentation, la figure 7 représente le processus de segmentation par classification super-pixellique.

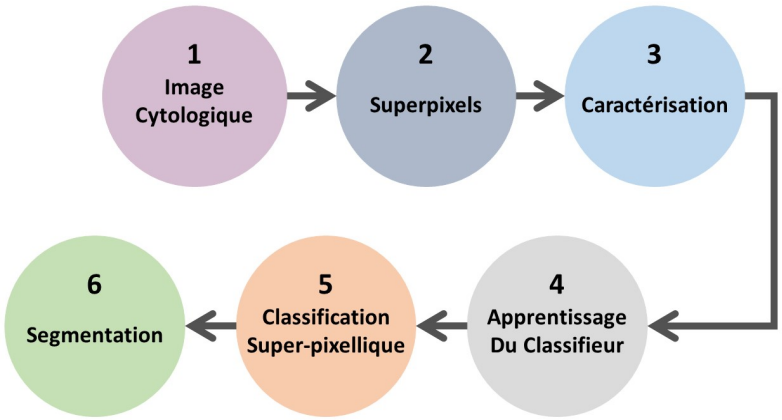


Figure 7 – Processus de segmentation par classification super-pixellique.

En se basant sur la figure 7, le succès de cette procédure de segmentation réside dans deux facteurs clefs, qui sont : le premier facteur, une bonne opération d'extraction des caractéristiques adaptées au problème de segmentation (étape 3), et le deuxième facteur, le choix du classifieur (étape 4) qui permet de générer une hypothèse robuste afin de réaliser une segmentation fiable. Dans ce contexte, Cernadas et al. [11] et González-Rufino et al. [12] présentent une évaluation exhaustive de différentes techniques de caractérisation se basant sur la couleur et la texture. Cernadas et al. [11] ont étudié l'influence de la normalisation de couleur en segmentation. González-Rufino et al. [12] ont présenté une liste exhaustive de plusieurs modes de caractérisation pour une meilleure reconnaissance du noyau en imagerie histologique.

L'intégration de superpixel dans le processus de segmentation a fait découvrir un nouvel axe de recherche dans la segmentation des images, et envisager la porte à de nouvelles perspectives. Dans cette deuxième partie de thèse, nous cherchons à répondre à des perspectives sérieuses qui peuvent influencer positivement ou négativement la qualité de segmentation. Ces perspectives peuvent être résumées dans les questions suivantes :

– La normalisation de couleur a-t-elle une influence sur la segmentation ?
– Quel est le meilleur espace couleur ?
– Quel mode de caractérisation doit-on utiliser ?
– Quel classifieur doit-on choisir ?

Suivant ces perspectives, nous proposons alors dans cette deuxième partie de thèse une série d'expérimentation partagée en deux volets. Dans le premier volet, en mettant en avant les spécificités de trois traitements qui sont : la normalisation de couleur, la diversité des composantes couleurs et la caractérisation super-pixellique. Les traitements cités permettent une étude exhaustive sur la qualité des caractéristiques mesurées. En effet, cette qualité est indispensable pour aborder la classification super-pixellique, l'efficacité de la classification repose essentiellement sur la fiabilité de la caractérisation super-pixellique (Chapitre 1).

Le deuxième volet concerne la réponse à la question *«Quel classifieur doit-on choisir ?»*, donc, nous exploitons les meilleures performances de caractérisation mesurées dans le premier volet pour détailler le concept de la classification super-pixellique. Les modes d'apprentissage supervisé et semi-supervisé seront étudiés sous forme d'exposition comparative, en discutant les performances de la classification super-pixellique (Chapitre 2). Cette exposition à pour but de mettre en évidence les qualités d'apprentissage semi-supervisé dans la segmentation des images médicales.

CHAPITRE 1

Caractérisation super-pixellique

1 Objectifs

L'avancée des systèmes d'acquisition dans le domaine d'imagerie microscopique a permis une facilité dans la numérisation et le stockage des données image. En raison de la difficulté du traitement manuel par les médecins, on fait appel à des outils informatiques en analyse d'image pour atteindre des méthodes de traitement d'images automatiques. Les défis soulevés dans les systèmes d'aide au diagnostic se concentrent sur le développement des outils précis, performants et rapides en exécution, ainsi la capacité de manipuler une grande masse de données lors des exercices expérimentaux [77].

La réussite des systèmes d'analyse d'image réside dans l'efficacité de l'opération de segmentation, qui est l'action de subdiviser l'image en régions homogènes. Chaque subdivision a généralement des propriétés similaires telles qu'une valeur d'intensité ou une texture similaire. Toutefois, les autres analyses telles que l'extraction des caractéristiques et la classification dépendent essentiellement de la qualité de segmentation.

De ce fait, la segmentation des globules blancs présente un grand défi dans la conception des systèmes d'aide au diagnostic, cela peut être fait par plusieurs techniques de segmentation. Dans notre travail, nous relevons le défi d'effectuer la segmentation et la reconnaissance des globules blancs (*noyau* et *cytoplasme*) dans la même phase de traitement. Pour cela, il serait intéressant d'intégrer des techniques issues de l'intelligence artificielle dans la segmentation, nous allons procéder à la classification super-pixelique en faisant appel aux techniques de super-pixel dans le processus de segmentation.

Dans ce chapitre, une étude comparative de plusieurs techniques est proposée, le principe repose sur l'application de super-pixel pour subdiviser l'image, par la suite nous exploitons l'information couleur pour caractériser chaque super-pixel, et nous classifions ce dernier pour la reconstruction de l'image segmentée.

Ce chapitre est organisé de la manière suivante : dans la section 2, nous présentons les techniques de segmentation existantes dans la littérature en expliquant le principe de chacune. Par la suite, nous détaillons le principe de la classification super-pixellique dans la section 3. Dans la section 4, nous décrivons les outils d'extraction des caractéristiques. Un plan d'expérimentations est exposé dans la section 5, et les résultats de chaque étape du plan expérimental seront présentés dans la section 6. Nous terminons par une synthèse générale qui met en évidence les principales propriétés de cette étude ainsi que ses points forts.

2 Techniques de segmentation

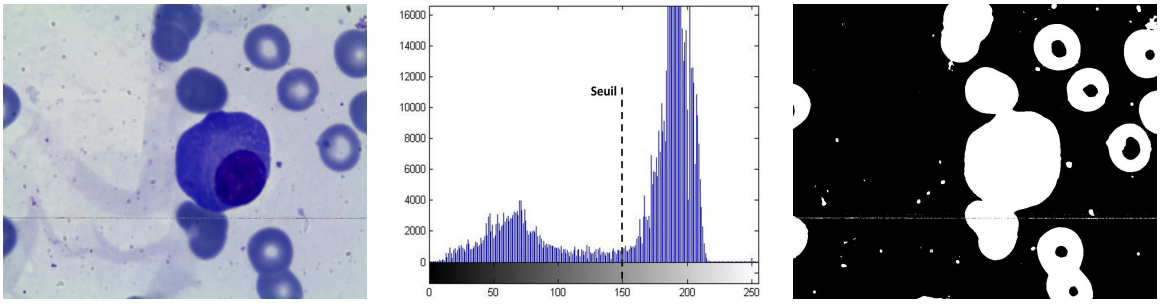
Les techniques de segmentation existantes dans la littérature peuvent être regroupées en cinq familles [78], qui sont :

1. Seuillage,
2. Segmentation à base de région,
3. Segmentation à base de contour,
4. Segmentation à base de graphe,
5. Segmentation basée sur la classification.

2.1 Seuillage

Le seuillage [79] est l'une des techniques basiques de segmentation. Il peut être classé en deux catégories : le seuillage global et le seuillage local (adaptatif), sur la base des critères de sélection des seuils. Une région est caractérisée par sa distribution de niveaux de gris. A chaque pic de l'histogramme est associée une région.

La méthode de seuillage global permet de trouver un seuil S via l'histogramme de l'image, de telles façons que les pixels dont la valeur d'intensité est supérieure ou égale à S sont classés comme région "1" et le reste des pixels comme région "2". Ce genre d'algorithme présente une faible complexité de calcul et une simplicité d'implémentation. Cependant, le seuillage global suppose que l'intensité de l'image a une distribution bi-modale. Cette hypothèse ne tient pas pour la plupart des images, comme a été bien démontré dans la figure 8, un simple seuillage ne peut pas séparer les cellules sanguines.



(a) **(b)** **(c)**

Figure 8 – Exemple d'un seuillage global, (a) image original, (b) histogramme et (c) résultat de seuillage avec un seuil $S = 150$.

Une autre limite du seuillage, qui est très sensible aux effets de bruit, et qui ne prend pas en considération l'information spatiale. Parfois, l'image segmentée peut ne pas correspondre aux objets de l'image d'origine.

2.2 Segmentation à base de région

Cette famille de segmentation fonctionne de manière itérative en regroupant les voisinages de pixels qui portent une similarité en fonction de l'intensité, l'information spatiale . . . etc. Trois algorithmes de segmentation appartient à cette famille : division et fusion (split and merge) [80], croissance de région (region growing) [81] et ligne de partage des eaux (watershed) [82].

Division et fusion

Introduit en 1980 par Horowitz et Pavlidis [80], cet algorithme se compose de deux phases de traitement. La première, division (split), a pour objectif de subdiviser l'image en bloc homogène. Initialement, l'image doit être de taille 2xn ainsi considérée comme un bloc. Suivant un processus itératif, ce bloc est divisé en quatre quadrants, si un critère d'homogénéité n'est pas vérifié, le processus de division sera répété récursivement jusqu'à ce que chaque quadrant ne contienne que des pixels homogènes, comme le montre la Fig 9. Une fois l'image est subdivisée en petit quadrant homogène, le processus de fusion permet de fusionner chaque quadrant adjacent qui vérifie un critère d'homogénéité. L'algorithme compare alors tous les quadrants avec leurs voisins et fusionne les quadrants qui sont homogènes entre eux selon certains critères, comme le montre la figure 9.

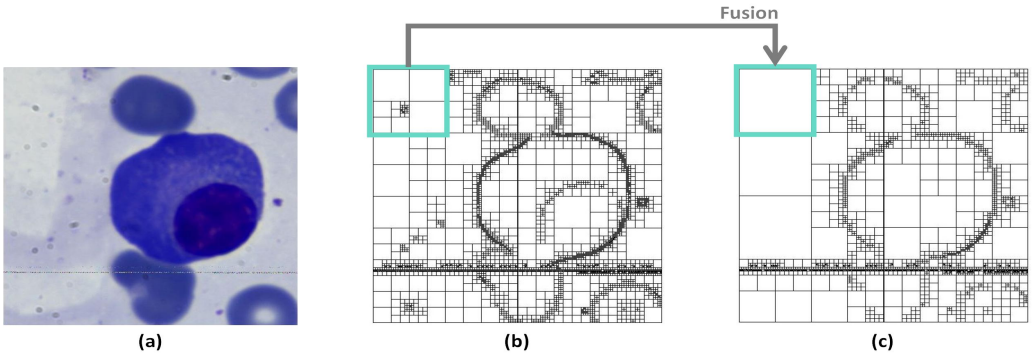


Figure 9 – Principe de division et fusion, (a) image originale, (b) division et (c) fusion.

Croissance de région

Contrairement à la mèthode division et fusion, les approches de croissance de région ont une autre philosophe de segmentation. Elles partent de quelques pixels de l'image essayant par la suite de regrouper les pixels voisins en appliquant un critère de regroupement. Des germes sont sélectionnés manuellement et correspondent aux points de départ de l'algorithme. Ainsi on commence à associer aux germes les seuls pixels qui sont en très bon accord avec les points de départ. La figure 10 illustre le principe de segmentation par croissance de région. Le choix du critère d'addition est crucial pour le succès de cet algorithme. Un critère d'homogénéité proposé par Adams et al. [83] est la différence entre la valeur d'intensité du pixel candidat et la valeur d'intensité moyenne de la région construite.

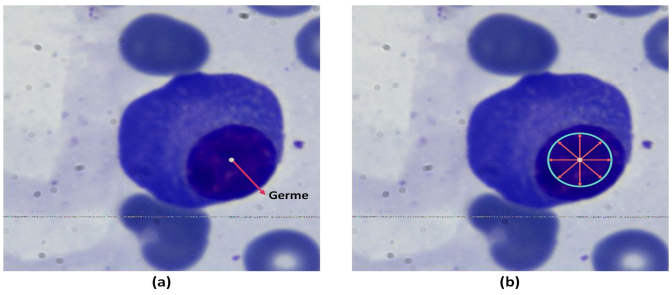
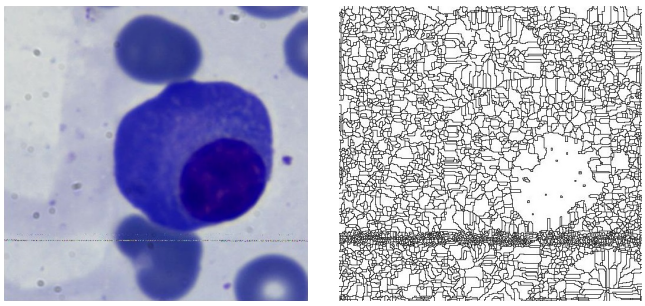


Figure 10 – Principe de croissance de région, (a) initiation de germe et (b) croissance de région.

Ligne de partage des eaux

L'algorithme ligne de partage des eaux [82] est une autre approche de segmentation d'image à base de région qui a connu une large exploitation dans la segmentation des

images médicales. Dans cet algorithme, l'image est considérée comme une carte topographique, dont on simule des inondations à partir de minima locaux. La ligne de partage des eaux est représentée par les points où deux lacs disjoints se rejoignent au cours de l'inondation. Cependant, la limite majeure de cette technique réside dans l'effet de sur-segmentation. Un histogramme d'une image présente plusieurs minima locaux et aussi plusieurs altitudes, donc le résultat si on applique à la lettre l'algorithme ligne de partage des eaux sera une image qui contient de nombreuses petites régions, cet effet est considéré comme la sur-segmentation (Fig. 11). Plusieurs solutions ont été apportées pour éviter la sur-segmentation dans le but de négliger les minima et les altitudes moins intéressants, par exemple, Najman et al [84] ont proposé d'utiliser des opérations morphologiques pour réduire la sur-segmentation.



(a) **(b)**

Figure 11 – Exemple de ligne de partage des eaux, (a) image original et (b) sur-segmentation.

2.3 Segmentation à base de contour

Contrairement aux approches régions, la segmentation par approche contour cherche à trouver les primitives d'une région. Les contours sont des courbes de transition qui séparent les régions d'intérêt. Généralement les contours se caractérisent par des valeurs de gradient élevées.

Détection de contour

La notion de contour étant reliée à celle de variation, il est évident qu'une telle définition nous amène tout naturellement vers une évaluation de la variation en chaque pixel. Une variation existera si le gradient est localement maximum ou si la dérivée seconde présente un passage par zéro. Les principaux algorithmes connus (Sobel, Prewitt, Kirsh, Canny, Dériche, ...) se focalisent sur cet aspect du contour.

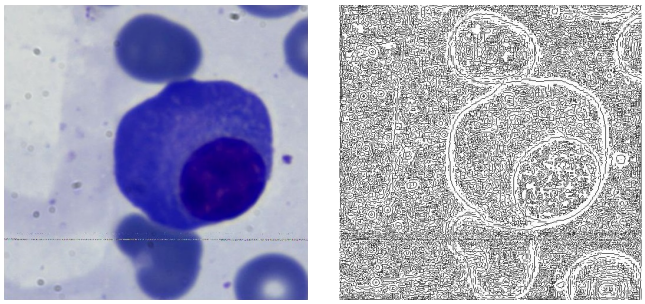
Le gradient d'une image se calcule comme suit :

$$\nabla I_x = \frac{\partial I_f(x,y)}{\partial x} \quad et \quad \nabla I_y = \frac{\partial I_f(x,y)}{\partial y} \tag{1.1}$$

Ces dérivées sont calculées par convolution de l'image avec un masque de différences. La figure 12 illustre un exemple de détection de contour par l'opérateur de Sobel.

Contours actifs

Les contours actifs apparaissent comme un outil intéressant dans la famille des détecteurs de contours, son principe est de faire évoluer un contour initial vers le contour de la région. De plus, il est possible d'imposer des contraintes de régularisation sur le contour. Plusieurs aspects sont utilisés pour faire évoluer le contour actif tel que la minimisation ou maximisation de l'énergie comme utilisé dans Level Set [85].



(a) **(b)**

Figure 12 – Exemple de détection de contour, (a) image original et (b) détection de contour par l'opérateur de Sobel.

2.4 Segmentation à base de graphe

Wu et Leahy [86] et Zahn [86] sont les premiers intégrateurs de la théorie des graphes dans le traitement d'image. Cela a obtenu un grand intérêt dans les applications de segmentation. Un graphe $G = (V, E)$ est défini par un ensemble de sommets $V = \{v_i\}$ et un ensemble d'arêtes $E = \{e_{ij}\}$. à savoir l'ensemble A et B, où $A \cup B = V$ et $A \cap B = \emptyset$. Les algorithmes basés sur les graphes tentent de minimiser certaines fonctions de coût, qui permet par la suite de couper certaines arêtes du graphe, le résultat de découpe peut construire la région d'intérêt, comme il est introduit par Shi et Malik [87].

2.5 Segmentation basée sur la classification

Une nouvelle famille de segmentation a vu le jour, la classification pixellique, ces techniques consistent à classer chaque pixel de l'image à une classe de région, et la classification se fait en utilisant les approches issues de l'intelligence artificielle. En effet, ces méthodes effectuent une séparation dans l'espace caractéristique de chaque pixel, l'espace de représentation peut être constitué par l'information couleur ou texture, de telle sorte qu'une projection dans cet espace réalise des frontières linéaires ou non-linéaires entre les régions de l'image. De nombreux travaux sur la classification pixellique ont montré l'efficacité de la segmentation [88], [89], [90], [91], [92].

Les différentes approches de segmentation par classification pixellique peuvent être regroupées en quatre grandes familles :

Méthode de classification par histogrammes : C'est une technique couramment utilisée dans la segmentation d'images couleur car elle présente l'avantage de ne pas utiliser de connaissance a priori sur l'image. Les méthodes d'analyse d'histogrammes se différencient par l'espace couleur choisi ou la composante couleur la plus significative [93] [94].

Méthode de classification par apprentissage non supervisé (Clustering) : Elle consiste à regrouper l'espace de représentation en région homogène suivant un critère de vraisemblance entre les pixels, cela est appliqué sans connaitre a priori les classes de région (sans apprentissage), souvent utilisant les algorithmes des k-moyennes [95], C-moyennes floues [96], [97] et Fisher [98], [99].

Méthode de classification par apprentissage supervisé : C'est la famille la plus efficace en segmentation, grâce à la phase d'apprentissage qui permis de générer une hypothèse robuste prenant en évidence les classes de chaque pixel. L'hypothèse fournie est capable de discriminer les régions d'intérêt avec succès, notamment lors d'utilisation

des classifieurs puissants, nous citons dans ce contexte : les Réseaux de Neurones Multi-Couches, les Machines à Supports de Vecteurs (SVM), les k plus proches voisins (k-PPV) et les arbres de décision. Et les travaux de Panda et Rosenfeld [92] et Vandenbroucke et al. [90] prouvent l'efficacité de cette famille de segmentation.

Méthode de classification par apprentissage semi-supervisé : En apprentissage supervisé, les algorithmes infèrent un modèle de prédiction à partir de données préalablement étiquetées. Cependant, l'étiquetage est un processus long et coûteux qui nécessite souvent l'intervention d'un expert. Cette phase contraste avec une acquisition automatique des données. Ce n'est alors pas rare de se retrouver avec un volume important de données dont seule une petite partie a pu être étiquetée. Par exemple, en recherche d'images par contenu, l'utilisateur souhaite étiqueter le minimum d'images pour fouiller une base aussi grande que possible.

La fiabilité de la segmentation à base de classification réside dans l'exploitation des techniques de l'intelligence artificielle pour procéder à la classification, notamment la classification par apprentissage supervisé. Évidemment, l'hypothèse apprise peut classer efficacement les pixels de l'image en région homogène. Deux facteurs clés qui provoquent la robustesse de la classification sont la qualité de caractérisation et la taille de la base d'apprentissage. La figure 13 illustre le protocole de segmentation par classification pixellique supervisée, le protocole est divisé en deux parties (apprentissage et test), si l'apprentissage est appliqué sur une base avec beaucoup d'échantillon, l'hypothèse fournie améliore significativement la qualité de segmentation, en contrepartie, cette famille nécessite un temps de calculs importants afin d'atteindre l'apprentissage et le test, cela démontre la complexité de cette technique. De ce fait, le passage de l'échelle pixel à l'échelle super-pixel peut résoudre la complexité provoquée par les techniques de segmentation par classification pixellique, ce dernier est abordé dans la section 3.

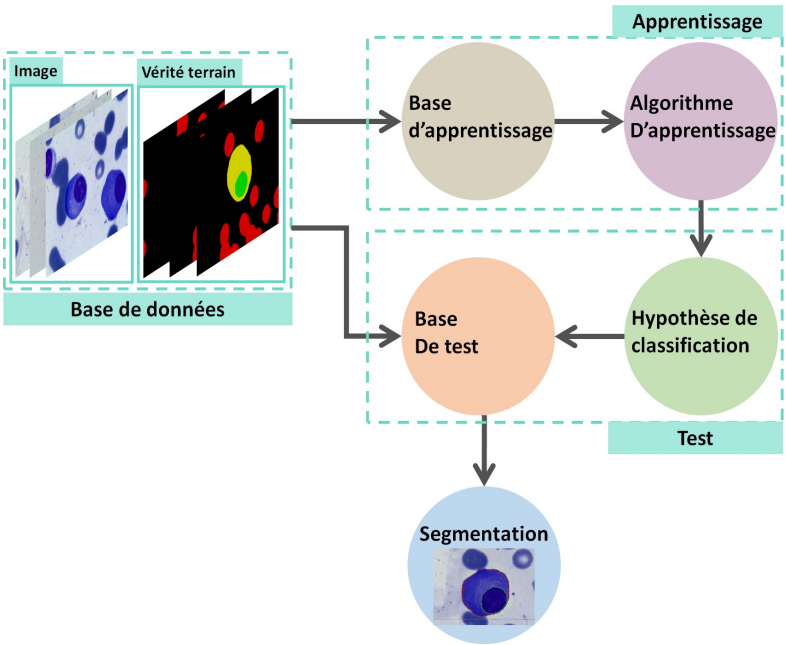


Figure 13 – Protocole de classification pixellique.

3 Segmentation par classification super-pixellique

Le terme super-pixel a été introduit par Ren et Malik en 2003 [100]. Son principe repose à décrire des groupes de pixels similaires en couleur ou autres propriétés de bas niveau. Le concept de super-pixel est motivé par deux aspects importants [100] : en premier

lieu, les images de taille importante qui peut causer un obstacle devant l'application de nombreux algorithmes de traitement d'images. Deuxièmement, les pixels ne sont pas des entités naturelles, ils ne sont qu'une conséquence des représentations discrètes.

Les super-pixels fournissent une primitive pratique à partir de laquelle on peut calculer les caractéristiques locales dans l'image. Ils capturent la redondance dans l'image et réduisent considérablement la complexité des tâches ultérieures de traitement d'images. Ils sont devenus des éléments essentiels dans de nombreux algorithmes de vision par ordinateur, tels que l'estimation de la profondeur [101], la segmentation d'images [102, 103], la squelettisation [104] et la localisation d'objets [105].

3.1 Algorithme du super-pixel

Récemment, les techniques de super-pixel ont reçu une attention particulière dans le domaine du traitement d'images et de la vision par ordinateur, notamment pour accélérer le processus de segmentation d'image [102]. La segmentation d'image en super-pixels vise à regrouper les pixels en régions homogènes partageant des caractéristiques similaires (texture, contour, couleur, etc ...) au sein de régions polygonales. La génération de super-pixel donne une description plus représentative comparée à celle du pixel. Il existe de nombreuses approches pour générer des super-pixels, chacune avec ses propres avantages et inconvénients. Les propriétés que l'on souhaite préserver dans la génération de super-pixel sont :

- Les super-pixels doivent bien adhérer au contour des régions dans l'image.
- Réduire la complexité de calcul en tant qu'étape de pré-traitement, les super-pixels doivent être rapides à calculer, efficaces et simples à utiliser.
- Les super-pixels doivent augmenter la vitesse et améliorer la qualité des résultats.

Une comparaison empirique a été étudiée par Achanta, et al. [8], prenant comme critère la vitesse et la qualité de sur-segmentation pour évaluer la segmentation obtenus [87], [106], [107], [108], [109], la conclusion est que la technique SLIC (Simple Linear Interactive Clustering) [110] a présenté son efficacité en terme de rapidité et de qualité de sur-segmentation.

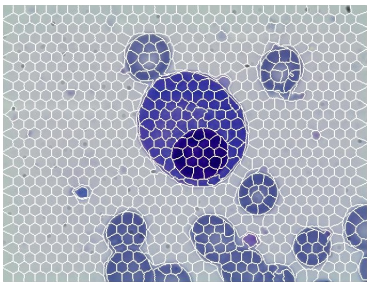


Figure 14 – Sur-segmentation obtenue à l'aide de l'algorithme SLIC

Dans ce travail, nous utilisons l'algorithme $SLIC$ (Simple Linear Iterative Clustering) qui a été introduit par Achanta et al. [110], l'algorithme $SLIC$ surpasse tous les algorithmes concurrents aussi bien d'un point de vue qualité de la segmentation que d'un point de vue vitesse d'exécution [110]. Il produit des super-pixels compacts et uniformes en couleur sans requérir à l'ajustement de multiples paramètres, comme le montre la Fig 14.

Cette méthode récente, permet de construire des super-pixels réguliers en surface. Tout d'abord, les centres des super-pixels sont initialisés sur une grille régulière, espacée de S pixels avec $S = \sqrt{\frac{N}{K}}$, où N est le nombre total de pixels dans l'image et K le nombre de super-pixels souhaités. Ils peuvent être éventuellement déplacés afin d'éviter de se trouver sur un contour de l'image. Cette méthode est itérative et comprend deux étapes :

1. L'assignation des pixels à un centre C_k suivant un critère d'appartenance,
2. La mise à jour des centres.

Cette approche tente de minimiser dans l'étape 1, le critère d'appartenance correspondant à une distance entre C_k et le pixel courant p définie par :

$$D_s(C_k; p) = d_{lab}(C_k; p) + \frac{m}{S} d_{xy}(C_k; p) \tag{1.2}$$

où d_{lab} est la distance calorimétrique et d_{xy} est la différence entre les positions dans l'image courante, telles que :

$$d_{lab}(C_k; p) = \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2}$$

$$d_{xy}(C_k; p) = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}$$

Ainsi, les paramètres sont le nombre approximatif de super-pixels K et leur compacité m. D'après Achanta et al. [110], $m \in [0 - 20]$ lorsque nous travaillons sur l'espace couleur Lab. Le terme m joue un rôle de pondération entre la couleur et la position. Quand $c = 0$ les super-pixels peuvent être très souples et adhèrent aux contours de l'image et quand $m = 20$ ils se rapprochent d'une forme régulière. Les auteurs proposent de fixer m à 10 car cette valeur permet d'obtenir des performances supérieures à celles réalisées par Felzenszwalb et Huttenlocher [111]. Pour chaque centre est défini une zone de recherche d'appartenance de taille (2S × 2S) et centrée sur C_k. Seuls les pixels p appartenant à cette zone sont parcourus comme le montre la Fig 15.

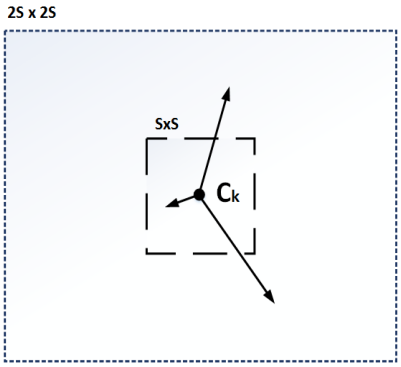


Figure 15 – Zone de recherche de pixels similaires au centre C_k de référence

3.2 Protocole de classification super-pixellique

L'objectif de ce travail est la segmentation automatique du *noyau* et *cytoplasme* pour la reconnaissance des globules blancs. Pour ce faire ; nous proposons une approche basée essentiellement sur une classification supervisée des super-pixels de l'image. L'intervention d'un hématologue expert est importante dans l'identification des régions *noyaux*, *cytoplasme*, *globule rouge* et *fond* d'images cytologiques. Le processus de segmentation proposé est illustré dans la Fig 16.

La première étape dans notre proposition est l'application de l'algorithme *SLIC* qui permet de générer un ensemble de super-pixels dans l'image, un super-pixel est composé d'un groupe de pixels partageant une similarité en fonction de la couleur.

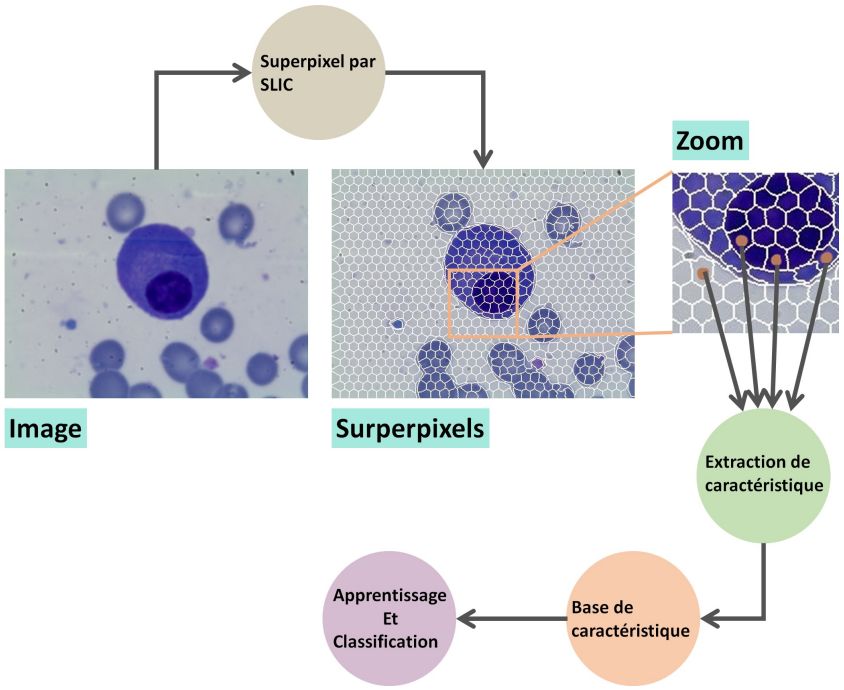


Figure 16 – Processus de segmentation par classification super-pixellique

L'extraction de caractéristique vient par la suite, cette étape a pour objectif de présenter chaque super-pixel sous forme d'un vecteur de descripteur. Le succès de l'apprentissage dépend essentiellement de la qualité de la base de caractéristique, de ce fait, le choix des descripteurs dans le processus de classification est primordial pour aboutir à une meilleure segmentation.

Dans ce chapitre, nous allons décrire les détails de la phases de l'extraction des caractéristiques. Cette dernière a pour rôle de traduire l'information du niveau de gris en vecteur caractéristique, différentes techniques ont été mises au point pour cette action. La section qui suit sera consacrée à exposer la caractérisation étudiée dans ce chapitre.

4 Extraction des caractéristiques

4.1 Normalisation d'image

Une caméra spécifique, un microscope et un logiciel sont les outils nécessaires pour une conception d'un environnement d'acquisition des images cytologiques. L'espace couleur RGB est le format fondamental utilisé dans la plupart des environnements d'acquisition des images couleurs. Des effets extérieurs et intérieurs de l'environnement peuvent modifier considérablement la qualité de l'image acquise (exemple : l'éclairage de l'environnement, dispositif de détection et les propriétés de la caméra). En imagerie cytologique, les cellules sanguines sont les régions cibles pour être bien discrétisées dans l'image acquise. Des effets secondaires d'acquisition sont observés sur l'information de texture et de couleur, qui par la suite, dégradent la distinction des régions d'intérêt (ROI). Cette dégradation cause des limites dans des processus de traitement d'images, notamment la segmentation et la classification basée sur l'information texture et couleur.

Dans ce contexte, certaines recherches procèdent à la normalisation comme phase de pré-traitement qui vise à produire une image plus uniforme en fonction de l'information couleur. Certainement, la normalisation des couleurs améliore significativement l'analyse de texture et de couleur comme a été prouvé par Cernadas et al. [11].

Dans la normalisation de couleur [112], des opérations mathématiques sont appliquées pour modifier l'intensité de la couleur de chaque pixel afin de réduire les effets issus de l'acquisition. Cernadas et al [11] ont étudié l'influence de la normalisation de couleur pour analyser cette dernière en utilisant les techniques les plus populaires dans la littérature, qui sont : Chroma, GWN, CGWN, HEQ, CLAHE, RGBcb, MV et Lmax.

Chroma est une normalisation qui est souvent utilisée par les systèmes de vison par ordinateur, elle est obtenue indépendamment de l'intensité d'éclairage observée utilisant la chromaticité des canaux rouge, vert et bleu suivant la formule 1.3.

$$(R', G', B') = \left(\frac{R}{R+G+B}, \frac{G}{R+G+B}, \frac{B}{R+G+B}\right) \tag{1.3}$$

GWN (Grey world Normalization) est une normalisation basique qui suppose que les changements dans le spectre d'éclairage peuvent être modélisés par trois facteurs constants (R_{moy}, V_{moy} et B_{moy}) appliqués aux canaux de couleur rouge, vert et bleu. Les constances sont alors obtenues en divisant chaque canal de couleur par sa valeur moyenne (R_{moy}, V_{moy} et B_{moy}) et l'image normalisée est calculée comme indiquée dans la formule 1.4 :

$$R' = \frac{R}{R_{moy}}, G' = \frac{G}{G_{moy}}, B' = \frac{B}{B_{moy}} \tag{1.4}$$

CGWN (Comprehensive Gray World Normalized), Finlayson et al [113] ont proposé de combiner la technique **CHROMA** et **GWN** en exploitant l'information locale de chaque canal de l'espace RGB, et l'information globale de la chromaticité par la sommation des trois canaux rouge, vert et bleu. La normalisation est calculée itérativement et successivement en appliquant l'équitation 1.3 et 1.4. Cela converge vers une représentation normalisée en fonction de l'intensité et la luminance de couleur.

HEQ (Histogram EQualization) ou égalisation d'histogramme est une approche développée à l'origine pour des images en niveau de gris, dont l'objectif est d'augmenter le contraste global de l'image. Une transformation est appliquée pour avoir un histogramme uniforme, ce qui maximise l'information d'entropie dans l'image. Finlayson et al [113] ont proposé d'utiliser le même procédé pour chaque canal de l'image couleur comme une représentation normalisée, améliorant la qualité de l'image et éliminant les effets d'éclairage produits par l'acquisition.

CLAHE (Contrast Limited Adaptive Histogram Equalization) est une extension améliorée d'égalisation d'histogramme, essentiellement développée pour l'imagerie médicale et s'est révélée efficace pour l'amélioration des images à faible contraste, mais également aussi pour résoudre les problèmes d'amplification du bruit. Exactement comme **HEQ**, **CLAHE** peut être utilisée comme une technique de normalisation en appliquant le **CLAHE** pour chaque espace de l'image couleur, les résultats sont des images normalisées et améliorées en fonction du contraste et entropie.

RGBcb (RGB color balance) est une famille d'algorithme qui sert à corriger des images prises dans un environnement lumineux naturellement ou artificiellement. L'hypothèse de cet algorithme suppose que les intensités élevées de l'espace RGB correspondent au blanc et les faibles intensités à l'obscurité. L'objectif est d'étaler autant que possible les valeurs d'intensité de l'espace RGB, de sorte qu'ils occupent la plage maximale possible entre [0, 255]. La proposition de Lumire et al [114] est d'appliquer une transformation linéaire à chaque espace.

MV (Mean Variance), statistiquement une image normalisée se caractérise par une moyenne qui tend vers 0 et une variance unitaire, ce procédé est appliqué sur des images en niveau de gris. Cernadas et al [11] ont proposé l'application de la technique **MV** pour chaque espace de l'image couleur.

L$_{max}$ est une simple procédure appliquée sur chaque espace couleur afin de minimiser les effets d'éclairage produits en phase d'acquisition, soit $L_{i,max}, i \in r, g, b$ la luminance maximale de chaque espace i, et chaque pixel de l'image est contrôlé par cette luminance maximale de la manière suivante :

$$(R', G', B') = \left(\frac{R}{L_{r,max}}, \frac{G}{L_{g,max}}, \frac{B}{L_{b,max}}\right) \tag{1.5}$$

Les techniques de normalisation citées ci-dessus peuvent être divisées en deux catégories, la première fournit des représentations normalisées tout en préservant ou améliorant la qualité visuelle de l'image selon la perception humaine comme **GWN** et L_{max} (Fig 17), tandis que la deuxième fournit des résultats moins appropriés à la visualisation comme **CGWN** et **Chroma** (Fig 17), par contre cette normalisation améliore significativement les résultats de traitement.

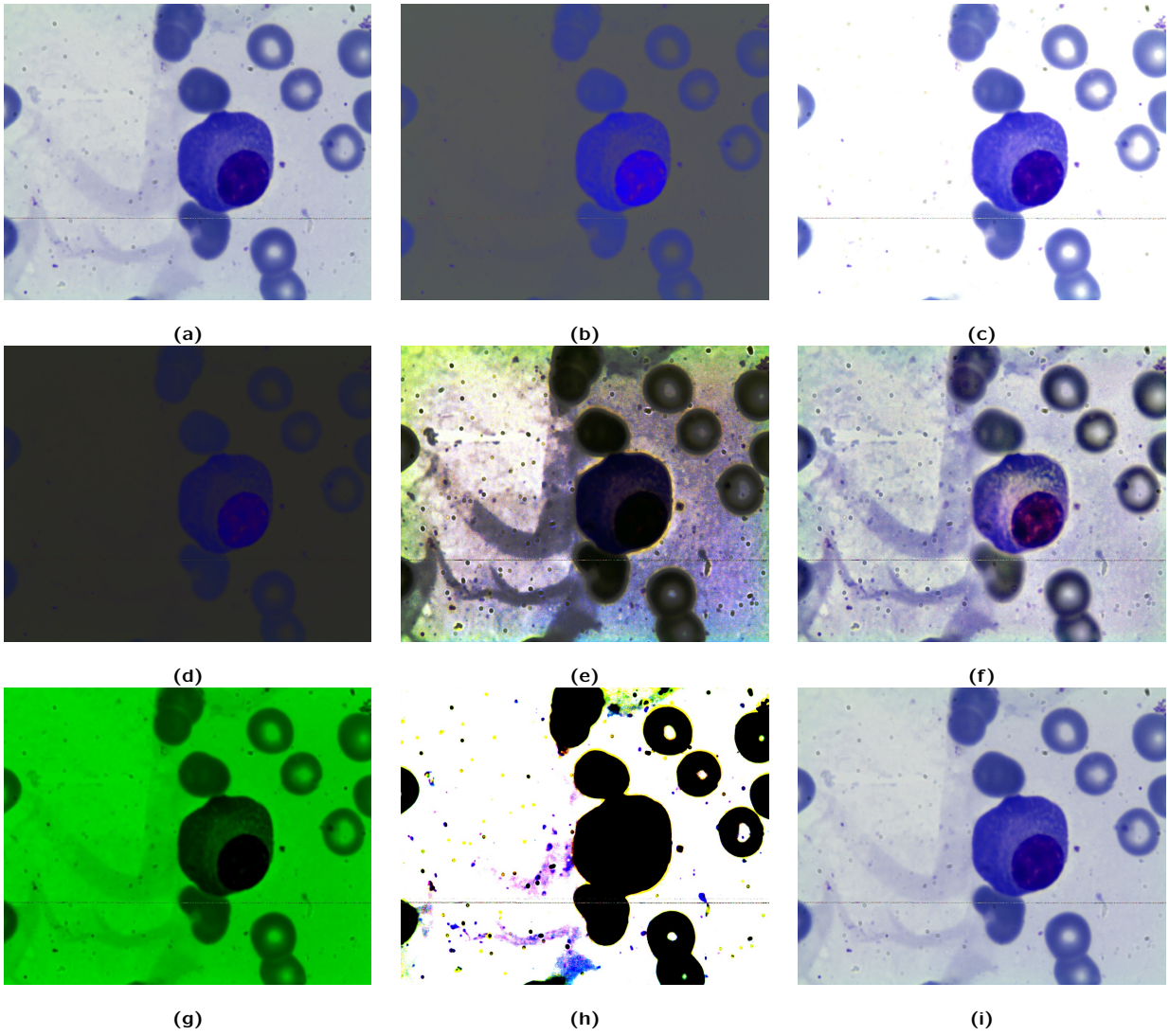


Figure 17 – Exemple de normalisation, (a) : sans normalisation, (b) : Chroma, (c) : GWN, (d) : CGWN, (e) : HEQ, (f) : CLAHE, (g) : RGBcb, (h) : MV et (i) : L_{max}.

4.2 Les espaces couleurs

Une caméra couleur enregistre des images couleurs où chaque pixel de l'image est caractérisé par une composante rouge, vert et bleu. Dans la littérature, l'analyse des images couleurs ne se base pas que sur l'espace RGB, de nombreuses manipulations ont été proposées pour générer des propriétés de couleurs différentes par un mélange approprié de couleurs primaires.

Trois couleurs primaires existantes sont alors nécessaires et suffisantes pour reproduire une liste exhaustive d'espace couleur, cette liste est synthétisée par Vandenbroucke et al. [1]. Récemment, l'analyse des espaces couleurs est un sujet qui a attiré l'attention de plusieurs chercheurs [1, 11, 115], réalisant des études comparatives de la classification basée sur différents espaces couleur. La synthèse de ces travaux ne permet pas de conclure sur la définition d'un seul espace colorimétrique bien adapté à la classification, évidemment les résultats de classification peuvent être améliorés en manipulant différents espaces couleur.

La composante couleur est l'élément essentiel dans notre problématique de segmentation, qui provoque la question suivante : *quel est le meilleur espace couleur pour aboutir à une bonne segmentation des cellules sanguines ?*. Néanmoins, le bon choix de la composante couleur peut apporter des améliorations considérables aux résultats de segmentation, selon Vandenbroucke et al. [1], les espaces couleurs peuvent être classés en quatre familles en fonction de leurs propriétés.

Les espaces primaires

La perception de la couleur se définit dans un contexte tridimensionnel comme énoncée par Young-Helmholtz en 1866, l'échelle couleur est la synthèse d'un stimulus de couleur (ensemble de rayons lumineux réfléchis ou transmis par un matériau) en mélangeant les quantités appropriées des trois couleurs primaires : le rouge, le vert et le bleu. Cette famille est divisée en deux groupes :

– Les espaces primaires réels : Ils représentent les propriétés physiques des couleurs primaires, reproduites et dépendantes de la caméra d'acquisition, qui est bien l'espace RGB. L'espace rgb est la version normalisée de RGB, définie par :

$$(r, g, b) = \left(\frac{R}{R+G+B}, \frac{G}{R+G+B}, \frac{B}{R+G+B}\right) \tag{1.6}$$

– Les espaces primaires virtuelles : Indépendamment du dispositif d'acquisition, la propriété primaire réelle RGB peut être convertie en espace de couleur virtuelle CIE [1] (X, Y, Z) par des transformations linéaires. Les coordonnées normalisées (x, y, z) peuvent également être déduites de CIE (X, Y, Z) pour caractériser la couleur.

Les espaces luminance-chrominance

La particularité des espaces de luminance-chrominance réside dans la composante dite luminance qui représente l'information achromatique de couleur, ainsi que deux autres composants de chrominance (**Chr1** et **Chr2**) qui quantifient le caractère de couleur. Une transformation linéaire et non linaire des espaces primaires produisent les escapes luminance-chrominance. Différents espaces couleurs de cette famille sont mis à disposition dans la littérature, on trouve, les espaces perceptuellement uniformes (où la luminance est notée **L**), les espaces antagonistes (où la luminance est notée **A**) ou encore les espaces de télévision (où la luminance est notée **Y**).

1. CIE : Commission Internationale de l'Eclairage

– Les espaces perceptuellements uniformes : Visuellement, Des couleurs primaires perceptuellement proches sont définies par un écart important dans l'espace de représentation adoptée, tandis qu'un écart faible peut correspondre à des couleurs perceptuellement différentes, d'où l'intérêt d'utiliser des espaces de représentation perceptuellement uniformes. Comme les espaces (Luv) et (Lab) sont deux espaces perceptuellement uniformes définis par le CIE.

– Les espaces antagonistes : Cette famille d'espace couleur a été créé dont l'objectif est de modéliser le système visuel humain. Basée sur la théorie des couleurs opposées initialement proposée par Hering. L'hypothèse de cette théorie est que l'œil humain joue le rôle d'un capteur de l'information couleur transmise par la suite au cerveau sous forme de trois composantes, une luminance et deux chrominances, correspondant respectivement à un signal d'opposition vert-rouge et à un signal d'opposition jaune-bleu.

– Les espaces de télévision : Cette famille a été conçue pour une question de compatibilité entre les nouveaux téléviseurs en couleur et les anciens téléviseurs noir et blanc. Une transformation de type linéaire des composantes trichromatique du système (Rc,Gc,Bc) est appliquée afin de séparer l'information de luminance de celle de chrominance, ainsi La luminance correspond à la composante Y du système (X,Y,Z).

Les espaces perceptuels

Les espaces perceptuels tentent de quantifier la perception subjective de la couleur humaine en utilisant l'intensité, la teinte et les composantes de saturation. La teinte correspond aux dénominations des couleurs telles que le rouge, le vert, le bleu, jaune, etc... La saturation est une grandeur permettant d'estimer le niveau de coloration d'une teinte indépendamment de sa luminosité. On distingue deux types d'espaces perceptuels :

– Les espaces de coordonnées polaires (ou cylindriques) qui correspondent aux expressions dans les coordonnées polaires des composantes de luminance-chrominance dans lesquels, la représentation de la couleur se fait avec un axe pour la luminosité et un plan pour la chrominance.

– Les espaces de coordonnées perceptuelles qui sont directement évalués à partir des espaces primaires et représentent la couleur en termes d'intensité I, de saturation S et de teinte T. L'intensité correspond à l'information de luminosité, mais elle est désignée ainsi principalement pour différencier des espaces de coordonnées polaires.

Les espaces d'axes indépendants

Les composantes primaires se caractérisent par une forte corrélation entre les composantes R, G et B, cette corrélation est considérée comme inconvénient majeur car les composantes partagent une information commune, puisque ne possèdent qu'une information de luminance. Plusieurs travaux existants ayant comme objectif de générer des espaces couleurs de plus en plus indépendants, t ainsi les composantes générées portent des informations différentes et non redondantes. Une solution considérée comme efficace est proposée elle réside dans l'application de l'analyse en composantes principales (ACP) [116], afin d'obtenir un système de représentation dont les composantes sont statistiquement indépendantes.

Nous prenons comme référence les travaux de Vandenbroucke et al. [1] afin de regrouper les espaces couleurs appropriés à chaque sous-famille, comme il est indiqué dans le tableau 11.

Famille	Sous-famille	Espace couleur	Nom et/ou référence
Espaces primaires	Espaces primaires réels	(R, G, B)	Espace d'acquisition
		(r, g, b)	La normalisation de (R,G,B) [117]
	Espaces primaires virtuelles	(X, Y, Z)	CIE 1931 [118]
		(x, y, z)	CIE 1931 [118]
Espaces luminance-chrominance	Espaces antagonistes	$(A, C1, C2)$	Espace de Garbay [119]
		(bw, rg, by)	Espace de Ballard [120]
	Espaces de télévision	(Y', I', Q')	Espace télévision NTSC [121]
		(Y', U', V')	Espace télévision EBU [117]
	Espaces perceptuellements uniformes	(L, a, b)	Espace CIELAB [118]
		(L, u, v)	Espace CIELUV [118]
	Autres	(Y, Ch_1, Ch_2)	Espace de Carron [122]
		(Y, x, y)	CIE 1931 [118]
		$(I1, r, g)$	[123]
Espaces perceptuels	Espaces de coordonnées polaires	$(I1, S1, H1)$	Modèle triangle HSI [124]
		$(I1, S2, H1)$	Modèle triangulaire modifié HSI [124]
		$(I4, S3, H2)$	Modèle HSV hexcone [124]
		$(I5, S4, H2)$	Modèle HLS double hexcone }citeshih1995reversibility
		$(I6, S5, H1)$	Modèle HLS amélioré [125]
		(L, S_{UV}, H_{UV})	CIE 1931 [118]
	Espaces de coordonnées perceptuelles	(A, C_{C1C2}, h_{C1C2})	Espace de coordonnées polaires de Garbay [119]
		(bw, C_{rgby}, h_{rgby})	Espace de coordonnées polaires de Ballard [120]
		(Y', C'_{IQ}, h'_{IQ})	Espace de coordonnées polaires de NTSC [121]
		(Y', C'_{UV}, h'_{UV})	Espace de coordonnées polaires de EBU [117]
		(L', C_{ab}, h_{ab})	Espace de coordonnées polaires de CIELAB [118]
		(L', C'_{UV}, h_{UV})	Espace de coordonnées polaires de CIELUV [118]
		$(L, C_{Ch_1Ch_2}, h_{Ch_1Ch_2})$	Espace de coordonnées polaires de Carron [122]
		$(I1, C_{I2I3}, h_{I2I3})$	Espace de coordonnées polaires de Ohta [117]
Espaces d'axes indépendants		$(I1, I2, I3)$	Espace de Ohta [123]

Table 11 – Espaces couleurs [1].

Dans ce contexte, une question est souvent posée : *Quel est le meilleur espace couleur ?*, de nombreux travaux ont essayé de répondre à cette question par des études comparatives exploitants plusieurs espaces couleurs dans un processus de traitement d'image. Vandenbroucke et al. [1] ont présenté un résumé de certains travaux qui propose des réponses à la question posée, se basant sur une comparaison sélective, le tableau 12 présente le résumé proposé par [1].

Les travaux cités dans le tableau 12 ont étudié la composante couleur dans un contexte de segmentation et de classification afin de répondre à la question précédente, plusieurs espaces couleurs ont été exploités certainement de chaque famille, et chaque travail a reposé sur une étude comparative pour sélectionner l'espace le plus pertinent, comme il est indiqué dans la troisième colonne du tableau 12. La synthèse de ces travaux peut se mettre en accord avec 3 espaces dites pertinents, qui sont : (R,G,B) de la famille espaces primaires, (L,a,b) de la famille espaces luminance-chrominance et (I1,I2,I3) de la famille espaces d'axes indépendants, la même conclusion a été constatée par Cernadas et al. [11], qui se sont focalisés seulement sur l'étude de 4 espaces couleurs partagés sur les 4 familles, qui sont :

- Espace primaire : (R,G,B).
- Espaces luminance-chrominance : (L,a,b).
- Espaces perceptuels : (H,S,V).
- Espaces d'axes indépendants : (I1,I2,I3).

Référence	**Espaces étudiés**	**Espaces sélectionnés**
[123]	(R,G,B), (X,Y,Z), (Y',I',Q'), (U,V,W), (L,a,b), (I1,r,g), (I1,I2,I3), (I1,S1,H1)	(I1,I2,I3), (L,a,b)
[126]	(R,G,B), (X,Y,Z), (Y',I',Q'), (U,V,W), (I1,I2,I3)	(I1,I2,I3), (R,G,B)
[127]	(R,G,B), (X,Y,Z), (Y',I',Q'), (L,a,b), (I1,I2,I3), (I4,S3,H2) ;	(I1,I2,I3)
[128]	(R,G,B), (X,Y,Z), (Y',I',Q'), (U,V,W), (L,a,b), (I1,r,g), (I1,I2,I3), (I1,S1,H1)	(I1,I2,I3)
[129]	(R,G,B), (Y,U,V), (Y,QRG,QRB)	(R,G,B)
[130]	(R,G,B), (Y',I',Q'), (U,V,W), (L,a,b), (I1,I2,I3)	(I1,I2,I3)
[131]	(R,G,B), (Y',U',V'), (Y',C'b,C'r), (L,a,b), (L,u,v)	(L,a,b)
[132]	(R,G,B), (Y',C'b,C'r), (L,a,b), (I1,r,g), (I4,S3,H2)	(Y',C'b,C'r)
[133]	(R,G,B), (X,Y,Z), (Y',I',Q'), (Y,U,V), (Y,Cb,Cr), (L,a,b), (L,u,v), (Y,Ch1,Ch2), (I1,I2,I3), (I1,S1,H1)	(I1,I2,I3)
[134]	(R,G,B), (X,Y,Z), (Y',U',V'), (I4,S3,H2), (C,M,Y)	(R,G,B)
[135]	(R,G,B), (L,a,b), (L,u,v)	(L,u,v)

Table 12 – Résumé d'une comparaison des meilleurs espaces couleurs [1].

4.3 Caractérisation super-pixellique

La représentation numérique d'une caractéristique visuelle telle que la couleur, la texture ou la forme est l'étape la plus importante dans un processus de reconnaissance automatique, ce procédé est reconnu par l'effet d'extraction de variable. Plusieurs modes de caractérisation sont proposés dans la littérature qui peuvent être exploités dans de nombreuses applications. Toutefois, le choix des caractéristiques efficaces pour résoudre une problématique bien précise est devenue une question importante à laquelle nous nous intéressons. Notre travail repose essentiellement sur l'étude de l'effet de la composante

couleur dans la classification super-pixellique pour la segmentation des globules blancs à partir des images cytologiques.

Une fois l'image est subdivisée en super-pixels, chacun peut se caractériser par 3 composantes couleurs $SP_j = \{E_1, E_2, E_3\}$ (Fig 18), et dans chaque composante couleur, le super-pixel est caractérisé par des intensités en niveau de gris ($SP_i = \{I_1, I_2, ..., I_N\}$). On fait appel à des techniques de caractérisation de type couleur pour traduire la composante couleur en un vecteur caractéristique (VC) (Fig 18), qui sera présenté dans un classifieur par la suite.

L'objectif de cette phase est l'identification des caractéristiques dites pertinentes, ayant le pouvoir de bien discriminer les régions d'intérêts. Certainement, la qualité de classification dépend de la qualité de caractérisations, et la caractérisation non pertinente ou redondante peut dégrader les résultats de segmentation.

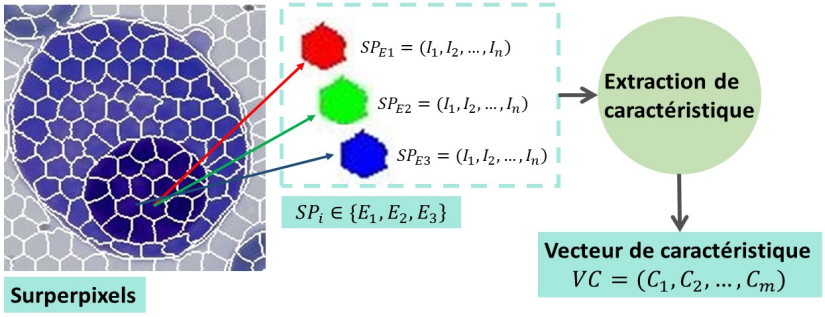


Figure 18 – Caractérisation super-pixellique.

Couleur moyenne (Color Mean (CM))

Est l'une des caractérisations les plus basiques existantes dans la littérature, Le vecteur caractéristique CM inclus simplement 3 paramètres $\{\mu^{E_1}, \mu^{E_2}, \mu^{E_3}\}$ issues de la moyenne des intensités I_n de chaque espace couleur $E_j \in \{E_1, E_2, E_3\}(Eq.1.7)$.

$$\mu^{E_j} = \frac{1}{N}\sum_{i=1}^{N} I_i \quad , j = 1, 2, 3 \tag{1.7}$$

Statistiques de premier ordre (First-Order Statistics (FOS))

Des mesures statistiques peuvent être appliquées pour chaque super-pixel, elles permettent de fournir des informations sur la répartition des niveaux de gris dans chaque espace couleur $E_j \in \{E_1, E_2, E_3\}$. Ce sont des mesures simples mais considérées comme efficaces. Le vecteur de caractéristique (FOS) comprend 5 mesures statistiques sur chaque composante couleur (15 caractéristiques au total) : le niveau de gris moyen (μ^{E_j}), la variance (σ^{E_j}) le troisième ($m_3^{E_j}$) et le quatrième ($m_4^{E_j}$) moments statistiques avec l'entropie (H^{E_j}).

$$\mu^{E_j} = \frac{1}{N}\sum_{i=1}^{N} I_i \quad , j = 1, 2, 3 \tag{1.8}$$

$$\sigma^{E_j} = \sqrt{\sum (i - \mu^{E_j})^2 I_i} \quad , j = 1, 2, 3 \tag{1.9}$$

$$m_3^{E_j} = \sum (i - \mu^{E_j})^3 I_i \quad , j = 1, 2, 3 \tag{1.10}$$

$$m_4^{E_j} = \sum (i - \mu^{E_j)^4} I_i \quad , j = 1, 2, 3 \tag{1.11}$$

$$H^{E_j} = -\sum I_i log(I_i) \quad , j = 1, 2, 3 \tag{1.12}$$

Contraste

Par définition, le contraste produit la différence d'information entre une région donnée et son voisinage en fonction de l'intensité du niveau de gris. Dans notre application, le super-pixel est considéré comme la région donnée. Plus la différence de valeur du niveau de gris est faible, plus le contraste est faible, ce qui indique une forte similarité avec les super-pixels voisins. Domingo Mery [136] a défini le contraste comme suit :

$$C_1 = \frac{\mu_{SP_i} - \mu_{SP_j}}{\mu_{SP_j}}, \qquad C_2 = \frac{\mu_{SP_i} - \mu_{SP_j}}{\mu_{SP_i} + \mu_{SP_j}}, \qquad C_3 = ln(\mu_{SP_i}/\mu_{SP_j}) \tag{1.13}$$

Où μ_{SP_i} et μ_{SP_j} indiquent la valeur moyenne de niveau de gris dans le super-pixel SP_i et dans le super-pixel voisin SP_j respectivement.

Deux autres caractéristiques de contraste sont suggérées par [137], qui peuvent être résumé en quatre étapes :

1. Nous prenons le profil dans la direction x et dans la direction y positionnée au centre de gravité de la région (SP_i et SP_j respectivement).
2. Nous calculons les rampes R_1 et R_2, considérés comme l'estimation d'une fonction du premier ordre qui contient l' intensité maximale et minimale d'un super-pixel SP_i.
3. Nous calculons l'ensemble de profil $Q = \{Q_1, Q_2\}$:

$$Q_1 = I_{SP_i} - R_1, \qquad Q_2 = I_{SP_i} - R_2 \tag{1.14}$$

4. Les nouvelles caractéristiques de contraste sont données par :

$$C_4 = \sigma_Q \quad (\sigma_Q = carttype(Q)), \qquad C_5 = ln(Q_{max} - Q_{min}) \tag{1.15}$$

Moments de Hu avec intensité

Les moments et les invariants associés ont été largement analysés dans de nombreuses applications de reconnaissance de formes pour caractériser une région dans l'image, nous citons : les moments géométriques [138], des moments proches [139], des moments de rotation [140] et les moments complexes [141].

Les moments invariants ont été introduits par Hu. [138], Hu a défini 6 invariants orthogonaux absolus et un invariant orthogonal oblique, ces mesures sont basées sur des invariants algébriques qui sont non seulement indépendants de la position, de la taille et de l'orientation, mais également indépendants de la projection parallèle. Statistiquement le moment est défini par :

$$m_{rs} = \sum_{i,j} i^r j^s \qquad avec \quad r, s \in N \tag{1.16}$$

Domingo Mery [136] a considéré l'information du niveau de gris dans les moments de Hu comme une caractérisation pertinente dans la reconnaissance automatique, moments de Hu avec intensité sont calculés comme suit :

$$m_{rs} = \sum_{i,j} i^r j^s I(i,j) \tag{1.17}$$

Avec $r, s \in N$ et I intensité du niveau de gris. Le paramètre $r + s$ est appelé l'ordre du moment.

5 Plan d'expérimentations

La classification super-pixellique est devenue une approche très attrayante dans la segmentation des images couleurs. Les épreuves de l'intelligence artificielle en segmentation provoquent une efficacité appréciable de segmentation via un concept de classification pixellique. Généralement ce concept procède à un parcours de pixel par pixel pour l'action de segmentation, cela cause des limites et des complexités de calcul. Une solution est proposée, il est intéressant de manipuler un sous-ensemble de pixel (super-pixel) au lieu de pixel par pixel. La solution est donc de passer de l'échelle du pixel à l'échelle du super-pixel intégrant un traitement qui nous génère des sous-régions homogènes dans l'image, dite super-pixel.

Pour ce faire, il est nécessaire de passer par une phase de pré-traitement du supe-pixel, et une phase de caractérisation super-pixellique (Fig.16), et ceci dans le but de préparer les données qui seront présentées à un classifieur par la suite. Le succès de la classification réside dans l'efficacité des données préparées, de ce fait, il nous est apparu évident d'étudier attentivement la phase de caractérisation dans laquelle nous nous attelons à répondre aux questions liées à ce contexte, qui sont :

- La normalisation de couleur a-t-elle une influence sur la segmentation ?
- Quel est le meilleur espace couleur ?
- Quelle est le meilleur mode de caractérisation ?

Un plan d'expérimentations a été mis en place afin de mieux répondre aux questions posées, la figure 19 résume les étapes de notre plan d'expérimentation. Les traitements appliqués commencent à partir de **l'étape 1** jusqu'à **l'étape 5**.

- **L'étape 1** a pour but d'étudier l'influence de la normalisation couleur en utilisant différente technique de normalisation, ceci nous permet de comparer les résultats par rapport à la segmentation sans normalisation.
- **L'étape 2** sert uniquement à traiter individuellement chaque espace couleur à part dans le processus de segmentation.
- **L'étape 3** a pour but d'application l'algorithme SLIC pour générer les super-pixels dans l'image.
- **L'étape 4** est étape clef dans le processus de classification, elle consiste à analyser différent mode de caractérisation en fonction de chaque espace couleur.
- **L'étape 5,** les données préparées issues de l'étape 4 sont présentées à un algorithme d'apprentissage supervisé, afin de générer une hypothèse de classification qui sera exploitée dans la classification super-pixellique.

Dans le plan présenté, 3 étapes importantes font l'objectif de cette étude, la normalisation couleur (**étape 1**), la composante couleur (**étape 2**) et la caractérisation super-pixellique (**étape 4**). Notre proposition consiste à manipuler 9 procédés de normalisation, 4 espaces couleurs et 4 techniques de caractérisations, ce qui fait en tous 144 combinaisons possibles pour analyser individuellement chaque aspect dans le processus de segmentation. La figure 20 présente l'exemple d'une expérience de la normalisation CGWN en fonction de 4 espaces couleurs et de 4 techniques de caractérisations.

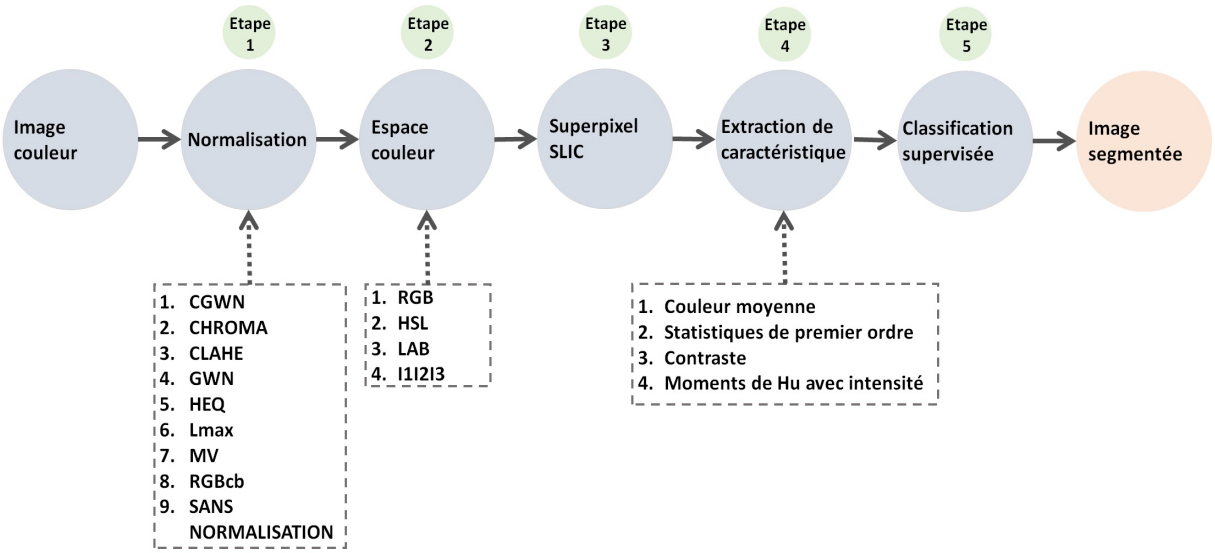


Figure 19 – Plan d'expérimentations.

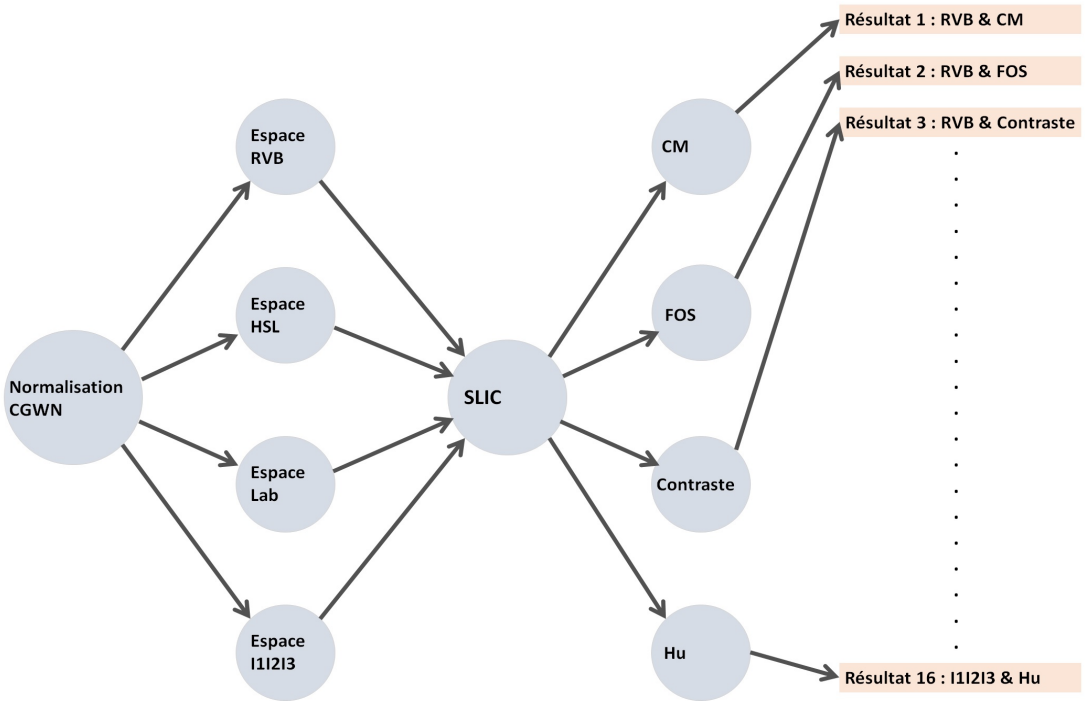


Figure 20 – Exemple d'une expérimentation.

6 Résultats et expérimentations

6.1 Base de données

La base de données des images cytologiques a été construite à partir d'images réelles acquises au sein du service d'hémobiologie (CHU Tlemcen), sur des lames la coloration de type MGG (May Grunwald Giemsa). L'environnement LEICA (caméra et microscope) permet d'obtenir des images couleur RGB de taille 768x1024. 87 images cytologiques avec vérité terrain ont été mises en place pour aborder l'étude expérimentale.

6.2 Expérimentations

La génération des super-pixels représente l'étape initiale dans la classification super-pixellique. Dans nos expérimentations, on fait appel à l'algorithme SLIC pour cette action, c'est un algorithme qui a été exploité dans de nombreuses applications de segmentation. En effet, SLIC possède deux contraintes qui sont la compacité des super-pixels m et le nombre de super-pixel k (section 3). Achanta et al. [8] ont fixé un intervalle de $m = [10 - 20]$ pour générer une bonne qualité de super-pixel.

Afin de trouver le bon équilibre entre la similitude des couleurs et la proximité spatiale en super-pixel des images cytologiques, nous avons simulé des super-pixels en faisant varier les différentes contraintes de $m = \{10, 15, 20\}$ et $k = \{100, 500, 1000\}$ comme il est présenté dans la figure 21. Qualitativement, nous estimons que le meilleur compromis est obtenu avec $m = 20$ et $k = 500$, ces dernières valeurs vont servir pour réaliser nos expérimentations comparatives. Concernant la question quantitative, nous réservons dans le chapitre suivant une étude approfondie détaillant l'influence de la compacité m et le nombre k sur la segmentation.

La réussite de notre proposition de classification super-pixellique réside d'une part dans l'exactitude des données préparées, et d'autre part, dans la robustesse de l'hypothèse de classification. Dans cette première partie d'expérimentation, nous avons décidé d'utiliser le classifieur de type SVM (Support Vector Machine) car il est considéré comme un classifieur de référence pour une variété d'applications de vision par ordinateur.

Pour la suite de notre expérimentation, nous avons sélectionné 15% de la base de données (13 images) pour préparer une base d'apprentissage, 144 (9x4x4) bases ont été préparées manipulant toutes les combinaisons possibles entre 9 types de normalisation, 4 espaces couleurs et 4 techniques de caractérisation.

L'expert hématologiste interviendra dans l'étiquetage de ces treize images par une sélection de super-pixel de chaque région d'intérêt (ROI), permettant ainsi une meilleure perception des régions *noyau*, *cytoplasme*, *globule rouge* et *fond*. Dans la partie d'apprentissage supervisé de SVM, une fonction gaussienne avec une séparation non-linéaire a été choisie. L'évaluation est réalisée par une validation croisée égale à 5.

6.3 Résultats

Le taux de reconnaissance (accuracy) est utilisé comme paramètre de référence pour la validation quantitative d'un système de classification. Néanmoins, dans la segmentation des images, le taux de reconnaissance n'est pas un paramètre pertinent pour juger de la bonne qualité de segmentation. Dans notre application, dans la segmentation du *noyau* et du *cytoplasme*, l'estimation du taux de reconnaissance peut donner une mesure qui n'a aucun rapport avec la segmentation obtenue et celle de l'expert.

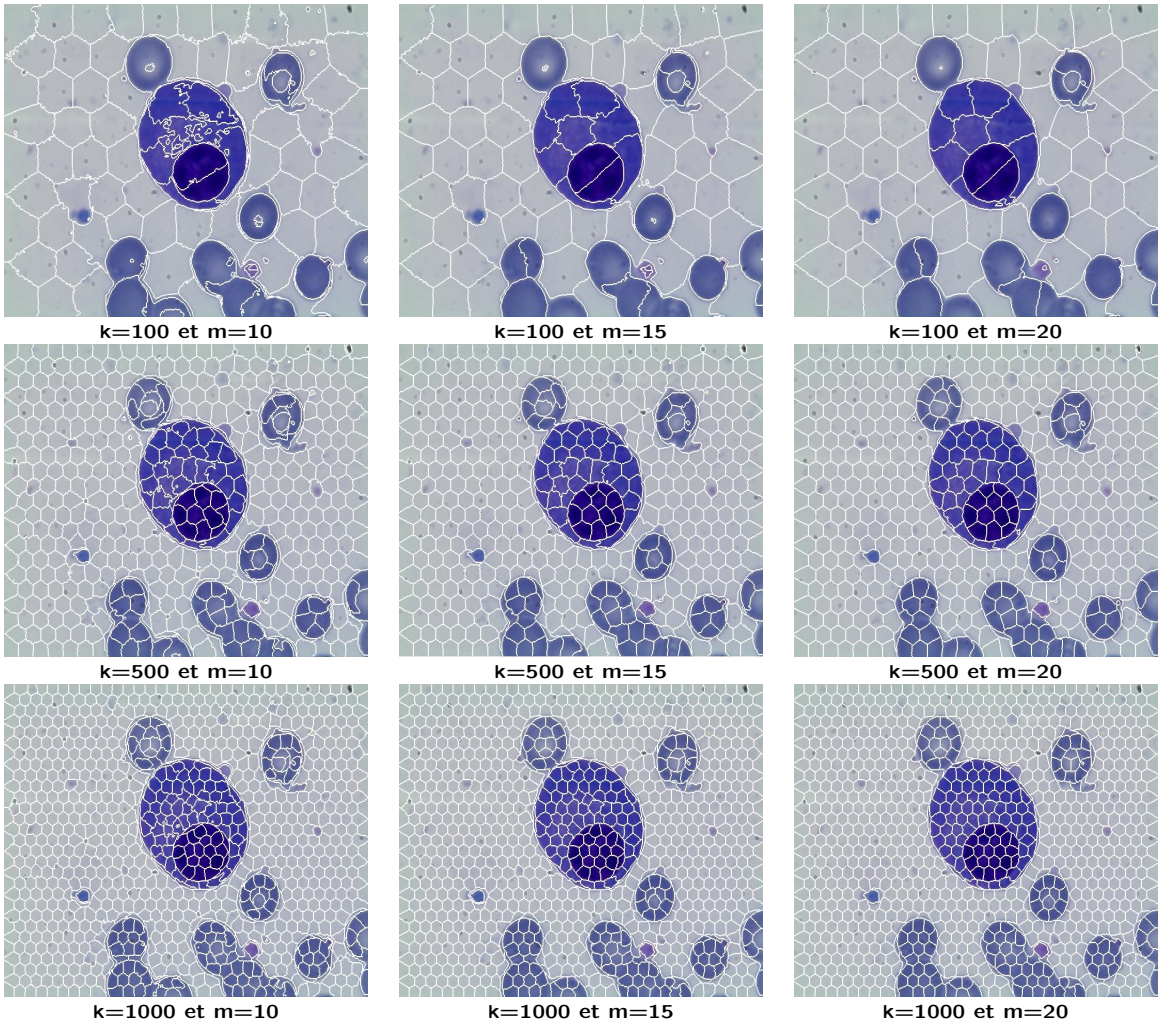


Figure 21 – Exemple de super-pixel avec différente contrainte de SLIC.

Pour démontrer cette non pertinence, nous proposons l'exemple de la figure 22 qui représente un exemple de segmentation du *cytoplasme*. La matrice de confusion de cet exemple est résumée dans la table 13.

Image couleur **Vérité terrain de la région cytoplasme** **Résultats de segmentation de cytoplasme**

Figure 22 – Exemple du calcul de taux de reconnaissance (accuracy).

		Segmentation	
		Cytoplasme	Non-cytoplasme
Vérité terrain	Cytoplasme	31066	30281
	Non-cytoplasme	0	725085

Table 13 – Matrice de confusion

Suite aux résultats de la matrice de confusion, nous obtenons un taux de reconnaissance de **96.14%**. On remarque que le taux obtenu est très élevé par contre, il ne reflète pas vraiment la qualité de la segmentation, il est clair que le résultat de la segmentation (figure 22) a quasiment négligé toute une région du *cytoplasme*.

En conséquence dans notre étude, nous avons fait appel à un autre paramètre pour analyser quantitativement les performances de segmentation. Cette évaluation est basée sur la similitude entre les régions et les contours détectés par notre méthode en comparaison avec ceux des experts. Pour estimer le chevauchement entre les régions segmentées et celles vérité terrain, les mesures de précision et du rappel des super-pixels sont calculées.

$$Precision = \frac{VP}{VP + FP} \qquad Rappel = \frac{VP}{VP + FN}$$

Avec : VP : Vrai Positif : nombre de super-pixels positifs classés positifs. FP : Faux Positif : nombre de super-pixels négatifs classés positifs. FN : Faux Négatif : nombre de super-pixels positifs classés négatifs.

Pour mieux apprécier la qualité des résultats de notre méthode, nous faisons appel à une mesure de performance appelée F-score (F) qui est la moyenne harmonique de la précision et du rappel. Elle est définie comme suit :

$$F - score = 2 \ \frac{Precision \ . \ Rappel}{Precision + Rappel}$$

La valeur de $F - score$ se situe entre $[0 - 1]$, une valeur F-score élevée nous renseigne sur la pertinence de la technique. Comme exemple de calcul de $F - score$, nous prenons le même exemple précédent (table 13), la mesure de $F - score$ nous donne une valeur de **0.6723**, qui est une mesure raisonnable par rapport à la segmentation obtenue.

Protocole de comparaison

Dans nos expérimentations, nous avons mené différentes manipulations en utilisant toutes les combinaisons issues de 9 techniques de normalisations, 4 espaces couleurs et 4 techniques de caractérisations, une combinaison se représente comme suit :

$$N_i E_j C_k \tag{1.18}$$

Avec N, E et C signifient respectivement Normalisation, Espace couleur et Caractérisation ($i = 9$ et $j = k = 4$), la codification des paramètre i,j et k est présenté dans le tableau 14. Pour plus de clarté, la combinaison $N_5E_2C_3$ correspond aux résultats de la segmentation avec la normalisation HEQ, espace couleur I1I2I3 et caractérisation par contraste.

Normalisation N		Espace couleur E		Caractérisation C	
i	Codification	j	Codification	k	Codification
$i = 1$	CGWN	$j = 1$	HSV	$k = 1$	CM
$i = 2$	CHROMA	$j = 2$	I1I2I3	$k = 2$	FOS
$i = 3$	CLAHE	$j = 3$	Lab	$k = 3$	Contraste
$i = 4$	GWN	$j = 4$	RGB	$k = 4$	Moment de Hu (Hu)
$i = 5$	HEQ				
$i = 6$	Lmax				
$i = 7$	MV				
$i = 8$	RGBcb				
$i = 9$	Sans Normalisation (SN)				

Table 14 – La codification de i,j et k dans une combinaison.

Nous disposons d'un nombre important de résultats issus de plusieurs combinaisons (144 résultats des 144 combinaisons possibles). Pour analyser ce genre de résultat, il est recommandé d'utiliser les tests non-paramétriques comme il a été expliqué par Janez Demsar [71]. L'objectif principal de ce type d'analyse est de savoir comment tirer des conclusions à partir d'un tableau de résultats, et par la suite identifier la segmentation pertinente [70].

Dans cette étude, nous nous concentrerons sur l'utilisation du test Friedman Aligned-Ranks (FAR) Post-Hoc [72], comme outil permettant de comparer le comportement de chaque approche. Son application nous permettra de mettre en évidence l'existence de différences significatives entre les résultats de segmentation et ainsi trouver l'approche la plus performante.

La procédure Post-hoc d'une comparaison multiple NxN [72] est un outil intéressant, son principal objectif est d'observer les différences de performance, en procédant à des comparaisons entre chaque paire d'approche. De cette façon, nous pouvons comparer tous les approches entre elles, et certainement à la fin, nous pouvons faire un classement de meilleures approches.

Le test Friedman est composé de trois étapes de calcul, la première étape consiste à convertir les résultats originaux en **Rangs**, l'approche la plus performante doit avoir le Rang 1, la deuxième Rang 2, etc. . . , pour une meilleure clarté un exemple est donné dans le tableau 15.

Le test de Friedman nécessite le calcul des rangs moyens de chaque approche, $R_j = \frac{1}{n}\sum_i r_i^j$, avec r_i^j le rang de j^{eme} approches et i^{eme} images. Si deux approches portent le même rang, cela signifie que les deux approches se comportent de la même façon. La

	$N_1E_1C_1$	Rang	$N_1E_1C_2$	Rang	$N_1E_1C_3$	Rang	$N_1E_1C_4$	Rang
Image 1	0.2831	(4)	0.9818	(2)	0.9821	(1)	0.6439	(3)
Image 2	0.8422	(4)	0.9806	(1.5)	0.9806	(1.5)	0.8961	(3)
Image 3	0.8225	(4)	0.8246	(2)	0.8241	(3)	0.8356	(1)
Image 4	0.9544	(3)	0.9812	(1)	0.9589	(2)	0.6405	(4)
Image 5	0.9289	(2)	0.9809	(1)	0.8091	(3)	0.5711	(4)
Rang moyen		3.4		1.5		2.1		3

Table 15 – Exemple des résultats de segmentation en fonction de $F - score$.

mesure de Friedman est distribuée selon z (Eq. 1.19), R_i et R_j sont les rangs moyens des deux approches à comparer.

$$z = \frac{R_i - R_j}{\sqrt{\frac{k(k+1)}{6n}}} \tag{1.19}$$

Nous présentons un exemple dans le tableau 15, qui compare 4 combinaisons appliquées à 5 images, la normalisation N_1 avec l'espace couleur E_1 appliqués sur les 4 techniques de caractérisations (C_1, C_2, C_3 et C_4). Les rangs moyens fournissent eux-mêmes une comparaison utile. A titre d'exemple, le test Friedman entre $N_1E_1C_1$ et $N_1E_1C_2$ du tableau 15 mesuré par l'équation 1.19 est calculé comme suit :

$$z = \frac{3.4 - 1.5}{\sqrt{\frac{4(4+1)}{6 \cdot 5}}} = 2.327015 \tag{1.20}$$

Nous nous basons aussi sur le paramètre p-value, qui donne plus de précision en terme de comparaison [72], permettant par la suite de confirmer qu'une hypothèse est nulle ou non (hypothèse nulle signifie que les résultats sont similaires, pas de différence de performance). Plus le paramètre p-value est élevé, plus la preuve est forte face à l'hypothèse nulle.

	Comparaison		Friedman	p-value
	$N_1E_1C_1$ vs.	$N_1E_1C_2$	2.327015	0.019964
Rang	3.4	1.5		
	$N_1E_1C_2$ vs.	$N_1E_1C_4$	1.837117	0.066193
Rang	1.5	3		
	$N_1E_1C_1$ vs.	$N_1E_1C_3$	1.592168	0.111347
Rang	3.4	2.1		
	$N_1E_1C_3$ vs.	$N_1E_1C_4$	1.10227	0.270344
Rang	2.1	3		
	$N_1E_1C_2$ vs.	$N_1E_1C_3$	0.734847	0.462433
Rang	1.5	2.1		
	$N_1E_1C_1$ vs.	$N_1E_1C_4$	0.489898	0.624206
Rang	3.4	3		

Table 16 – Exemple d'une comparaison non-paramétrique multiple.

Le tableau 16 présente les tests non-paramétriques des résultats de segmentation qui figure dans le tableau 15. Une comparaison entre deux techniques de segmentation qui donne une mesure élevée de p-value, signifie que ces deux techniques produisent des résultats similaires, nous pouvons remarquer cela entre $N_1E_1C_1$ et $N_1E_1C_4$ (p-value=0.624206). Pour mettre en évidence cette conclusion, nous pouvons observer les moyennes des Rangs qui présentent une forte similitude. Contrairement à la comparaison entre $N_1E_1C_1$ et $N_1E_1C_2$ qui donne une mesure de p-value=0.019964, ce qui prouve qu'il existe un large écart de performance entre ces deux techniques. Cet écart de performance signifie qu'une technique entre $N_1E_1C_1$ et $N_1E_1C_2$ produit les meilleurs résultats de segmentation, la moyenne du Rang nous indique que $N_1E_1C_2$ est la technique la plus performante

($Rang = 1.5$) et $N_1E_1C_1$ est la technique la moins performante ($Rang = 3.4$).

L'analyse de nos résultats suit le même protocole qui a été présenté dans l'exemple précédent, des comparaisons multiples seront effectuées afin de mettre en avant les meilleures performances.

L'organisation de la discussion de nos résultats est scindée en trois parties, dans chacune d'elle nous rependrons à la question de la pertinence de traitement appliqué. L'analyse de normalisation, l'analyse d'espace couleur et l'analyse des techniques de caractérisation seront étudiées séparément.

L'Analyse de l'effet de normalisation

La normalisation de couleur a-t-elle une influence sur la segmentation ?, c'est la question qui sera traitée dans cette partie du résultat. Pour ce faire, nous procédons à une comparaison multiple entre les différentes techniques de normalisation, mais avant cela, il est nécessaire d'observer la meilleure combinaison en fonction de la couleur E_j et la caractérisation C_k pour chaque technique de normalisation.

Pour chaque normalisation N_i, des comparaisons multiples ont été réalisées en fonction de E_j et C_k afin de mettre en avant les meilleures performances de chaque technique de normalisation. Les tableaux 17 et 18 indiquent la moyenne du Rang de chaque normalisation N_i en fonction de E_j et C_k, les meilleurs classements sont affichés en caractère **gras**.

Dans le tableau 17, les résultats de la segmentation en utilisant la normalisation N_1, la combinaison $N_1E_4C_2$ (normalisation CGWN, espace couleur RGB et caractérisation FOS) est classée au premier Rang pour la segmentation du *noyau*, car elle présente la valeur minimale des moyennes des Rangs. Pour la segmentation du *cytoplasme* (tableau 18), la combinaison $N_1E_4C_2$ est classé aussi au premier Rang dans la normalisation N_1.

Afin de confirmer les observations du classement obtenu dans les tableaux 17 et 18, nous procédons à des comparaisons multiples entre E_j et C_k pour chaque normalisation N_i. Les mesures de Friedman et de $p - value$ sont affichées dans les tableaux 34, 36, 38, 40, 42, 44, 46, 48 et 50 (Annexe A) pour la segmentation du *noyau*, et les tableaux 35, 37, 39, 41, 43, 45, 47, 49 et 51 (Annexe A) pour la segmentation du *cytoplasme*.

	Classement								
Combinaison	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	N_9
E_1C_1	9.7414	13.1379	8.0172	11.4483	8.8448	13.5345	**3.2471**	13.908	13.6322
E_1C_2	**2.9425**	**4.2644**	**5.0862**	**4.6609**	9.023	**4.4598**	8.9425	10.5345	**4.6667**
E_1C_3	7.0977	9.2299	11.2126	11.4368	8.5977	11.7126	11.8966	6.1667	10.7874
E_1C_4	8.5172	14.6437	10.3103	10.8736	8.4195	10.7299	7.2471	10.6379	11.1897
E_2C_1	12.7126	**5.1207**	**5.6552**	9.8506	**8.0057**	7.1437	8.2529	7.3736	**5.8333**
E_2C_2	15.362	15.954	14.6839	15.3391	8.8218	15.2184	13.2069	8.1552	15.3391
E_2C_3	15.3621	11.4023	14.6379	15.3391	8.4828	15.2184	13.1897	**5.5345**	15.3391
E_2C_4	8.1667	**5.2759**	10.9253	**4.6667**	8.569	5.7356	**2.3276**	**3.6782**	**5.2184**
E_3C_1	11.023	9.2184	11.4253	5.8678	8.3161	6.5805	**2.9943**	13.3218	**5.9598**
E_3C_2	**3.546**	**4.908**	6.2759	**4.4253**	8.6724	6.6149	8.0057	**4.5805**	**5.9253**
E_3C_3	7.1609	9.2069	8.3391	8.0805	8.7644	9.408	12.4598	6.8161	8.8851
E_3C_4	10.2356	11.0575	**5.8448**	13.2644	**7.908**	8.8391	5.8333	**5.477**	12.2069
E_4C_1	**3.1207**	**3.6264**	**4.3161**	5.5575	**7.7356**	**5.3391**	8.5057	8.7529	**5.2356**
E_4C_2	**2.6839**	**3.2989**	6	5.2816	8.7471	**5.2816**	13.2069	15.1839	**5.2241**
E_4C_3	**4.6034**	**5.6494**	6.9828	**4.6437**	8.3908	**5.0517**	12.5517	6.1207	**5.2299**
E_4C_4	13.7241	10.0057	6.2874	5.2644	8.7011	**5.1322**	**4.1322**	9.7586	**5.3276**

Table 17 – Classement de Friedman selon la segmentation du *noyau* en fonction de chaque normalisation N_i.

L'analyse non-paramétrique de la normalisation N_1 dans les tableaux 34 et 35 (Annexe A) montrent clairement la variabilité des performances. Dans la segmentation du *noyau*, les comparaisons entre (E_4C_2 vs. E_2C_2), (E_4C_2 vs. E_2C_3), (E_1C_2 vs. E_2C_2) et (E_1C_2 vs. E_2C_3) présentent des estimations importantes dans l'écart de performance de segmentation ($Friedman > 17$). En effet, les combinaisons E_4C_2 et E_1C_2 peuvent conduire à de

	Classement								
Combinaison	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	N_9
E_1C_1	13.9253	13.9023	9.4253	14.3391	8.2644	14.5575	9.3391	13.2184	13.7011
E_1C_2	**2.908**	**2.6724**	5.5287	**3.5172**	8.7816	**2.7241**	8.6609	13.2184	**2.7241**
E_1C_3	6.1954	6.2874	7.6207	7.6379	9.1264	6.6264	6.7874	5.3276	6.7989
E_1C_4	9.6609	12.7356	10.2471	13.8046	7.6264	8.6897	10.1322	7.3793	11.0517
E_2C_1	8.8276	9.569	8.8218	10.0287	8.2011	10.3448	9.2931	6.9598	9.0517
E_2C_2	13.1379	14.2356	13.523	14.0287	9.5402	14.5345	10.6322	5.7069	14.046
E_2C_3	12.2586	14.2356	12.7126	9.2126	9.2874	14.454	6.5575	7.1437	13.8448
E_2C_4	12.2874	9.5172	7.2586	9.2069	7.431	6.7184	8.2874	6.4943	7.1897
E_3C_1	9.1839	8.7644	10.0172	7.046	8.0115	8.7759	9.3448	13.2184	9.2989
E_3C_2	**2.9138**	**3.3506**	**3.7356**	**2.9598**	8.3333	**2.8678**	8.4138	**2.9655**	**3.5172**
E_3C_3	5.6494	5.5115	7.3276	6.431	9.0747	5.6494	**4.523**	**4.3621**	6.3966
E_3C_4	11.8563	9.3851	10.954	12.7414	7.2069	11.2011	10.8276	7.5517	13.5862
E_4C_1	5.8851	6.477	7.9425	5.1207	8.5057	9.1322	9.454	7.7069	6.1839
E_4C_2	**2.0977**	**2.5862**	**4.8103**	**3.408**	8.7759	**3.3966**	10.4598	13.2184	**3.6954**
E_4C_3	5.2874	5.4598	6.9253	6.1667	9.2011	6.2471	7.8391	8.7759	6.0115
E_4C_4	13.9253	11.3103	9.1494	10.3506	8.6322	10.0805	**5.4483**	12.7529	8.9023

Table 18 – Classement de Friedman selon la segmentation du *cytoplasme* en fonction de chaque normalisation N_i.

meilleurs résultats de segmentation du *noyau* en utilisant la normalisation N_1. Dans le tableau 35 (Annexe A), les meilleures performances de la segmentation du *cytoplasme* ont été obtenues par les combinaisons E_4C_2 et E_1C_2. Ces résultats confirment le classement de $Friedman$ observé dans les tableaux 17 et 18.

Les tableaux 42 et 43 (Annexe A) résument les tests statistiques de $Friedman$ et de $p-value$, concernant les résultats de segmentation par la normalisation N_5. Les mesures statistiques estiment qu'il y a une similarité de performance entre toutes les combinaisons E_j et C_k, car avec cette normalisation les meilleurs classements indiquent un $p-value$ élevé par rapport aux autres techniques de normalisation.

D'après cette analyse statistique, nous pouvons sélectionner pour chaque technique de normalisation N_i la meilleure combinaison en fonction de E_j et de C_k celle qui permet de conduire à de meilleurs résultats de segmentation. Cette sélection est résumée dans le tableau 19 qui se base sur l'analyse non-paramétrique précédente, un classement de **Rang 1** jusqu'au **Rang 6** a été effectué pour chaque N_i.

Après avoir sélectionné les meilleures combinaisons de E_j et de C_k pour chaque normalisation N_i, nous effectuons une comparaison entre les techniques de normalisation pour discuter l'influence de la normalisation des images sur la segmentation. Nous procédons à une comparaison non paramétrique, en se basant sur l'étude précédente et nous mettons en comparaison uniquement la liste du **Rang 1** de tableau 19.

Les tableaux 20 et 21 montrent la comparaison d'un point de vue statistique entre toutes les techniques de normalisation. Le tableau 20 résume les tests statistiques de la segmentation du *noyau*, la normalisation N_1 (CGWN) est classée en premier position comme la normalisation la plus performante pour la segmentation du *noyau*, en deuxième position la normalisation N_8 (RGBcb) et en troisième position la normalisation N_2 (CHROMA). Une mesure élevée de $p-value$ est observée concernant les normalisations N_8 et N_2 ce qui signifie que les normalisations N_1, N_8 et N_2 peuvent conduire à des résultats similaires. Dans le même classement, la normalisation N_9 se trouve à la quatrième position, elle représente les performances d'une segmentation **sans normalisation**, un important écart de performance ($p-value = 0.000004$) est observé par rapport à la normalisation N_1. A partir de ces observations, on remarque que la normalisation N_1 a apporté des améliorations considérables pour la segmentation du *noyau*.

Les tests non-paramétriques de la segmentation du *cytoplasme* sont affichés dans le tableau 21. De même, la segmentation du *noyau* avec la normalisation N_1 est classée en première position. La normalisation N_9 est classée en troisième position qui représente les performances d'une segmentation **sans normalisation**. Donc, nous considérons que la normalisation N_1 comme une technique intéressante à exploiter pour améliorer les

résultats de segmentation du *noyau* et du *cytoplasme*.

	Classement	Rang1	Rang2	Rang3	Rang4	Rang5	Rang6
N_1	Noyau	E_4C_2	E_1C_2	E_4C_1	E_3C_2	_	_
	Cytoplasme	E_4C_2	E_1C_2	E_3C_2	_	_	_
N_2	Noyau	E_4C_2	E_4C_1	E_1C_2	E_3C_2	E_2C_1	E_4C_3
	Cytoplasme	E_4C_2	E_1C_2	E_3C_2	_	_	_
N_3	Noyau	E_4C_1	E_1C_2	E_2C_1	E_3C_4	E_4C_2	_
	Cytoplasme	E_3C_2	E_4C_2	E_1C_2			_
N_4	Noyau	E_3C_2	E_4C_3	E_1C_2	E_2C_4	E_4C_4	_
	Cytoplasme	E_3C_2	E_4C_2	E_1C_2	_	_	_
N_5	Noyau	E_4C_1	E_3C_4	E_2C_1	_	_	_
	Cytoplasme	E_3C_4	E_2C_4	E_1C_4	_	_	_
N_6	Noyau	E_1C_2	E_4C_3	E_4C_4	E_4C_2	E_4C_1	_
	Cytoplasme	E_1C_2	E_3C_2	E_4C_2	E_3C_3		_
N_7	Noyau	E_2C_4	E_3C_1	E_1C_1	_	_	_
	Cytoplasme	E_3C_3	E_4C_4	_	_	_	_
N_8	Noyau	E_2C_4	E_3C_2	E_3C_4	E_2C_3	E_4C_3	E_1C_3
	Cytoplasme	E_3C_2	E_3C_3	_	_	_	_
N_9	Noyau	E_1C_2	E_2C_4	E_4C_2	E_4C_3	E_4C_1	
	Cytoplasme	E_1C_2	E_3C_2	E_4C_2	_	_	_

Table 19 – Synthèse des meilleures combinaisons E_jC_k pour chaque technique de normalisation N_i.

En dépit de ces différences observées entre la segmentation avec normalisation et celle sans normalisation, l'analyse non-paramétrique montre que la normalisation de couleur peut influencer sur la qualité de segmentation. De ce fait, la normalisation de type **CGWN** est la technique la plus performante, c'est celle qui va être exploitée dans la segmentation du *noyau* et du *cytoplasme*.

Classement	Comparaison	$Friedman$	p-value
1	$N_1E_4C_2$		
2	$N_8E_2C_4$	0.415227	0.677975
3	$N_2E_4C_2$	0.775091	0.438286
4	$N_9E_1C_2$	4.622865	0.000004
5	$N_6E_1C_2$	4.816638	0.000001
6	$N_4E_3C_2$	5.204183	0
7	$N_3E_4C_1$	8.429116	0
8	$N_7E_2C_4$	13.024299	0
9	$N_5E_4C_1$	14.782095	0

Table 20 – Classement des meilleures techniques de normalisation pour la segmentation du *noyau*.

Classement	Comparaison	$Friedman$	p-value
1	$N_1E_4C_2$		
2	$N_6E_1C_2$	0.179932	0.857206
3	$N_9E_1C_2$	0.193773	0.846354
4	$N_2E_4C_2$	1.439455	0.150022
5	$N_4E_3C_2$	2.629774	0.008544
6	$N_8E_3C_2$	5.107297	0
7	$N_3E_3C_2$	8.069252	0
8	$N_7E_3C_3$	10.283799	0
9	$N_5E_3C_4$	12.830527	0

Table 21 – Classement des meilleures techniques de normalisation pour la segmentation du *cytoplasme*.

Analyse de couleur

Dans cette seconde partie de nos expérimentations, nous complétons notre étude par une analyse en fonction de chaque espace couleur E_j. Nous nous intéressons essentiellement à quantifier l'influence de la composante couleur sur la qualité de la segmentation. Nous appliquons toutes les combinaisons possibles entre les normalisations N_i et les techniques de caractérisation C_k, et pour chaque espace couleur E_j séparément afin d'effectuer une comparaison multiple non-paramétrique.

Les tests statistiques sont présentées dans les tableaux 52, 54, 56 et 58 (Annexe B) pour la segmentation du *noyau*, et dans les tableaux 53, 55, 57 et 59 (Annexe B) pour la segmentation du *cytoplasme*. A partir de ces résultats, nous observons la meilleure combinaison entre N_i et C_k qui peut conduire à une bonne segmentation. Par exemple, dans l'espace couleur E_1, la normalisation N_1 combinée avec la caractérisation C_2 (tableau 52 (Annexe B)), c'est la combinaison qui donne les meilleurs résultats en terme de segmentation du *noyau* selon l'estimation de $Friedman = 14.36$.

Le bilan de cette analyse est résumé dans un tableau récapitulatif (tableau 22), un classement de la meilleure combinaison pour chaque espace couleur E_j a été mis en évidence. Une deuxième comparaison est effectuée en tenant compte uniquement du classement du **Rang 1**, dans l'objectif d'identifier le meilleur espace couleur E_j qui peut conduire à un meilleur résultat de classification super-pixellique.

La comparaison obtenue est présentée dans les tableaux 23 et 24. Dans la segmentation du *noyau* (tableau 23) les espaces E_4, E_1 et E_2 (RGB, HSV et I1I2I3, respectivement) donnent des résultats similaires, cela est justifiée par l'estimation de $Friedman$. En outre, l'espace E_3 (Lab) présente un écart important de performance en comparaison avec E_4 ($p - value = 0.00066$).

Concernant la segmentation du *cytoplasme* (tableau 24), la comparaison entre les espaces E_4 et E_1 a donné une estimation de $p - value = 0.79$, ce que signifie une forte similarité de performance. Par contre, les espaces E_3 et E_2 présentent un écart important de performance en comparaison avec E_4.

On conclu que l'espace E_4 (RGB) est classé en première position comme étant l'espace le plus performant dans la segmentation du *noyau* et du *cytoplasme*. De même, l'espace E_1 (HSV) est classé en deuxième position.

	Classement	Rang1	Rang2	Rang3	Rang4	Rang5	Rang6
E_1	Noyau	N_1C_2	N_2C_2	N_8C_3			
	Cytoplasme	N_6C_2	N_9C_2	N_1C_2	N_2C_2	—	—
E_2	Noyau	N_8C_4	N_2C_1	N_2C_4	N_8C_3	N_2C_2	N_4C_4
	Cytoplasme	N_2C_2	N_6C_4				
E_3	Noyau	N_8C_2	N_1C_2	N_2C_2	N_8C_4	N_4C_2	N_9C_2
	Cytoplasme	N_6C_2	N_1C_2	N_4C_2			
E_4	Noyau	N_1C_2	N_2C_2	N_1C_1	N_2C_1		
	Cytoplasme	N_1C_2	N_2C_2	N_6C_2		—	—

Table 22 – Synthèse des meilleures combinaisons N_iC_k pour chaque espace couleur E_j.

Classement	Comparaison	$Friedman$	p-value
1	$E_4N_1C_2$		
2	$E_1N_1C_2$	0.704664	0.481019
3	$E_2N_8C_4$	0.704664	0.481019
4	$E_3N_8C_2$	3.405877	0.00066

Table 23 – Classement des meilleurs espaces couleurs pour la segmentation du *noyau*.

Classement	Comparaison	$Friedman$	p-value
1	$E_4N_1C_2$		
2	$E_1N_6C_2$	0.264249	0.791588
3	$E_3N_6C_2$	2.084632	0.037103
4	$E_2N_2C_2$	9.043191	0

Table 24 – Classement des meilleurs espaces couleurs pour la segmentation du *cytoplasme*.

Analyse de caractérisation

La caractérisation super-pixellique est le traitement appliqué pour chaque super-pixel, permettant de traduire la propriété couleur en une caractéristique représentative capable d'être exploitée dans une hypothèse de classification. Plus la caractérisation est fiable plus la classification est bonne. De ce fait, il est impératif d'analyser attentivement chaque technique de caractérisation C_k.

Utilisant le même protocole de test, une analyse en fonction de la normalisation N_i et de la couleur E_j est effectuée pour chaque technique de caractérisation C_k séparément comme il'est représentée dans les tableaux 60, 62, 64 et 66 (Annexe C) pour la segmentation du *noyau*, et les tableaux 61, 63, 65 et 67 (Annexe C) pour la segmentation du *cytoplasme*. La synthèse de cette analyse est résumée dans le tableau 25.

Par la suite, nous nous intéressons dans notre étude aux résultats du **Rang 1**. Une deuxième analyse non-paramétrique est effectuée dans l'objectif de comparer les techniques de caractérisation. Les tableaux 26 et 27 indiquent les statistiques de $p - value$ en comparaison avec la technique de caractérisation la plus performante. Dans la segmentation du *noyau* (tableau 26), la caractérisation C_2 de la technique FOS est classée dans la première position, la deuxième position est réservée pour la caractérisation C_4 (Hu) avec un écart de $p - value = 0.21$, après vient le classement de C_1 et C_3 avec un écart de performance important ($p - value = 0.000152$).

Concernant la segmentation du *cytoplasme* (tableau 27), nous observons également que la caractérisation C_2 (FOS) est la technique la plus fiable dans l'obtention d'une bonne segmentation du *cytoplasme*, les caractérisations C_1 (CM) et C_3 (contraste) ont obtenu le même classement avec un écart de performance important ($p - value = 0.013$)$parrapport$C$_2$, la caractérisation C_4 (Hu) est classée en quatrième position.

	Classement	Rang1	Rang2	Rang3	Rang4	Rang5	Rang6
C_1	Noyau	N_2E_4	N_1E_4	N_2E_2			
	Cytoplasme	N_4E_4	N_4E_3	N_1E_4	N_2E_4	N_9E_4	
C_2	Noyau	N_1E_4	N_2E_4	N_8E_2	N_9E_4	N_6E_4	N_8E_1
	Cytoplasme	N_1E_4	N_6E_3	N_6E_4			
C_3	Noyau	N_1E_4	N_2E_4	N_8E_2	N_9E_4	N_6E_4	N_8E_1
	Cytoplasme	N_6E_3	N_1E_4	N_6E_4			
C_4	Noyau	N_8E_2	N_2E_2	N_8E_3			
	Cytoplasme	N_6E_2	N_9E_2	N_6E_1	N_3E_2		

Table 25 – Synthèse des meilleures combinaisons N_iE_j pour chaque technique de caractérisation C_k.

Classement	Comparaison	$Friedman$	p-value
1	$C_2N_1E_4$		
2	$C_4N_8E_2$	1.233162	0.217515
3.5	$C_1N_2E_4$	3.78757	0.000152
3.5	$C_3N_1E_4$	3.78757	0.000152

Table 26 – Classement des meilleures techniques de caractérisation pour la segmentation du *noyau*.

Classement	Comparaison	$Friedman$	p-value
1	$C_2N_1E_4$		
2.5	$C_1N_4E_4$	$2.46\overline{6}325$	$0.01\overline{3}651$
2.5	$C_3N_6E_3$	2.466325	0.013651
4	$C_4N_6E_2$	3.053545	0.002262

Table 27 – Classement des meilleures techniques de caractérisation pour la segmentation du *cytoplasme*.

Synthèse des expérimentions

Les expérimentations réalisées se sont basée essentiellement sur trois volets d'analyses : la normalisation couleur, la composante couleur et la caractérisation super-pixellique. Par le biais des tests non-paramétriques ($Friedman$ et $p - value$), nous avons réalisé des comparaisons entre toutes les combinaisons possibles de $N_iE_jC_k$, avec $i = 9$, $j = 4$ et $k = 4$. Les tests mesurés nous ont permis d'effectuer la sélection des meilleures combinaisons $N_iE_jC_k$, qui donnent les meilleures mesures de $F - score$ en matière de segmentation. Cela est synthétisé dans le tableau 28, les analyses de normalisation, de couleur et de caractérisation ont démontré que $N_1E_4C_2$ est la combinaison qui vient en première position comme combinaison la plus fiable pouvant d'être exploitée dans la segmentation du *noyau* et du *cytoplasme*.

Afin de confirmer cette observation, nous avons effectué une dernière comparaison statistique entre les combinaisons sélectionnées dans le tableau 28. Les mesures effectuées dans le tableau 29 confirment que $N_1E_4C_2$ (normalisation **CGWN**, espace couleur **RGB** et caractérisation **FOS**) est la combinaison la plus performante pour la segmentation.

	Classement	Rang1	Rang2	Rang3	Rang4	Rang5	Rang6	Rang7	Rang8	Rang9
Normalisation	*Noyau*	$N_1E_4C_2$	$N_8E_2C_4$	$N_2E_4C_2$	$N_9E_1C_2$	$N_6E_1C_2$	$N_4E_3C_2$	$N_3E_4C_1$	$N_7E_2C_4$	$N_5E_4C_1$
	Cytoplasme	$N_1E_4C_2$	$N_6E_1C_2$	$N_9E_1C_2$	$N_2E_4C_2$	$N_4E_3C_2$	$N_8E_3C_2$	$N_3E_3C_2$	$N_7E_3C_3$	$N_5E_3C_4$
Couleur	*Noyau*	$N_1E_4C_2$	$N_1E_1C_2$	$N_8E_2C_4$	$N_8E_3C_2$	–	–	–	–	–
	Cytoplasme	$N_1E_4C_2$	$N_6E_1C_2$	$N_6E_3C_2$	$N_2E_2C_2$	–	–	–	–	–
Caractérisation	*Noyau*	$N_1E_4C_2$	$N_8E_2C_4$	$N_2E_4C_1$	$N_1E_4C_3$	–	–	–	–	–
	Cytoplasme	$N_1E_4C_2$	$N_4E_4C_1$	$N_6E_3C_3$	$N_6E_2C_4$	–	–	–	–	–

Table 28 – Synthèse des meilleures combinaisons $N_iE_jC_k$ pour la segmentation du *noyau* et du *cytoplasme*.

	Classement	Comparaison	p-value	Normalisation	Couleur	Caractérisation
Noyau	1	N1E4C2		CGWN	RGB	FOS
	2	N1E1C2	$0.51\overline{8}317$	CGWN	HSV	FOS
	3	N8E2C4	0.347447	RGBcb	I1I2I3	Hu
	4	N2E4C1	0.022012	CHROMA	RGB	CM
Cytoplasme	1	N1E4C2		CGWN	RGB	FOS
	1	N6E1C2	$0.88\overline{3}286$	Lmax	HSV	FOS
	2	N1E4C1	0.000001	CGWN	RGB	CM
	3	N1E3C3	0	CGWN	Lab	Contraste

Table 29 – Classement de meilleur technique de normalisation, de caractérisation et couleur pour la segmentation du *noyau* et du *cytoplasme*.

7 Conclusion

Dans ce chapitre, nous avons procédé à de multiples expérimentations permettant une analyse détailler du concept de la classification super-pixellique, appliquées aux images cytologiques pour la reconnaissance automatique des globules blancs. Le but de ce travail est orienté vers la qualité de la caractérisation super-pixellique en exploitant l'information de la composante couleur. En effet, la normalisation couleur, la composante couleur et la caractérisation super-pixellique sont les principaux sujets de ce chapitre.

La normalisation couleur a-t-elle une influence sur la segmentation ? C'est la première question qui a été traitée dans la section 6.3, en tenant compte des résultats obtenus,

nous répondons par *"OUI"*. 8 techniques de normalisation ont été analysées et comparées avec les résultats sans normalisation, les résultats de normalisation **CGWN** a montré l'influence positive sur la segmentation du *noyau* et du *cytoplasme*.

Quel est le meilleur espace couleur ? L'espace couleur représente un choix important dans la segmentation des images couleurs. Nous avons étudié 4 espaces couleurs les plus cité dans la littérature (section 6.3). Dans l'analyse non-paramétrique effectuée, l'espace **RGB** a donné les meilleures performances de segmentation.

Quelle est le meilleur mode de caractérisation ? La fiabilité de la classification super-pixellique réside dans la fiabilité de la phase de caractérisation. Nos expérimentations se sont basées sur 4 techniques de caractérisation (CM, FOS, Contrast et Hu), la caractérisation **FOS** nous a donné des résultats de segmentation très proches de la segmentation de l'expert. De ce fait, la caractérisation **FOS** est la plus fiable pour représenter le super-pixel dans la classification.

CHAPITRE 2

Classification Super-pixellique

1 Objectif

La vision par ordinateur ou la vision artificielle est la discipline permettant la compréhension conceptuelle d'une scène désignée à partir d'informations image. Cette discipline devisé en différents concepts comme : suivi du mouvement, reconstruction, segmentation et reconnaissance [142].

Dans notre travail, nous traitons principalement le concept de la segmentation des globules blancs, ainsi que la reconnaissance du *noyau* et du *cytoplasme*. La reconnaissance et l'extraction des globules blancs dans le sang peuvent aider les hématologues à diagnostiquer les maladies comme le sida, la leucémie et le cancer du sang, ce qui font d'elles les étapes les plus importantes dans les procédures hématologiques. Cette analyse peut être effectuée par des approches manuelles et automatiques.

Les méthodes automatiques impliquent généralement des instruments tels que la cytométrie en flux (CMF) et les machines automatiques de comptage. Ces instruments peuvent examiner les globules blancs quantitativement mais pas qualitativement ; ils ne bénéficient pas des techniques de traitement d'image. L'application des systèmes automatiques qui incluent des techniques de traitement d'image peut fournir une évaluation qualitative et ainsi améliorer la segmentation. En outre, certaines de ces tâches telles que scruter manuellement les cellules sanguines par des experts sont fastidieuses et susceptibles d'erreurs. Par conséquent, un système automatique basé sur des techniques de traitement d'image peut aider les hématologues.

Au cours des dernières années, de nouveaux développements en termes d'extraction de variables [11], [12], [143], [144], [145], [146] et de nouveaux algorithmes d'apprentissage [11], [12], [147], [148], [149] ont apporté des améliorations considérables dans la précision de reconnaissance. L'intégration des techniques du super-pixel améliore significativement les performances de segmentation et de reconnaissance, cette intégration se base sur le concept de la classification super-pixellique (figure **??**), qui est le sujet de notre étude.

Le super-pixel est le traitement principal dans la conception du processus de classification super-pixellique. Un super-pixel peut être défini comme une région compacte de pixels similaires et connectés, qui représentent localement une même structure d'image. La mesure de similarité peut être définie de nombreuses façons, en utilisant l'intensité, la couleur, la texture et la position comme caractéristiques. Les pixels contenus dans le même super-pixel sont considérés égaux par définition, et ils ont des avantages par

rapport aux primitives de pixel simples, comme l'efficacité de calcul, puisque le nombre de primitives est fortement réduit au niveau du super-pixel. La production de super-pixel par l'algorithme *SLIC* est devenue de plus en plus populaire dans les applications de segmentation, cette production dépend de deux contraintes, la compacité m et le nombre de super-pixel k (figure 21).
Le succès d'un système de reconnaissance automatique dépend clairement de la qualité de chaque traitement compris dans le processus. Comme il est présenté dans la figure 7, une bonne représentation des variables (caractérisation), un bon paramétrage de super-pixel et une classification précise déterminent la qualité de notre système de reconnaissance.

La caractérisation super-pixellique a pour mission de représenter le super-pixel de l'image en vecteur de variable. L'importance de cette phase de traitement nous a obligée à réserver tout le chapitre 1 pour étudier attentivement la caractérisation super-pixellique. Dans ce présent chapitre, nous exploitons la caractérisation pertinente tirée des conclusions du chapitre 1.

Les paramètres super-pixelliques et la classification super-pixellique retiendront notre attention dans cette partie de la thèse. Concernant les paramètres super-pixelliques, la génération des super-pixels à partir de l'algorithme *SLIC* nécessite le paramétrage de la compacité m et du nombre de super-pixel k. Comme il est montré dans la figure 21, le choix des paramètres m et k influent sur la qualité et la quantité des super-pixels permettant ainsi une dégradation ou une amélioration des résultats de la segmentation.
Les techniques d'apprentissage deviennent de plus en plus pertinentes pour la vision par ordinateur. Pour contribuer à ce domaine, dans cette thèse nous nous concentrons principalement sur l'application de nouveaux procédés d'apprentissage automatique.
L'apprentissage supervisé est la référence fondamentale pour obtenir de meilleure performance de classification, notamment l'utilisation des techniques ensemblistes dans les systèmes de vision par ordinateur [11, 12]. Dans cette discipline, les échantillons d'apprentissage doivent être accordés par leurs étiquettes correspondantes fournies par un expert humain. Si les échantillons d'apprentissage sont suffisants, ces approches peuvent obtenir des performances très élevées de reconnaissance et de classification.

Dans nombreuses applications de vision par ordinateur [11,12], le classifieur de type forêt aléatoire est jugé comme le classifieur le plus performant dans ce domaine. Ce classifieur est basé sur le principe de l'union fait la force, c'est-à-dire, un ensemble de classifieur de type ***arbre de décision*** qui se réunit dans une forêt, et la concaténation de l'ensemble sera prise comme une décision finale.

En pratique, l'obtention de suffisamment d'images cytologiques étiquetées par un médecin est une tâche très fastidieuse, longue et coûteuse. Le manque de données marquées et les problèmes liés au processus d'étiquetage ont orienté les recherches vers l'étude de nouvelle procédure d'apprentissage basé sur le principe semi-supervisé (SSL).

Dans l'apprentissage semi supervisé, une quantité raisonnablement de données étiquetées peut-être prise en charge, et avec une certaine manière d'apprentissage, l'hypothèse apprise peut être renforcée, enrichie et rendue robuste en intégrant les données non étiquetées dans l'apprentissage.

Dans ce chapitre, nous allons vous présenter une comparaison entre la classification supervisée et la classification semi-supervisée, appliquée sur les super-pixels. En tenant compte des travaux de la littérature, nous avons décidé d'exploiter l'algorithme *Random Foreste* [9] en version supervisée et en version semi-supervisée (*co-Forest*). Afin d'exposer une validation complète de nos résultats, nous présentons une comparaison avec le mono classifieur en mode supervisé et en mode semi-supervisé.

Pour ce faire, ce chapitre est organisé comme suit : Dans la section 2, nous introduisons les travaux connexes concernant la segmentation classification pixellique et super-pixellique,

en mode supervisée et semi-supervisée. L'approche proposée sera détaillée dans la section 3, avec la présentation des techniques de classification abordées dans cette étude, en la technique du multi-classificeur / mono-classifieur et du supervisé / semi-supervisé. La section 4 présente les résultats expérimentaux. Et dans la section 5 nous concluons par l'apport de cette étude.

2 État de l'art du domaine

2.1 Classification pixellique / super-pixellique supervisée

En intégrant l'intelligence artificielle dans les mécanismes de segmentation on se permet de relever le défi d'obtenir une performance robuste et efficace, notamment si on exploite la puissance de l'apprentissage supervisé dans la classification pixellique / super-pixellique. Dans la littérature, un certain nombre de travaux ont apporté des contributions dans la classification pixellique / super-pixellique appliquèe à la segmentation automatique dans de différente modalité d'imagerie médicale IRM, Scanner, Rétinienne, Microscopique, etc. . .

Dans [150], Fuyong Xing et Lin Yang fournissent un récent état de l'art de la segmentation et la détection du *noyau / cellule* de différents types d'images microscopiques. 7 groupes de segmentation ont été classé dans [150], seuillage d'intensité, opération morphologique, ligne partage des eaux, modèle déformable, clustering, graphe et classification supervisée. En outre, la discussion des auteurs se focalise sur les méthodes d'actualité dans le domaine de segmentation, et les potentiels perspectives de détection / segmentation du *noyau / cellule*. Notamment la classification pixellique, qui présente un potentiel remarquable dans la segmentation, tandis que, ce genre de technique présente des limitations en terme de coût de calcul. Par conséquent, ces limites peuvent être résolues avec l'intégration du super-pixel dans la classification, ceci peut conduire à la conception d'un algorithme robuste de segmentation.

Une classification pixellique supervisée des images cytologiques a été pratiquée par Benazzouz et al. [151] pour la reconnaissance des globules blancs. Les auteurs ont établi la sélection des composantes couleurs par des techniques de sélection ReliefF, LDA, RFE et mRMR. Cette sélection a pour objectif de produire les caractéristiques pertinentes pour l'apprentissage artificiel, et la segmentation est exécutée par une classification en utilisant la méthode SVM (Support Vector Machine).

Sur les mêmes images cytologiques mais par une approche différente, Ismahan Baghli et al. [152] ont traité un nouveau cadre de segmentation basée sur la théorie d'évidences, appelés ESA (Evidential Segmentation Algorithm) pour la segmentation des globules blancs. L'algorithme proposé permet de résoudre le problème de la segmentation d'images de cellules sanguines. L'objectif est d'extraire les composants d'une image de cellule donnée en utilisant la théorie d'évidence, qui permet une plus grande flexibilité pour classer les pixels. Les résultats obtenus montrent l'efficacité de l'algorithme proposé par rapport à d'autres procédés concurrents.

Kong et al. [153] ont exploité le succès de l'apprentissage supervisé dans la segmentation du *noyau* des images histopathologiques par classification pixellique. En premier lieu, les auteurs ont proposé de passer de l'espace couleur RGB à l'espace couleur du type discriminant (MDC). En deuxième lieu, appliquer de la Transformée de Fourier Local (LFT) [154] pour l'extraction des caractéristiques, cette dernière présente une puissance discriminante pour la classification, à la fin, un classifieur supervisé de type KNN est mis en place pour procéder à la classification pixellige.

Yin Zhou et al. [155] ont proposé une approche nommée : SCCR (sparsity constrained convolutional regression), destinée à la segmentation des images histologiques. Une banque de filtre convolutionnel est utilisée pour l'extraction de variable, appliquée sur

chaque pixel de l'image. Par la suite, l'application d'une fonction de régression permet de générer une hypothèse capable de produire une classification pixellique.

Dans la classification pixellique, Zhaozheng Yin et al. [156] ont démontré l'efficacité de ce contexte dans la segmentation des cellules, exploitant l'information pixellique pour faire apprendre une hypothèse bayésienne, qui sera utilisée pour la classification pixellique par la suite.

Charisma (Cellular Histological Automated Robust Image Segmentation of Muscle Algorithm) [157] est un nouveau cadre de segmentation de cellules histologiques qui combine des techniques de traitement d'image, un classifieur supervisé et un nouvel algorithme robuste de clump splitting. La proposition combine l'efficacité du traitement d'image avec la capacité d'adaptation de l'intelligence artificielle. En outre, Thomas Janssens et al. [157] ont introduit une nouvelle technique de clump splitting, qui est basée sur l'analyse de concavité et la détection du chemin le plus court pour l'application de la classification pixellique. Les auteurs ont démontré également que l'amélioration de la performance de segmentation se traduit par une meilleure représentation de la distribution des propriétés cellulaires.

Cheng et al. [158] ont proposé une classification super-pixellique par le classifieur SVM, les super-pixels ont été générés par un algorithme appelé Turbopixels [159]. Une extraction de variable pour chaque super-pixel est appliquée à base de texture et de forme, afin de préparer les données d'apprentissage et de classification.

Dans le contexte super-pixel, Shiyong Ji et al. [160] ont proposé une nouvelle approche de segmentation, mettant en combinaison le super-pixel et le fuzzy clustering. La proposition est notée MSFCM (multi-stage segmentation method based on super-pixel and fuzzy clustering), destinée à la segmentation des tumeurs dans les images cérébrales. L'image IRM est traitée en super-pixel à savoir des sous-régions homogènes dans l'image par l'application de Turbopixels [159]. Par la suite, la mise en exécution d'une règle floue pour mesurer le degré d'appartenance de chaque super-pixel, et une méthode de diffusion itérative basée sur la fonction Butterworth [160] sont utilisées pour définir la classe de chaque super-pixel, l'image segmentée est obtenue en fusionnant les super-pixels qui ont la même étiquette de classification.

De même, Wei Wu et al. [161] sont se reposés sur les avantages du super-pixel dans la segmentation des images cérébrales. Les auteurs ont adopté un nouvel algorithme basé sur la théorie des graphes, dans l'objectif de produire plus d'affinités dans les super-pixels cérébraux. La caractérisation super-pixellique est apportée à partir de la texture du supepixel, utilisant la caractérisation par l'ondelette de Gabor [162]. Concernant la classification super-pixellique, les auteurs ont fait appel au classifieur SVM pour cette action.

2.2 Classification pixellique / super-pixellique semi-supervisée

Les méthodes de classification pixellique sont fréquemment utilisées dans la segmentation d'images. Cependant, les méthodes de segmentation supervisées conduisent à une meilleure précision, cette dernière a besoin une richesse perceptible en termes de pixel étiqueté pour la phase d'apprentissage du classifieur. Obtenir un pixel étiqueté qui provient de la segmentation manuelle d'un expert du domaine est une action qui devient de plus en plus compliqué à cause du plusieurs raisons (coût, difficulté, compétence, etc. . .).

Une solution est apportée par l'utilisation de l'apprentissage semi-supervisé, qui utilise à la fois un ensemble de pixels étiquetés et non-étiquetés au cœur d'apprentissage, cette qualité peut conduire à un apprentissage efficace, permettant la classification et la segmentation avec une grande précision. Plusieurs travaux se sont intéressés à cette approche dans la segmentation des images médicales, nous pouvons citer quelques-uns.

La segmentation des images médicales échographiques est une problématique difficile à résoudre en raison de la présence du bruit d'atténuation du signal. Dans [164] Xbresson et al. présentent une méthode de segmentation intelligente basée sur le concept de la semi-supervision pour identifier les anomalies en imagerie échographique. Les auteurs ont proposé une initialisation des étiquettes assistée par un expert du domaine, suivi d'une utilisation du graphe de patchs d'image qui agit en tant que probabilités correctives d'intensité pour une bonne représentation de l'image ultra-sonore. Ensuite, ils ont formulé la segmentation comme un problème de réduction minimale continue et cela est résolu avec un algorithme d'optimisation de type « Continuous graph cut model ». Ce dernier est proposé comme un algorithme de segmentation qui vise à minimiser l'opérateur de coupe (ou norme à base de graphes) appliqué sur les objets d'intérêt. Ce travail est validé dans le cadre de segmentation des imageries cliniques échographiques (prostate, fœtus, tumeurs du foie et des yeux). Les résultats obtenus ont une grande similarité avec la vérité-terrain fourni par les délimitations d'experts médicaux dans toutes les applications.

Dans le domaine médical, Azmi et al. [165] ont adapté les approches semi-supervisées pour une segmentation par classification pixellique dans le but d'aide au diagnostic. Plus précisément, ils ont proposé une nouvelle approche améliorée de classification semi-supervisée interactive de type auto-apprentissage pour la segmentation des lésions suspectes en IRM du sein. L'idée de ce travail est de faire intervenir en premier lieu un radiologue expert pour une segmentation semi-complète de la région d'intérêt (sélection de 20% de la région pour quelques images). En second lieu, une caractérisation de texture pour chaque pixel de l'image est faite en utilisant trois catégories : matrice de co-occurrence, variable statistique d'histogramme et longueur d'exécution matricielle. Avant d'aborder la partie apprentissage semi-supervisée, ils ont proposé un seuillage des images d'apprentissage pour faire générer les pixels de potentiel.

Un auto-apprentissage est appliqué par la suite par un classifieur bayésien entraîné sur les pixels expertisés, une mesure de confiance probabiliste est appliqué pour améliorer la robustesse du classifieur, permettant par la suite de segmenter les lésions avec précision. En terme de performances, les auteurs ont démontré que la segmentation semi-supervisée est plus fiable que la segmentation supervisée (KNN, SVM et réseau bayésien) et non supervisée (FCM).

Par contre, ce que nous reprochons à ce travail est qu'il traite d'une manière non exhaustive la partie calcul de confiance, aucun détail n'est présenté sur la fonction de mesure de ce dernier, nous rappelons qu'en auto-apprentissage est basé essentiellement sur le ré-apprentissage du classifieur en utilisant les données nouvellement classées avec une confiance. Ces dernières sont sélectionnées après des mesures mathématiques adaptées telles que de la probabilité d'appartenance qui est utilisé en basic self training [166] ou celle de SETRED [6] qui utilise l'information du graphe de voisinage comme mesure de confiance.

Azmi et al. [167] reprennent l'idée de la segmentation semi supervisée mais cette fois sur les images (IRM) du tissu cérébral. Ils font appel à une méthode ensembliste en semi-supervisé pour la segmentation d'images IRM, par plusieurs classifieurs semi-supervisés simultanément. Ainsi, dans cet article, les auteurs utilisent deux algorithmes semi-supervisés : L'algorithme de filtrage espérance-maximisation et MCo_Training qui est une version améliorée de la méthode semi-supervisés Co_Training afin d'augmenter la précision de la segmentation. Les résultats expérimentaux montrent que la performance de segmentation dans cette approche est très élevée.

3 L'approche proposée

Dans le présent chapitre, nous abordons la question de la classification super-pixellique, qui consiste à former un classifieur permettant d'affecter une classe de région pour chaque super-pixel de l'image. En effet, la technique de classification est un choix crucial pour ce genre de technique de segmentation.

Il a été clairement constatée que, **le classifieur de Forêt Aléatoire (Random Forest) a conduit au meilleur résultat de classification notamment dans les applications types images**. Le concept se base sur des ensembles de classifieurs ou méthode ensembliste, qui est une des approches multi-classifieurs les plus populaires et les plus efficaces. Elle consiste à combiner les décisions individuelles de plusieurs hypothèses $h_1, \ldots, h_T$ pour classer de nouveaux exemples.

Afin de confirmer la conclusion qui dit : la méthode ensembliste Forêt Aléatoire est robuste en classification, nous proposons une étude comparative discutant les qualités d'apprentissage ensembliste RF en comparaison avec le mono classifieur arbre de décision. Cette comparaison est abordée en utilisant deux modes d'apprentissage, le supervisé et le semi-supervisé, dans l'objectif de démontrer d'une manière empirique l'efficacité des approches d'ensemble (de manière plus spécifique celle de l'algorithme des forêts Aléatoires) dans la segmentation des images. Ainsi, nous exposons la richesse du semi-supervision dans l'amélioration de l'apprentissage supervisé en exploitant les données non étiquetées. De ce fait, la comparaison est menée entre Forêt Aléatoire vs. Arbre de Décision en mode supervisé. Concernant le mode semi-supervisé, la comparaison est menée entre co-Forest vs. SETRED, le co-Forest est la version semi-supervisée de Forêt Aléatoire, et le SETRED est une technique d'auto-apprentissage considérée comme mono-classifieur en mode semi-supervisé.

L'objectif principal est la segmentation et la reconnaissance efficace des globules blancs dans la cytologie, la segmentation se base essentiellement sur la classification super-pixellique, comme il est représenté dans la figure 23. Une première étape consiste dans la production des super-pixels par l'algorithme SLIC, par la suite en exécutant la caractérisation super-pixellique, ainsi un apprentissage et une classification seront appliqués sur la base de caractéristiques préparées. L'étude réalisée dans le chapitre 1 nous a permis de sélectionner les caractéristiques les plus pertinentes pour notre application, qui se résume dans la normalisation CGWN, l'espace couleur RGB et la caractérisation FOS. Notre contribution consiste à réaliser une meilleure classification appliquée à la caractérisation effectuée pour une segmentation plus fiable.

3.1 Méthodes de classification

Avec la disponibilité des données non étiquetées et la difficulté d'obtenir leurs étiquettes, des méthodes d'apprentissage semi-supervisé ont acquis une grande importance. A la différence de l'apprentissage supervisé, l'apprentissage semi supervisé vise les problèmes avec relativement peu de données étiquetées et une grande quantité de données non étiquetées. La question qui se pose est alors de savoir si la seule connaissance des points avec labels est suffisante pour construire une fonction de décision capable de prédire correctement les étiquettes des points non étiquetés. Différentes approches proposent de déduire des points non étiquetés des informations supplémentaires et de les inclure dans le problème d'apprentissage.

En effet, la combinaison de différentes méthodes d'apprentissage/classification permet de tirer avantage de leurs forces tout en contournant leurs faiblesses. Aujourd'hui, force est de constater que ce qu'on appelle maintenant systèmes multi-classifieurs (désignés souvent par l'acronyme MCS pour Multiple Classifier Systems) constitue une des voies les plus prometteuses de l'apprentissage automatique [168].

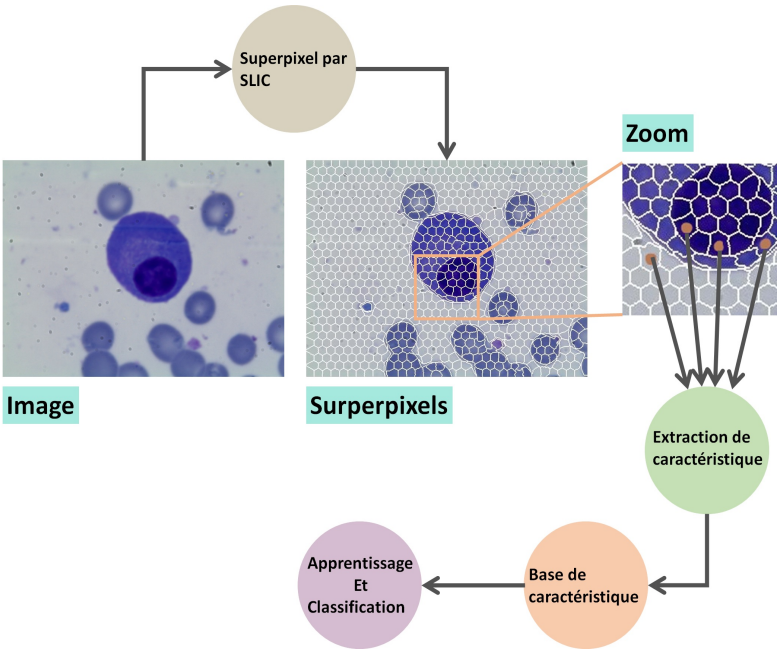


Figure 23 – Processus de segmentation par classification super-pixellique.

Les MCS ont fait l'objet de nombreux travaux et il existe aujourd'hui un grand nombre de méthodes capables de générer automatiquement des ensembles de classifieurs : Bagging [169], Boosting [170], Random Subspaces [171], ECOC [172], pour ne citer que les plus courantes, sont autant de méthodes différentes dont l'objectif commun est de créer de la diversité au sein d'un ensemble de classifieurs performants tout en cherchant à établir le meilleur consensus possible entre ces classifieurs.

De ce fait, nous proposons dans ce travail d'appliquer les systèmes multi-classifieur de type méthode d'ensemble en comparaison avec les méthodes mono-classifieurs en apprentissage supervisé et semi-supervisé : Arbre de décision, forêt aléatoire, auto-apprentissage SETRED, et forêts en apprentissage semi supervisé *co-Forest*.

Arbre de décision

Les arbres de décision représentent une méthode très efficace dans l'apprentissage supervisé. Il s'agit de partitionner un ensemble de données en des groupes les plus homogènes possible du point de vue de la variable à prédire. On prend en entrée un ensemble de données classées, et on fournit en sortie un arbre qui ressemble beaucoup à un diagramme d'orientation. Un arbre de décision est composé d'une racine qui est le point de départ de l'arbre, des nœuds et des branches qui relient : la racine avec les nœuds, les nœuds entre eux et les nœuds avec les feuilles.

Il existe plusieurs algorithmes dans la littérature CART [173], ID3 [174] et C4.5 [175], dans ce travail nous nous limiterons à l'application de l'algorithme CART (Classification and Regression Tree)

Méthode d'auto-apprentissage "SETRED"

SETRED (self-training with data editing) est un algorithme d'auto-apprentissage proposé par Li and Zhou [6]. Les auteurs ont étudié le potentiel des techniques de filtrage des données comme une mesure de confiance, cela permet de diminuer le risque d'ajouter des données bruitées à la base d'apprentissage. Cet algorithme utilise les techniques d'édition des données pour identifier et supprimer les données mal classées. Dans le processus

d'auto-apprentissage, à chaque itération l'algorithme fait appel au principe de la règle du plus proche voisin et au principe de coupure des arêtes pondérée pour mesurer la confiance.

Le graphe de voisinage est un outil issu de la géométrie computationnelle qui a été exploité dans de nombreuses applications d'apprentissage automatique. Par définition un graphe de voisinage $G = (S, E)$ [58] est associé à un ensemble de données dont « S » sont les sommets et « E » sont les arêtes.

Muhlenbach et al. [4] ont exploité l'information des arêtes pour calculer un poids statistique afin de couper les arêtes des données connexes de différentes classes.

En premier lieu, une hypothèse est apprise sur les données étiquetées. Suivi d'une application du processus d'auto-apprentissage qui fait appel à l'algorithme de Cut Edge Weight Statistic (CEWS) [4] pour calculer le rapport R_i 2.1 (chapitre 2 section 3.1). Pour juger si la donnée est bien classée, le rapport R_i doit être supérieur à un seuil qui est fixé par l'utilisateur.

Méthode d'ensemble : Forêts Aléatoires

Une Forêt Aléatoire est un prédicteur constitué d'un ensemble de classifieurs élémentaires de type arbres de décision. Dans les cas spécifiques des modèles CART (arbres binaires), Breiman [9] propose une amélioration du bagging avec un algorithme d'induction de forêts aléatoires (Forest-RI — pour Random Forest - Random Input —) qui utilise le principe de randomisation "Random Feature Selection" proposé par Amit et Geman [176]. L'induction des arbres se fait sans élagage et selon l'algorithme CART [173], toutefois, au niveau de chaque nœud, la sélection de la meilleure partition, basée sur l'index de Gini, s'effectue uniquement sur un sous-ensemble d'attributs de taille préfixée (généralement égale à la racine carrée du nombre total d'attributs) sélectionné aléatoirement depuis l'espace original des caractéristiques [177]. La prédiction globale de la forêt aléatoire est calculé en prenant la majorité des votes de chacun de ses arbres. Cet algorithme appartient à la famille la plus large des forets aléatoires définis comme suit par Breiman [9].

Algorithm 5 Pseudo code de l'algorithme des forêts aléatoires

Entrée : L'ensemble d'apprentissage L, Nombre d'arbres N.
Sortie : $Ensemble\ d'arbres\ E$
Processus :
for $i = 1 \rightarrow N$ **do**
$T^i \leftarrow BootstrapSample(L)$
$C^i \leftarrow ConstructTree(T^i)$ où à chaque noeud :
- Sélection aléatoire de $K = \sqrt{M}$ Variables à partir de l'ensemble d'attributs M
- Sélection de la variable la plus informatif K en utilisant l'index de Gini
- Création d'un nœud fils en utilisant cette variable

$E \leftarrow E \cup \{C^i\}$
end for
Retourner E

Méthode des Forêts Aléatoires en apprentissage semi supervisé "*co-Forest*"

co-Forest a été proposé par Li et Zhou [20] il étend le paradigme *co-Training* [16] par la méthode d'ensemble *Random Forest* [9].

Nous désignons par L et U l'ensemble des données étiquetées et non étiquetées respectivement. En *co-Training*, deux classifieurs sont formés à partir de L, puis chacun d'eux sélectionne les exemples les plus confiants dans U pour la phase de labellisation, à partir de leur propre fonction de classement ou par séparation hyperplan. Ainsi, une partie importante de *co-Training* réside dans la façon d'estimer la confiance de la prévision, en

d'autres termes, *comment estimer ou obtenir la confiance d'un exemple non marqué*.

Dans *co-Forest*, un ensemble de N classifieurs désignés comme H^* est utilisé dans *co-Training* au lieu de deux classifieurs. De cette façon, nous pouvons estimer efficacement la confiance de chaque classifieur. Si nous voulons considérer l'exemple labellisé le plus confiant par un des classifieurs h_i $(i = 1, 2, \ldots, N)$ de l'ensemble H^*, nous employons tous les autres classifieurs à l'exception de h_i, appelé l'ensemble concomitance de h_i et notée par H_i. Par conséquent, la confiance de l'étiquetage peut être calculée comme le degré d'accords sur l'étiquetage, c'est à dire le nombre de classifieurs qui s'accordent sur le label attribué par H_i. L'idée générale de *co-Forest* est de former tout d'abord un ensemble des classifieurs de données étiquetées L et affiner chaque classifieur avec des données non marquées par son ensemble de concomitance.

Plus précisément, dans chaque itération d'apprentissage autour de *co-Forest*, l'ensemble de concomitance H_i permettra de tester chaque exemple dans U. Pour un exemple non labellisé xu, si le nombre de classifieurs qui s'accordent sur une étiquette particulière dépasse un seuil θ prédéfini, cette nouvelle étiquette lui sera affectée et par la suite il sera copié dans le nouveau ensemble L'. A l'itération suivante, L' est employé pour le raffinage de h_i. Delà, on notera que les exemples non étiquetés ne sont pas supprimés de U, de sorte qu'ils peuvent être sélectionnés par d'autres $H_j(ji)$ dans les itérations suivantes.

Un problème qui peut affecter la performance globale de *co-Forest* est que toutes les données non étiquetées dont la confiance est au-dessus de θ seront ajoutées à L_i, ce qui rend L_i assez grand dans l'avenir. Mais dans le cas où un classifieur ne peut représenter la distribution sous-jacente ; une énorme quantité de données étiquetées deviendra nuisible à la performance, au lieu d'aider la précision de la prédiction.

Ce phénomène a été découvert dans plusieurs algorithmes d'apprentissage semi- supervisé. Inspiré par Nigam et al. [178], *co-Forest* reprend aussi le principe en assignant un poids à chaque exemple non marqué. Un exemple est pondéré en fonction de la confiance prédictive d'un ensemble concomitant. Cette approche permet de réduire l'influence de θ, même si θ est petit, les exemples qui ont une faible confiance prédictive peuvent être limités.

$$\frac{\hat{e}_{i,t}}{\hat{e}_{i,t-1}} < \frac{w_{i,t-1}}{w_{i,t}} < 1 \tag{2.1}$$

Selon Li et Zhou [20], dans les deux itérations ($(t-1)$ avec $t > 1$), la condition (2.1) pour l'actualisation de l'ensemble d'apprentissage devrait être satisfaite pour améliorer de manière itérative la capacité de généralisation.

En supposant que le classificateur h_i est reconstruit (phase ré-apprentissage) sur l'ensemble de données $L_i U L'_{i,t}$ dans la t i ème l'itération. Et que $\hat{e}_{i,t}$ désigne le taux d'erreur de H^* sur $L'_{i,t}$ aussi que $\hat{e}_{i,t} w_{i,t}$ est la moyenne pondérée des exemples mal labellisés par H^*.

Les hypothèses que $\hat{e}_{i,t} < \hat{e}_{i,t-1}$ et $w_{i,t-1} < w_{i,t}$, $w_{i,t} < \frac{\hat{e}_{i,t-1} w_{i,t-1}}{\hat{e}_{i,t}}$ devront être réunis par l'obtention du sous-échantillon de U et ainsi satisfaire la condition en (2.1).
En résumé, le principe de l'algorithme *co-Forest* (figure 24), consiste en N arbres aléatoires qui sont d'abord formés à partir d'un ensemble d'apprentissage bootstrap [1] de l'ensemble labellisé L pour créer une forêt aléatoire. Ensuite, à chaque itération, chaque arbre aléatoire sera affiné avec les exemples nouvellement marqués par son ensemble de concomitance, où la confiance de l'exemple labellisé dépasse un certain seuil θ. Cette méthode permettra de réduire les chances qu'un arbre dans une forêt aléatoire soit biaisé lorsque nous utilisons les données non étiquetées.

1. Un échantillon bootstrap L est, par exemple, obtenu en tirant aléatoirement n observations avec remise dans l'échantillon d'apprentissage L_n, chaque observation ayant une probabilité 1/n d'être tirée.

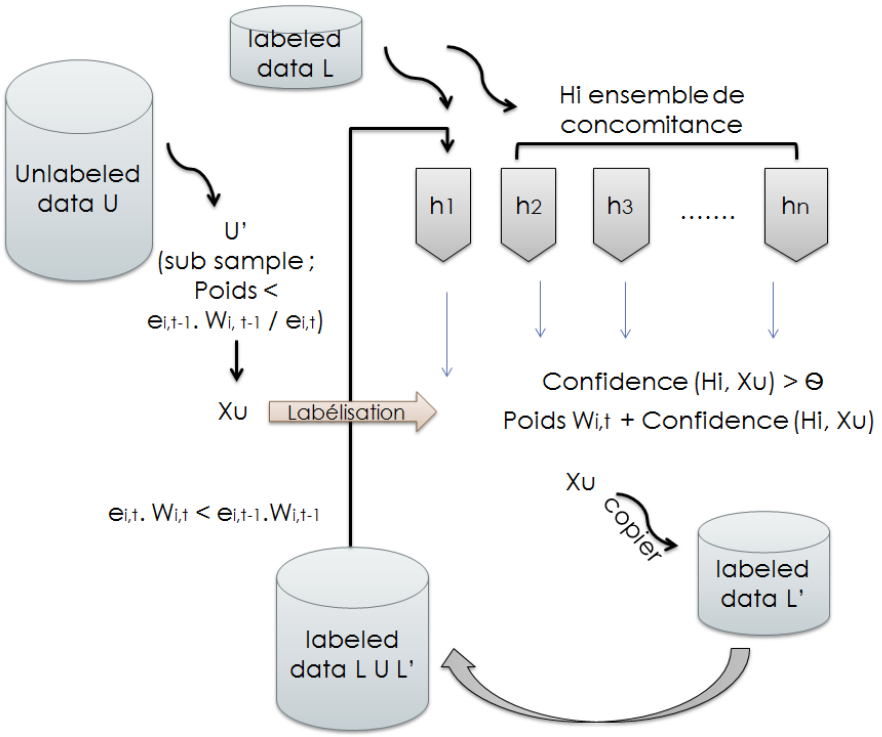


Figure 24 – Schéma représentatif du principe global du fonctionnement de *co-Forest*

Pour une description plus détaillée de l'algorithme *co-Forest*, nous référons le lecteur vers les articles [20] et [179] pour plus de détails.

4 Résultats et expérimentations

4.1 Base de données

La base de données des images cytologiques a été construite à partir d'images réelles acquises au sein du service d'hémobiologie (CHU Tlemcen), sur des lames la coloration de type MGG (May Grunwald Giemsa). L'environnement LEICA (caméra et microscope) permet d'obtenir des images couleur RGB de taille 768x1024. 87 images cytologiques avec vérité terrain ont été mises en place pour aborder l'étude expérimentale.

4.2 Expérimentations

La première étape dans nos expérimentations est la génération du super-pixel, donc on fait appel à l'algorithme SLIC pour ce faire. Cette génération nécessite un certain paramétrage de la compacité m et le nombre de super-pixel k, les paramètres m et k produisent des propriétés différentes en terme de super-pixels, comme il est affiché dans la figure 25. Une question qui se pose concernant l'influence de la qualité etla quantité du super-pixel dans la classification super-pixellique, afin de réspondre à cette question, nous teston la segmentation avec différents paramétrages de m et de k ($m = \{10, 15, 20\}$ et $k = \{100, 500, 1000, 1500, 2000\}$).

Une bonne classification nécessitera une efficacité dans la caractérisation et une robustesse dans l'apprentissage. La caractérisation super-pixellique a été étudiée attentivement dans le chapitre 1, et la configuration pertinente de caractérisation a été menée à l'utilisation de la normalisation CGWN, l'espace couleur RGB et la caractérisation FOS, qui sera de même utilisé dans cette partie d'expérimentation. Concernant la classification super-pixellique qui fait l'objet de cette étude, nous avons décidé d'utiliser les classifieurs RF, CART, co-FOREST et SETRED.

Par la suite de notre expérimentation, nous avons sélectionné 10% de la base de données (9 images) pour réaliser l'apprentissage. L'expert hématologiste interviendra dans l'étiquetage de ces 9 images par une sélection du super-pixel de chaque région d'intérêt (ROI), permettant ainsi une meilleure perception des régions *noyau*, *cytoplasme*, *globule rouge* et *fond*. Cette expertise est utilisée pour l'apprentissage supervisé et semi-supervisé. Concernant l'apprentissage de la forêt aléatoire supervisé et semi-supervisée, un nombre d'arbre égale à 100 a été choisi. L'évaluation a été réalisée par une validation croisée égale à 5. Les détails des paramètres d'expérimentation sont résumés dans le tableau 30.

Image étiquetée	9 images
Image non étiquetée	27 images
Image de test	51 images
Degré de confiance co-FOREST	75%
Nombre d'arbres RF / co-FOREST	100
Degré de confiance SETRED	75%
Validation croisée	5

Table 30 – Paramètres de classification

4.3 Résultats

Les résultats de notre segmentation sont validés en se basant sur la similitude des régions estimées par notre approche en comparaison avec celles de l'expert. Les mesures de précision et du rappel des super-pixels sont calculées pour estimer le chevauchement entre les régions segmentées et celles de vérité terrain.

$$Precision = \frac{VP}{VP + FP} \qquad Rappel = \frac{VP}{VP + FN}$$

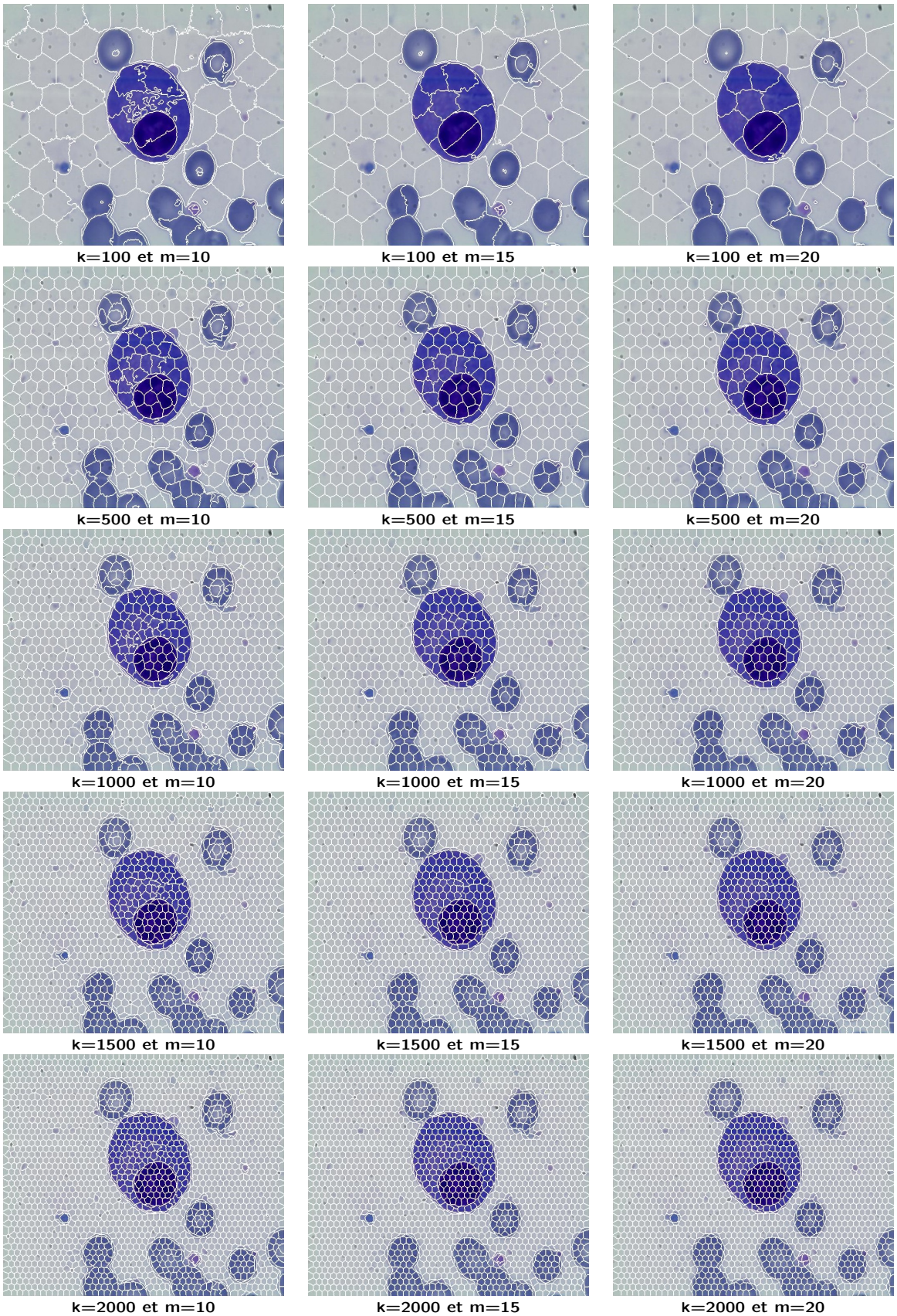


Figure 25 – Exemple du super-pixel avec différente contrainte de SLIC.

VP (Vrai Positif) : nombre de super-pixels positifs classés positifs.
FP (Faux Positif) : nombre de super-pixels négatifs classés positifs.
FN (Faux Négatif) : nombre de super-pixels positifs classés négatifs.

Pour mieux apprécier la qualité des résultats de notre méthode, nous faisons appel à une mesure de performance appelée F-score (F) qui est la moyenne harmonique de la précision et du rappel. Elle est définie comme suit :

$$F - score = 2. \frac{Precision\ .\ Rappel}{Precision + Rappel}$$

La valeur de $F - score$ se situe entre $[0 - 1]$, une valeur forte de F-score nous renseigne sur la pertinence de la segmentation.

Afin de présenter une discussion exhaustive du résultat de la segmentation, nous proposons de réaliser une analyse non-paramétrique de la mesure $F - score$, permettant ainsi une description détaillée des résultats de chaque technique de classification. Nous nous concentrerons sur l'utilisation du test FAR (Friedman Aligned-Ranks) [72]. Son application nous permettra de mettre en évidence les meilleurs résultats de segmentation. Un exemple de cette analyse est détaillé dans la section 6.3.

Une discussion en deux volets est proposée dans cette partie, le premier volet concerne l'étude super-pixellique de l'algorithme SLIC en fonction de la compacité m et du nombre de super-pixel k. Dans le deuxième volet on discute les résultats de la segmentation issus de différents modes de classification.

Analyse super-pixelliqe

La régularité super-pixellique est contrôlée par le paramètre de compacité m, de même la quantité super-pixellique est contrôlée par le paramètre k. En effet, une bonne classification super-pixellique repose aussi sur la qualité super-pixellique générée, et la variabilité de m et k peut produire différentes sortes de super-pixel. Cette variabilité nous a obligé à réaliser différentes expérimentations en faisant varier m et k afin d'aboutir à l'équilibre optimal qui conduit à la bonne classification super-pixellique.

La figure 26 représente 6 tracés (3 pour la segmentation du *noyau* et 3 pour la segmentation du *cytoplasme*), dans chacun, 4 courbes pour 4 classifieurs (RF, co-FOREST, SETRED et Arbre de Décision) analysent le Ranking de $Friedman$ en fonction du nombre de super-pixel k ($k = \{100, 500, 1000, 1500, 2000\}$), et dans chaque tracé la compacité m prend respectivement les valeurs 10, 15 et 20. A partir de la figure 26 nous distinguons :

- Entre $k = [100 - 1000]$, une amélioration considérable est observée dû au retour de ranking lors de la progression du super-pixel k, cette observation est la même pour toutes les valeurs de compacité m.
- Lorsque $k > 1000$, une légère variation du ranking est remarquée ce qui donne une certaine stabilité aux performances, quelques exceptions peuvent être signalées dans la segmentation du *cytoplasme* avec $m = \{10, 15\}$. En revanche, les meilleures performances sont obtenues avec $k = \{1000, 1500, 2000\}$.

La figure 27 présente un exemple de segmentation du *noyau* et du *cytoplasme*, la segmentation affichée est obtenue par le classifieur de forêt aléatoire (RF) en fonction de $k = \{100, 500, 1000, 1500, 2000\}$. Il est clair que plus on produit de super-pixel plus la classification est précise et les contours raffinés. Afin d'apprécier les différences de performance entre les différentes valeurs de m et k, des comparaisons non-paramétriques sont réalisées et affichées dans le tableau 31. Cette comparaison nous a permis de fixer les valeurs de m et k pour chaque classifieur qui conduit à la meilleure segmentation du *noyau* et du *cytoplasme*.

D'après cette analyse, la régularité m et la quantité super-pixellique k entrent en jeu dans les performances de segmentation, dans notre application qui concerne la segmentation des images cytologiques, nous fixons $m = [10 - 15]$ et $k = [1000 - 1500]$ comme un jeu de valeur pour aboutir à une bonne classification super-pixellique.

	Noyau		*Cytoplasme*	
Classifieur	k	m	k	m
Arbre de Décision	1000	10	1000	10
RF	1500	10	1500	10
SETRED	1500	10	1500	10
co-FOREST	1500	15	1500	15

Table 31 – Meilleur jeu de k et m pour chaque classifieur.

Analyse de classification super-pixellique

La classification super-pixellique consiste à affecter une classe de région pour chaque super-pixel produit dans l'image, et la segmentation finale est apportée en fusionnant les super-pixels de même classe. Avant d'établir la classification, il est nécessaire de faire apprendre une hypothèse qui sera confiée à cette action.

Dans ce contexte, nous mettons en comparaison le concept ensembliste et le concept semi-supervisé, afin de mettre en place une technique de classification robuste pour la classification super-pixellique.

Nous avons réalisé des tests non-paramétriques de $Friedman$ pour la segmentation du *noyau* (tableau 32) et pour la segmentation du *cytoplasme* (tableau 33). A partir des tableaux 32 et 33, on peut constater que la meilleure technique de classification avec laquelle nous avons obtenu une bonne segmentation est co-FOREST. La classification ensembliste supervisée de type RF est classé en deuxième position, en troisième position vient le classement de SETRED comme une technique mono-classifieur semi-supervisée et le mono-classifieur supervisé Arbe de Décision est classé en dernier.

Nous confirmons la conclusion de [11, 12] qui constate que les techniques ensemblistes sont les plus performants dans la classification, ceci est statistiquement justifié par un large écart de performance par rapport à SETRED et Arbre de Décision (tableaux 32 et 33). Notre deuxième observation : l'apprentissage semi-supervisé surpasse l'apprentissage supervisé comme il est indiqué dans le classement de Freidman (tableaux 32 et 33).
Le classifieur semi-supervisé co-FOREST a apporté un écart de $p - value = 0.37841$ en comparaison avec les performances du classifieur supervisé RF, qui reste un écart important et indique aussi l'amélioration apportée par l'effet d'intégration des données non-étiquetées dans l'apprentissage semi-supervisé. Cela a permis de produire une hypothèse robuste et efficace pour la classification super-pixellique. En terme de mono-classifieur, l'apprentissage semi supervisé SETRED est mieux classé par rapport à l'apprentissage supervisé Arbre de Décision, comme il est indiqué dans les tableaux 32 et 33.

Afin d'apprécier cette amélioration, nous exposons quelques exemples de segmentation dans les figures 28 et 29, et pour chaque image nous indiquons la mesure de $F - score$. À partir de ces figures, des globules blancs de différente nature et structure sont très bien segmentés, bénéficiant de la robustesse de co-FOREST dans la classification super-pixellique. Dans l'image 1 et 5, les contours obtenus sont extrêmement raffinés par co-FOREST en comparaison avec les autres classifieurs. Des améliorations sont nettement observées au niveau de la classification ensembliste (co-FOREST et RF) en comparaison avec le mono-classifieur (images 3, 4, 7 et 8).

La classification super-pixellique réalisé par tous les classifieurs donnent de bons résultats pour la segmentation du *noyau*, en raison de leur perception qui est uniforme, clair et facile à être séparé. Contrairement à la segmentation du *cytoplasme*, qui présente une certaine difficultée dans la segmentation. Dans nos résultats, une précision satisfaisante est obtenue par le classifieur co-FOREST, comme il est indiqué dans les images 3, 4, 7 et 9. Par exemple, la mesure $F - score$ du *cytoplasme* dans l'image 4 ($F - score_{cytoplasme} = 0,9339$) représente une grande précision de segmentation par co-FOREST en comparaison par rapport aux autres classifieurs ($F - score_{cytoplasme} < 0.8$).

Classement	Technique	*Friedman*	p-value
1	co-FOREST		
2	RF	$0.88\overline{0}83$	$0.3\overline{7}841$
3	SETRED	5.373065	0
4	Arbre de Décision	7.134726	0

Table 32 – Classement de meilleurs classifieurs pour la segmentation du *noyau*.

Classement	Technique	*Friedman*	p-value
1	co-FOREST		
2	RF	$1.11\overline{5}718$	$0.26\overline{4}543$
3	SETRED	3.082906	0.00205
4	Arbre de Décision	3.435238	0.000592

Table 33 – Classement de meilleurs classifieurs pour la segmentation du *cytoplasme*.

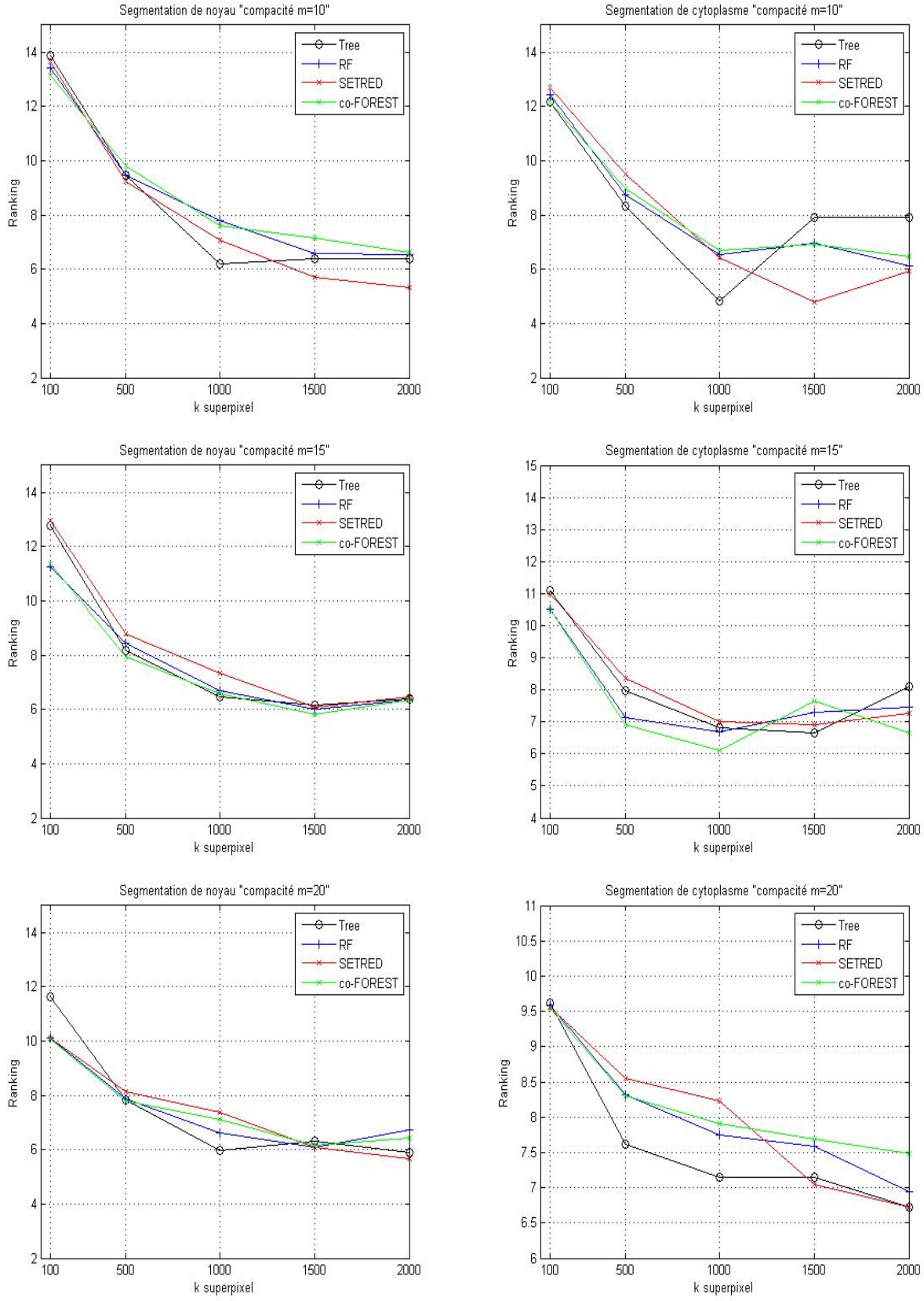


Figure 26 – Ranking de la classification super-pixellique du *noyau* et du *cytoplasme* en fonction de $k = \{100, 500, 1000, 1500, 2000\}$ pour chaque valeur de compacité $m = \{10, 15, 20\}$.

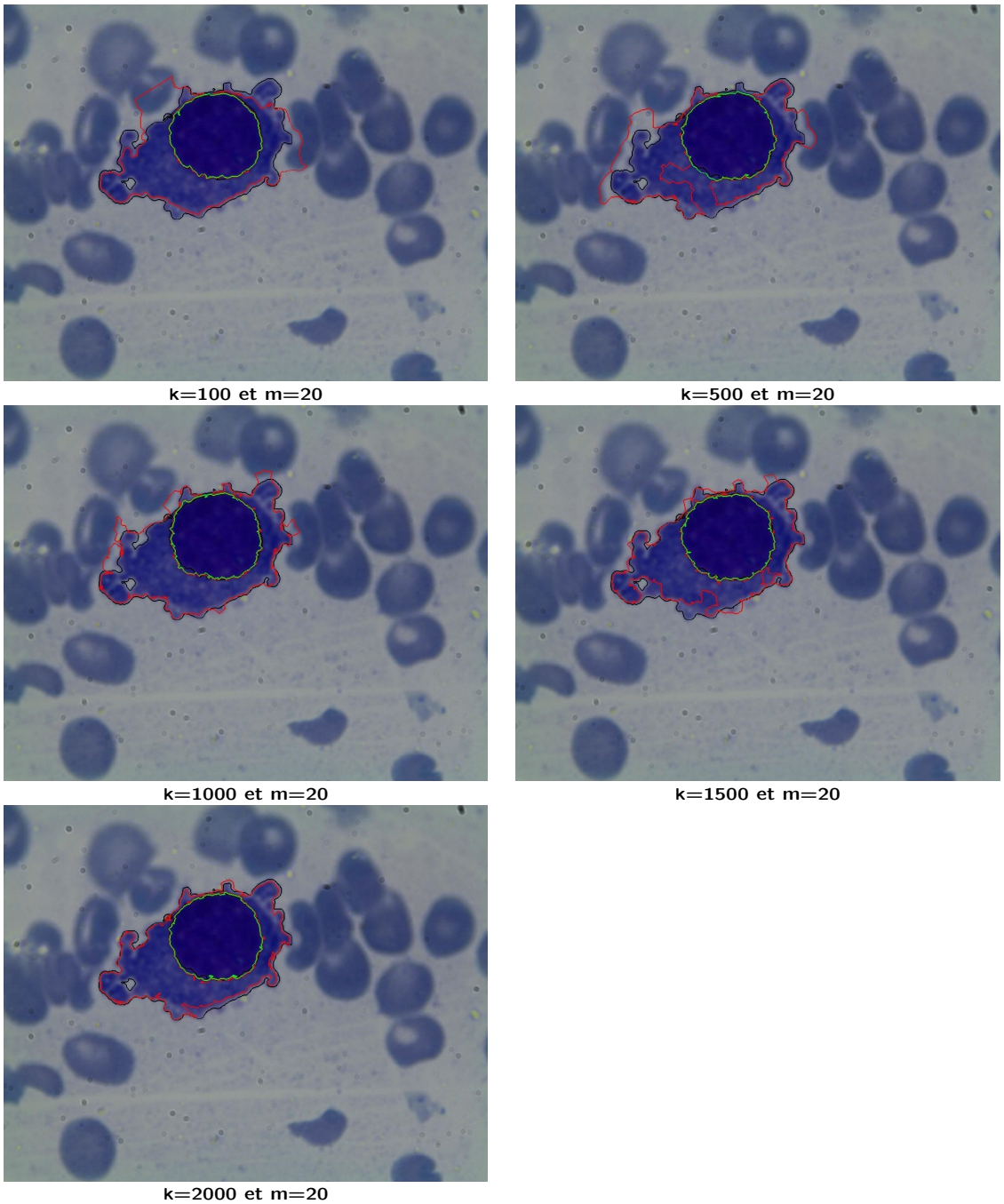


Figure 27 – Résultats de la segmentation en fonction de différente valeur de k. Contour vert pour la segmentation du *noyau*, contour rouge pour la segmentation du *cytoplasme* et la vérité terrain est donnée par le contour noir.

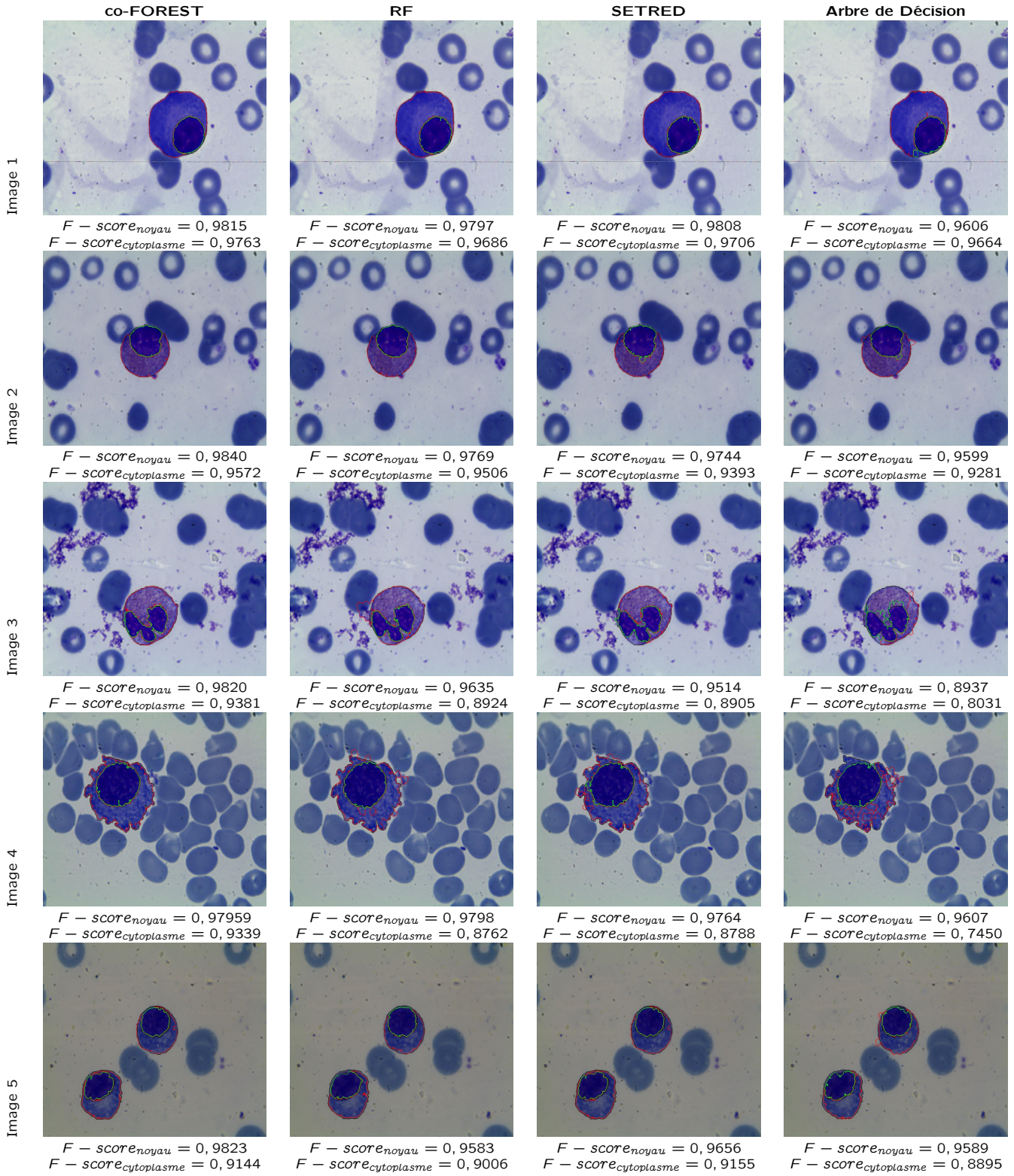


Figure 28 – Résultats de segmentation par classification super-pixellique de co-FOREST, RF, SETRED et Arbre de Décision. Contour vert pour la segmentation du *noyau*, contour rouge pour la segmentation du *cytoplasme* et la vérité terrain est donnée par le contour noir.

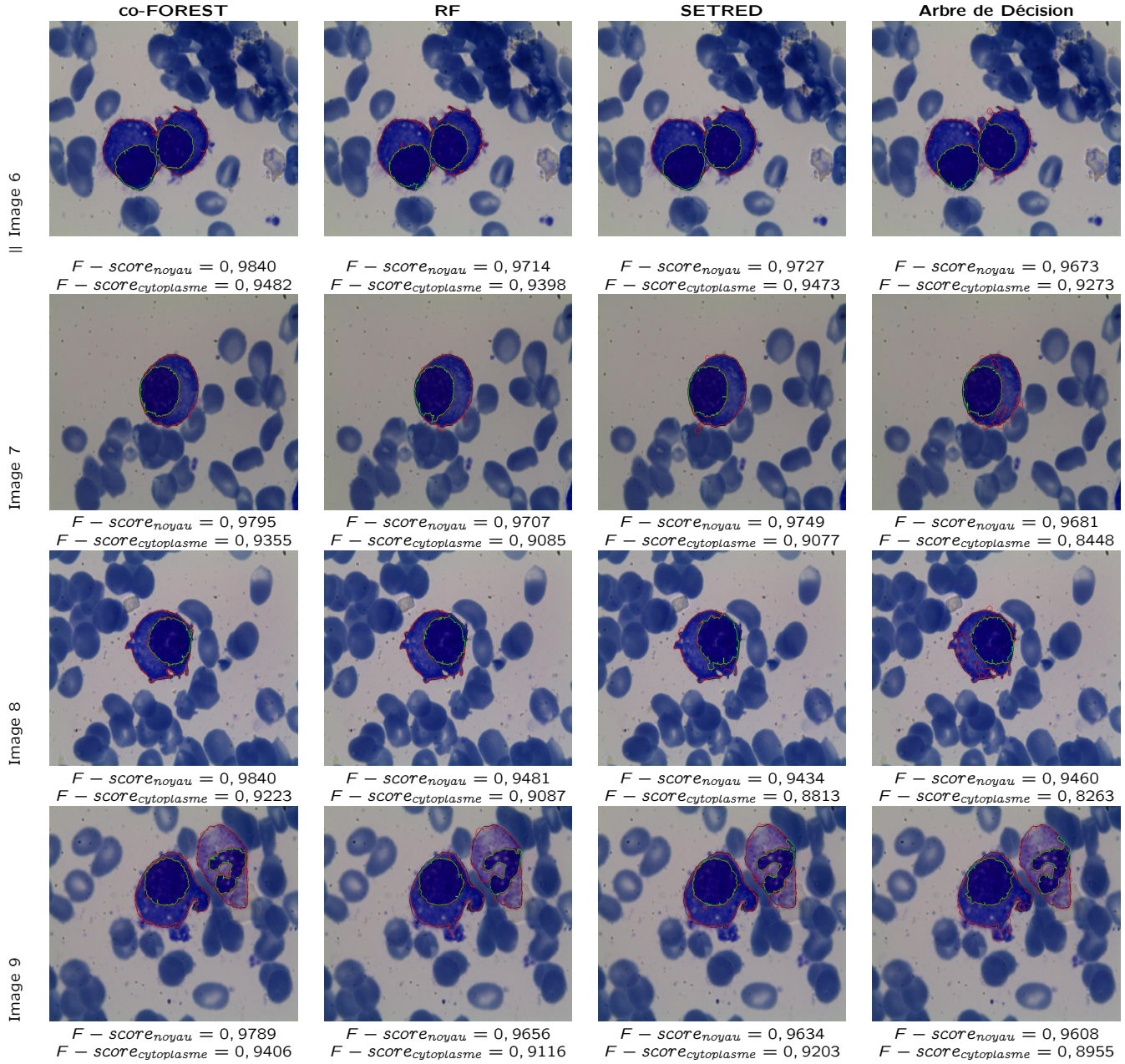


Figure 29 – Résultats de segmentation par classification super-pixellique de co-FOREST, RF, SETRED et Arbre de Décision. Contour vert pour la segmentation du *noyau*, contour rouge pour la segmentation du *cytoplasme* et la vérité terrain est donnée par le contour noir. (Suite)

Conclusion

Dans cette deuxième partie de la thèse, nous nous sommes intéressés à la problématique de la segmentation cellulaire en vue d'aide au diagnostic. L'objectif est de construire un processus permettant d'extraire le *noyau* et le *cytoplasme* d'une image couleur de la cytologie sanguine. Nous avons abordé le concept de la classification dans la segmentation en utilisant la classification super-pixellique. Le choix de ce concept revient à l'efficacité des tâches d'apprentissage artificiel, qui vise à étiqueter les super-pixels dont l'étiquette est inconnue, soit parce qu'elle est difficile ou couteuse à obtenir. Ainsi, pour entrainer un classifieur, on dispose souvent d'une base d'exemples étiquetés, et d'une masse d'exemples non étiquetés. Pour cela, l'intégration de l'apprentissage semi-supervisé a été proposée dans notre application pour produire un classifieur de segmentation fiable.

Le super-pixel est un prétraitement applicable sur l'image dans le but de produire des sous-régions, la propriété du super-pixel est que chacun décrit un groupe de pixel portant une similarité en terme de couleur. L'avantage de ce prétraitement réside dans la capacité à délimiter les contours des régions incluses dans l'image, et cela a permis de traiter des super-pixels au lieu des pixels dans de nombreuse application de vision par ordinateur. Dans procédure de traitement, on a fait appel à l'algorithme SLIC pour la génération des super-pixels, deux contraintes doivent être fixées par l'utilisateur, la compacité m pour contrôler la similarité du super-pixel et le paramètre k pour simuler le nombre de super-pixel souhaité dans l'image. La variabilité de entre m et k peut conduire à des caractéristiques super-pixelliques différentes qui peuvent influencee les résultats de la segmentation. Dans le chapitre 2, nous avons étudié la variabilité de la compacité m et de la quantité k afin de choisir les valeurs optimales qui conduisent à une segmentation fiable. Suite à différentes simulations entre $m = \{10, 15, 20\}$ et $k = \{100, 500, 1000, 1500, 2000\}$, nous avons constaté qu'il y a un compromis entre $m = [10 - 15]$ et $k = [1000 - 1500]$ comme jeu de valeur pour obtenirvà une meilleure segmentation de nos images cellulaires.

Selon la nature des données initiales d'apprentissage, l'idéal est de trouver et d'utiliser la technique de classification qui conduit aux meilleurs résultats de classification . En apprentissage, on distingue généralement deux approches : le cas supervisé et le cas non supervisé, on peut combiner celles-ci dans un apprentissage semi-supervisé. Dans ce cadre, on enrichit un ensemble de données étiquetées par l'intégration d'exemples non étiquetés et ceci en utilisant l'apprentissage dite semi-supervisé. Typiquement, ce type d'apprentissage se présente lorsque l'obtention d'une classe d'un exemple est coûteuse.

D'un point de vue de performance, nous avons étudié dans le chapitre 2 les avantages et la capacité prédictive des méthodes d'ensemble, plus spécifiquement des forêts aléatoires. Sur le même principe, nous avons proposé dans ce travail de reprendre l'algorithme Random Forest dans un contexte semi supervisé afin de produire une hypothèse de classification plus robuste que celle produite en supervisé.

Dans les expérimentations du chapitre 2, une comparaison est menée entre les multi-classifieurs et le mono-classifieur en mode supervisé (forêts aléatoires vs. arbre de décision) et en mode semi-supervisé (co-FOREST vs. auto-apprentissage SETRED), dans l'objectif de présenter une discussion exhaustive de la classification super-pixellique.

En mode supervisé, les résultats de segmentation obtenus par les forêts aléatoires surpassent nettement la segmentation obtenue par arbre de décision, grâce à la concaténation des décisions de multi-classifieur et l'efficacité ensembliste de forêts aléatoires. En mode semi-supervisé, l'utilisation des multi-classifieurs semi-supervisé co-FOREST s'est révélée efficace dans la segmentation du *noyau* et du *cytoplasme* en comparaison avec l'auto-apprentissage SETRED, les statistiques réalisées indiquent que l'apprentissage semi-supervisé améliore significativement l'apprentissage supervisé.

Dans des applications images, l'apprentissage se fait suite à une phase de caractérisation (extraction de variable), qui restera une étape primordiale dans n'importe quel système de vision par ordinateur. L'objectif de cette étape réside dans la manière de représenter une information couleur, texture ou forme en un vecteur de caractéristique capable d'être exploité par un classifieur.

Notre procédure de segmentation est basée sur la classification super-pixellique, donc la caractérisation se fait au niveau super-pixellique basée sur l'information couleur de l'image. Différentes manières de caractérisation ont été développées dans le chapitre 1, mettant en question l'effet de la normalisation couleur, la composante couleur et la technique de caractérisation. Notre proposition a été d'étudier l'influence de chaque point sur les résultats de segmentation.

Les conclusions constatées indiquent que la normalisation de couleur apporte des précisions en caractérisation et des améliorations importantes en termes de segmentation. Concernant la composante couleur, le choix de l'espace couleur à utiliser reste toujours en suspens à définir, de ce fait, nous avons étudié les pertinences de quatre espaces couleurs (Hsv, I1I2I3, Lab et RGB) de propriété différente, et les résultats réalisés ont mené sur le choix de RGB comme espace à exploiter dans le processus de segmentation. Dans la caractérisation super-pixellique on s'est basé sur l'information couleur de chaque super-pixel, on a fait appel à quatre techniques de caractérisation (CM, FOS, Contraste et moment de Hu) afin de réaliser la caractérisation demandée. Suite à une étude détaillée utilisant les tests non-paramétrique de *Friedman*, nous avons constaté que la caractérisation de FOS est notée comme la meilleur manière de représenter le super-pixel, et avec la combinaison de la normalisation CGWN et la couleur RGB, nous avons obtenu les meilleurs résultats de segmentation.

Quatrième partie

IV

Conclusion et Perspectives

Conclusion générale & Perspectives

Conclusion

Dans ce travail de thèse, la problématique de la classification partiellement supervisée des données médicales est abordée. L'étiquetage manuel d'une donnée médicale est de plus en plus difficile à obtenir. Les travaux de la littérature liés à ce sujet ont mené au développement d'un nouveau type d'apprentissage, dit semi-supervisé. La qualité majeure de ce type d'apprentissage est de combiner à la fois les données étiquetées et non étiquetées lors de l'apprentissage du classifieur, permettant de prédire les classes manquantes des données non étiquetées.

La classification semi-supervisée trouve ses racines dans les problèmes d'apprentissage en présence d'une majorité des données non étiquetées. De nombreux travaux ont été réalisés depuis les années 70. Ce thème de recherche a connu ensuite un regain d'intérêt à la fin des années 1990 dans la communauté du Machine Learning, avec la disponibilité croissante de grands jeux de données grâce aux nouvelles technologies.

Le premier algorithme développé dans ce domaine est intitulé algorithme d'auto-apprentissage. Ce principe a connu une large utilisation dans plusieurs applications, l'avantage de cette technique réside dans sa simplicité d'implémentation, sa rapidité d'exécution et ses performances. Nous nous sommes intéressé pour développer l'auto-apprentissage sur les données médicales et étudier en même temps ses avantages et ses limites sur la segmentation des images médicales, car ce domaine souffre de la complexité de l'expertise humaine.

La similarité est la notion la plus utilisée en apprentissage semi-supervisé, cette notion est d'un grand intérêt dans la définition des éléments de confiance. Dans une première partie, nous avons détaillé la mesure de similarité dans un processus d'auto-apprentissage (algorithme *SNNRCE*) dans l'objectif d'étudier l'influence de cette similarité sur la définition des éléments de confiance ainsi que sur les performances de classification.

Dans cette même partie, nous avons proposé un nouvel algorithme de classification semi-supervisée suivant le principe d'auto-apprentissage. Les difficultés majeures trouvées dans les premières expérimentations ont été pris en considération dans cette proposition. Dans le nouvel algorithme *R-COSET*, nous avons concentré nos efforts sur l'amélioration de la mesure de confiance en utilisant les outils de graphe de voisinage relatif et la technique de *CEWS*. En effet, *R-COSET* procède à un graphe de voisinage sur deux niveaux de construction permettant ainsi de donner une vue plus exhaustive de l'espace de représentation et plus d'informations sur les données voisines. Cette proposition a permis de renforcer la mesure de confiance en auto-apprentissage. Des tests statistiques et non paramétriques ont été réalisés pour tester la pertinence du modèle, prouvant ainsi sa

supériorité par rapport à d'autres modèles proposés dans le même domaine.

Un autre volet important dans cette thèse concerne principalement la segmentation des images médicales. L'intérêt majeur de cette partie dans le domaine du traitement d'images médicales nous a motivé pour mettre davantage en évidence le potentiel applicatif de semi supervision dans la segmentation des images cytologiques et la reconnaissance des globules blancs. La segmentation par classification super-pixellique est devenue une approche potentielle pour ces applications. La deuxième partie de cette thèse concerne notre deuxième contribution i.e la segmentation par classification semi-supervisé super-pixellique. La caractérisation super-pixellique est considérée comme la partie clef dans la réussite de la segmentation. En premier lieu, nous avons réalisé de multiples expérimentations avec différentes procédures de caractérisation super-pixellique. Celles-ci nous ont permis de mettre en évidence la multitude de questions posées dans ce cadre, ainsi que les nombreuses méthodes développées pour y répondre. Les comparaisons statistiques effectuées ont justifié le choix des caractérisations pertinentes puisque celles-ci permettent de prendre en considération l'information présente dans les super-pixels.

En deuxième lieu, nous avons abordé la partie de la classification super-pixellique en mode supervisé et semi-supervisé, afin de montrer la valeur ajoutée de l'apprentisage semi-supervisé dans des applications d'imagerie médicale. Cette fois-ci, nous avons intégré le concept ensembliste dans la classification, sachant que ce dernier permet de contourner les limites de la classification par mono-classifieur, en mode supervisé et semi-supervisé. Dans nos expérimentations, nous avons utilisé l'algorithme de la Forêt Aléatoire, ce dernier a été exécuté en mode supervisé (*Random Forest*) et en mode semi-supervisé (*co-Forest*). La segmentation obtenue a été comparée par rapport au mono-classifieur en mode supervisé (*Arbre de Décision*) et en mode semi-supervisé (auto-apprentissage *SETRED*). Lors cette comparaison et via les résultats de segmentation obtenus, l'apprentissage semi-supervisé a montré sa supériorité par rapport à l'apprentissage supervisé. Par contre, l' algorithme d' auto-apprentissage *SETRED* a obtenu moins de précision de segmentation, en comparaison avec l'approche ensembliste *co-Forest* qui a donné plus de précision dans la segmentation.

Perspectives

Ce travail de thèse constitue un début intéressant à des pistes de recherche futures, Nous pensons principalement aux points suivants :

- Actuellement, les procédures semi-supervisées visent à explorer les récentes techniques de classification supervisée, comme le Deep Learning [183]. Par définition cela signifie un apprentissage profond, qui est un ensemble de méthodes d'apprentissage automatique qui permettent de modéliser avec un haut niveau d'abstraction de données grâce à des architectures articulées avec différentes transformations non linéaires. Ces techniques ont permis des progrès importants et rapides dans les domaines de l'analyse du signal sonore ou visuel et notamment de la reconnaissance faciale, de la reconnaissance vocale, de la vision par ordinateur et du traitement automatisé du langage. La puissance du Deep Learning peut être exploitée dans le contexte semi-supervisé, et de bénéficier de sa robustesse pour résoudre la problématique des données partiellement supervisées, les premiers travaux ont été entamées dans [184], [185], et [186].

- Concernant la segmentation par classification super-pixellique, elle reste toujours une approche potentielle dans les tâches de segmentation d'images. Il faut noter que la caractérisation super-pixellique joue un rôle important dans le succès de la segmentation. Dans cette thèse nous avons exploité l'information couleur pour cette caractérisation. Comme travaux futurs et afin d'améliorer cette étude, il sera judicieux d'intégrer la caractérisation l'information de texture.

– Proposer une architecture d'annotation semi-automatique des images médicales pour extraire des informations médicales spécifiques (i.e. modalité médicale (image globale), région anatomique (au niveau des pixels de l'image)) à partir du contenu et du contexte des images. L'annotation des images médicales par apprentissage semi-supervisé en tenant compte de la labélisation partielle des experts peut être une autre piste de recherche à exploiter.

– Exploitation du contexte semi-supervisé dans une approche de classification non supervisée, et cela par l'intégration des données étiquetées dans le mécanisme d'apprentissage non supervisé, afin de renforcer le regroupement des données non étiquetées. Dans ce contexte, nous pourrons envisager de laisser les données étiquetées pour "guider" ou "ajuster" le processus de regroupement, à savoir fournir une forme limitée de supervision. L'approche résultante est appelée le *clustering semi-supervisé*. Cette approche peut être exploitée lorsque les données labellisées disponibles sont loin d'être représentatives pour aboutir à une classification ciblée des exemples, de sorte que l'apprentissage supervisé est difficile à réaliser, même sous une forme transductive. Contrairement au principe de regroupement traditionnel, l'approche de regroupement semi-supervisé se résume en quelques méthodes publiées jusqu'à présent. La principale distinction entre ces méthodes concerne la façon dont les deux sources d'information sont combinées : soit en adaptant la mesure de similarité [187–189] ou en modifiant la recherche de grappes (clusters) appropriées [190–192].

– Appliquer les algorithmes semi-supervisés existants pour résoudre la problématique de poly-pathologies chez les patients. Cette problématique de classification *multi-labels* est affirmée par les praticiens, vu que, sur le terrain, les patients peuvent être atteints de plusieurs pathologies simultanément.

Productions scientifiques

De nombreux résultats ont été obtenus dans cette thèse, nos contributions réalisées ont abouti à plusieurs travaux scientifiques (publications et communications internationales). Notons aussi qu'une partie de nos travaux ont été réalisé dans le cadre d'une collaboration scientifique avec des chercheurs de l'équipe CREDOM (Settouti Nesma) et des chercheurs de l'équipe Machine Leraning (Vincent Barra, Clermont Ferrand - France) et de l'équipe de Groupe Signaux Multidimensionnels (Adel Mouloud, Institut FRESNEL).

Journaux internationaux

– **Bechar, M. E. A.**, Settouti, N., Barra, V., and Chikh, M. A. (2017). *Semi-supervised super-pixel classification for medical images segmentation : application to detection of glaucoma disease*. Multidimensional Systems and Signal Processing, 1-20. (First Online : 14 March 2017)
– Settouti, N., **Bechar, M. E. A.**, and Chikh, M. A. (2016). *Statistical comparisons of the top 10 algorithms in data mining for classification task*. Int. J. Interact. Multimedia Artif. Intell., Special Issue Artif. Intell. Underpinning, 4, 46-51.
– **Bechar, M. E. A.**, Settouti, N., Chikh, M.A. and Adel, M. (xxxx) *Reinforced confidence in self-training for a semi-supervised medical data classification*, Int. J. Applied Pattern Recognition, Vol. X, No. Y, pp.xxx–xxx. (In press)
– Settouti, N., **Bechar, M. E. A.**, El HABIB DAHO, M., and Chikh, M. A. (201X). *An optimized pixel-based classification approach for automatic white blood cells segmentation*, Int. J. Biomedical Engineering and Technology, Vol. X, No. Y, pp.xxx–xxx. (In press)

Conférences Internationales

- Saidi, M., **Bechar, M. E. A.**, Settouti, N., Chikh, M. A. (2016, November). Application of pixel selection in pixel-based classification for automatic white blood cell segmentation. In Proceedings of the Mediterranean Conference on Pattern Recognition and Artificial Intelligence (pp. 31-38). ACM.
- Settouti, N., **Bechar, M. E. A.**, and Chikh, M. A. (2015). *Statistical comparisons of the Top 10 algorithms in data mining for classification task*. International conference Advanced Information Technology, Services and Systems (AIT2S-15). December 16-17, 2015 Faculty of Sciences & Technologies, Settat, Morocco.
- **Bechar, M. E. A.**, Settouti, N., El HABIB DAHO, M., and Chikh, M. A. (2015). *Croissance de région par classification pixellique : Application aux images cytologiques*. Systeme Conjoint de Compression et Indexation des Objets Videos : SCCIBOV'2015, December 02-03, 2015, Djillali Liabes University -Faculty of Technology, Sidi Bel Abbes, Algeria.
- **Bechar, M. E. A.**, Settouti, N., and Chikh, M. A. (2015) *L'impact de la mesure de similarité en auto-apprentissage*. Colloque sur l'Optimisation et les Systèmes d'Information COSI'2015, 1 au 3 juin 2015, Oran, Algérie Université d'Oran 1, Ahmed Ben Bella.
- Settouti, N., **Bechar, M. E. A.**, EL HABIB DAHO, M., LAZOUNI, M. A., and CHIKH, M. A. (2014) *Bagged Nearest Neighbor Classifiers in Semi Supervised Learning*. The 19 International Conference on Mechanics in Medecine and Biology (ICMMB'14), September 3-5 2014, Bologna, Italy.
- **Bechar, M. E. A.**, and CHIKH, M. A. (2014) *Semi-supervised learning based on nearest neighbor rule and weighted mean*. Biomedical Engineering International Conference (BIOMEIC'14), 15-16 october 2014, Tlemcen, Algérie Université de Tlemcen, Abou Bekr Belkaid.
- **Bechar, M. E. A.**, and CHIKH, M. A. (2014) *Classification pixellaire partiellement supervisée des images cytologiques*. La Troisième conférence International sur les Système Complexes (CISC'14), 09-10 Décembre 2014, Université de Jijel, Algerie. 2014.

Cinquième partie

V

Annexes

Annexe A : Résultat non-paramétrique de la normalisation couleur

Dans cette annexe, nous présentons le détail de l'analyse non-paramétrique concernant la normalisation couleur. Les tableaux 34, 36, 38, 40, 42, 44, 46, 48 et 50 affichent les mesures de Friedman de chaque combinaison entre E_j et C_k pour la segmentation du *noyau*, et les tableaux 35, 37, 39, 41, 43, 45, 47, 49 et 51 pour la segmentation du *cytoplasme*.

Normalisation N_1		
Comparaison	Friedman	p-value
E_4C_2 vs. E_2C_2	17.563338	0
E_4C_2 vs. E_2C_3	17.563338	0
E_1C_2 vs. E_2C_2	17.205065	0
E_1C_2 vs. E_2C_3	17.205065	0
E_4C_1 vs. E_2C_2	16.958255	0
E_4C_1 vs. E_2C_3	16.958255	0
E_3C_2 vs. E_2C_2	16.369095	0
E_3C_2 vs. E_2C_3	16.369095	0
E_4C_2 vs. E_4C_4	15.294276	0
E_1C_2 vs. E_4C_4	14.936003	0
.	.	.
.	.	.
.	.	.
E_1C_3 vs. E_3C_3	0.087578	0.930212
E_2C_2 vs. E_2C_3	0	1

Table 34 – Comparaison multiple de la segmentation du *noyau* en fonction de la normalisation CGWN.

Normalisation N_1		
Comparaison	Friedman	p-value
E_4C_2 vs. E_1C_1	16.385018	0
E_4C_2 vs. E_4C_4	16.385018	0
E_4C_2 vs. E_2C_2	15.294276	0
E_1C_2 vs. E_1C_1	15.262429	0
E_1C_2 vs. E_4C_4	15.262429	0
E_3C_2 vs. E_1C_1	15.254468	0
E_3C_2 vs. E_4C_4	15.254468	0
E_1C_2 vs. E_2C_2	14.171687	0
E_3C_2 vs. E_2C_2	14.163725	0
E_4C_2 vs. E_2C_4	14.115956	0
.	.	.
.	.	.
.	.	.
E_1C_2 vs. E_3C_2	0.007962	0.993648
E_1C_1 vs. E_4C_4	0	1

Table 35 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la normalisation CGWN.

Normalisation N_2		
Comparaison	Friedman	p-value
E_4C_2 vs. E_2C_2	17.531492	0
E_4C_1 vs. E_2C_2	17.077679	0
E_1C_2 vs. E_2C_2	16.193939	0
E_4C_2 vs. E_1C_4	15.716242	0
E_3C_2 vs. E_2C_2	15.302237	0
E_4C_1 vs. E_1C_4	15.262429	0
E_2C_1 vs. E_2C_2	15.007657	0
E_2C_4 vs. E_2C_2	14.792694	0
E_1C_2 vs. E_1C_4	14.378689	0
E_4C_3 vs. E_2C_2	14.275188	0
.	.	.
.	.	.
.	.	.
E_1C_3 vs. E_3C_1	0.015923	0.987296
E_3C_1 vs. E_3C_3	0.015923	0.987296

Table 36 – Comparaison multiple de la segmentation du *noyau* en fonction de la normalisation CHROMA.

Normalisation N_2		
Comparaison	Friedman	p-value
E_4C_2 vs. E_2C_2	16.138208	0
E_4C_2 vs. E_2C_3	16.138208	0
E_1C_2 vs. E_2C_2	16.018783	0
E_1C_2 vs. E_2C_3	16.018783	0
E_4C_2 vs. E_1C_1	15.676434	0
E_1C_2 vs. E_1C_1	15.557009	0
E_3C_2 vs. E_2C_2	15.079312	0
E_3C_2 vs. E_2C_3	15.079312	0
E_3C_2 vs. E_1C_1	14.617538	0
E_4C_2 vs. E_1C_4	14.060224	0
.	.	.
.	.	.
.	.	.
E_3C_3 vs. E_4C_3	0.071655	0.942877
E_2C_2 vs. E_2C_3	0	1

Table 37 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la normalisation CHROMA.

Normalisation N_3		
Comparaison	Friedman	p-value
E_4C_1 vs. E_2C_2	14.362766	0
E_4C_1 vs. E_2C_3	14.299073	0
E_1C_2 vs. E_2C_2	13.295909	0
E_1C_2 vs. E_2C_3	13.232216	0
E_2C_1 vs. E_2C_2	12.507708	0
E_2C_1 vs. E_2C_3	12.444015	0
E_3C_4 vs. E_2C_2	12.244975	0
E_3C_4 vs. E_2C_3	12.181282	0
E_4C_2 vs. E_2C_2	12.030011	0
E_4C_2 vs. E_2C_3	11.966318	0
.	.	.
.	.	.
.	.	.
E_2C_2 vs. E_2C_3	0.063693	0.949215
E_3C_2 vs. E_4C_4	0.015923	0.987296

Table 38 – Comparaison multiple de la segmentation du *noyau* en fonction de la normalisation CLAHE.

Normalisation N_3		
Comparaison	Friedman	p-value
E_3C_2 vs. E_2C_2	13.558642	0
E_3C_2 vs. E_2C_3	12.436053	0
E_4C_2 vs. E_2C_2	12.069819	0
E_1C_2 vs. E_2C_2	11.074616	0
E_4C_2 vs. E_2C_3	10.94723	0
E_3C_2 vs. E_3C_4	9.999797	0
E_1C_2 vs. E_2C_3	9.952027	0
E_4C_3 vs. E_2C_2	9.139942	0
E_3C_2 vs. E_1C_4	9.020518	0
E_3C_2 vs. E_3C_1	8.702053	0
.	.	.
.	.	.
.	.	.
E_1C_4 vs. E_3C_1	0.318465	0.750132
E_2C_4 vs. E_3C_3	0.095539	0.923886

Table 39 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la normalisation CLAHE.

Normalisation N_4		
Comparaison	Friedman	p-value
E_3C_2 vs. E_2C_2	15.11912	0
E_3C_2 vs. E_2C_3	15.11912	0
E_4C_3 vs. E_2C_2	14.816578	0
E_4C_3 vs. E_2C_3	14.816578	0
E_1C_2 vs. E_2C_2	14.792694	0
E_1C_2 vs. E_2C_3	14.792694	0
E_2C_4 vs. E_2C_2	14.784732	0
E_2C_4 vs. E_2C_3	14.784732	0
E_4C_4 vs. E_2C_2	13.956723	0
E_4C_4 vs. E_2C_3	13.956723	0
.	.	.
.	.	.
.	.	.
E_1C_2 vs. E_2C_4	0.007962	0.993648
E_2C_2 vs. E_2C_3	0	1

Table 40 – Comparaison multiple de la segmentation du *noyau* en fonction de la normalisation GWN.

Normalisation N_4		
Comparaison	Friedman	p-value
E_3C_2 vs. E_1C_1	15.764011	0
E_3C_2 vs. E_2C_2	15.334084	0
E_4C_2 vs. E_1C_1	15.143005	0
E_3C_2 vs. E_1C_4	15.023581	0
E_1C_2 vs. E_1C_1	14.991734	0
E_4C_2 vs. E_2C_2	14.713077	0
E_1C_2 vs. E_2C_2	14.561807	0
E_4C_2 vs. E_1C_4	14.402574	0
E_1C_2 vs. E_1C_4	14.251303	0
E_3C_2 vs. E_3C_4	13.550681	0
.	.	.
.	.	.
.	.	.
E_1C_2 vs. E_4C_2	0.151271	0.879762
E_2C_3 vs. E_2C_4	0.007962	0.993648

Table 41 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la normalisation GWN.

Normalisation N_5		
Comparaison	Friedman	p-value
E_4C_1 vs. E_1C_2	1.783403	0.074521
E_3C_4 vs. E_1C_2	1.544555	0.122454
E_4C_1 vs. E_1C_1	1.536593	0.124393
E_4C_1 vs. E_2C_2	1.504747	0.132389
E_4C_1 vs. E_3C_3	1.42513	0.15412
E_2C_1 vs. E_1C_2	1.409207	0.158774
E_4C_1 vs. E_4C_2	1.401245	0.161141
E_4C_1 vs. E_4C_4	1.337552	0.181042
E_3C_4 vs. E_1C_1	1.297744	0.194375
E_4C_1 vs. E_3C_2	1.297744	0.194375
.	.	.
.	.	.
.	.	.
E_1C_1 vs. E_2C_2	0.031846	0.974594
E_3C_3 vs. E_4C_2	0.023885	0.980944

Table 42 – Comparaison multiple de la segmentation du *noyau* en fonction de la normalisation HEQ.

Normalisation N_5		
Comparaison	Friedman	p-value
E_3C_4 vs. E_2C_2	3.232419	0.001227
E_2C_4 vs. E_2C_2	2.921915	0.003479
E_3C_4 vs. E_2C_3	2.882107	0.00395
E_3C_4 vs. E_4C_3	2.762683	0.005733
E_3C_4 vs. E_1C_3	2.659182	0.007833
E_1C_4 vs. E_2C_2	2.65122	0.00802
E_3C_4 vs. E_3C_3	2.587527	0.009667
E_2C_4 vs. E_2C_3	2.571604	0.010123
E_2C_4 vs. E_4C_3	2.45218	0.014199
E_2C_4 vs. E_1C_3	2.348678	0.01884
.	.	.
.	.	.
.	.	.
E_1C_3 vs. E_3C_3	0.071655	0.942877
E_1C_2 vs. E_4C_2	0.007962	0.993648

Table 43 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la normalisation HEQ.

Normalisation N_6		
Comparaison	Friedman	p-value
E_1C_2 vs. E_2C_2	14.904156	0
E_1C_2 vs. E_2C_3	14.904156	0
E_4C_3 vs. E_2C_2	14.084109	0
E_4C_3 vs. E_2C_3	14.084109	0
E_4C_4 vs. E_2C_2	13.972647	0
E_4C_4 vs. E_2C_3	13.972647	0
E_4C_2 vs. E_2C_2	13.765644	0
E_4C_2 vs. E_2C_3	13.765644	0
E_4C_1 vs. E_2C_2	13.686028	0
E_4C_1 vs. E_2C_3	13.686028	0
.	.	.
.	.	.
.	.	.
E_3C_1 vs. E_3C_2	0.04777	0.9619
E_2C_2 vs. E_2C_3	0	1

Table 44 – Comparaison multiple de la segmentation du *noyau* en fonction de la normalisation Lmax.

Normalisation N_6		
Comparaison	Friedman	p-value
$E1C2$ vs. E_1C_1	16.39298	0
$E1C2$ vs. E_2C_2	16.361133	0
$E1C2$ vs. E_2C_3	16.24967	0
$E3C2$ vs. E_1C_1	16.193939	0
$E3C2$ vs. E_2C_2	16.162093	0
$E3C2$ vs. E_2C_3	16.05063	0
$E4C2$ vs. E_1C_1	15.46147	0
$E4C2$ vs. E_2C_2	15.429623	0
$E4C2$ vs. E_2C_3	15.318161	0
$E3C3$ vs. E_1C_1	12.340514	0
.	.	.
.	.	.
.	.	.
E_2C_2 vs. E_2C_3	0.111463	0.911249
E_1C_1 vs. E_2C_2	0.031846	0.974594

Table 45 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la normalisation Lmax.

Normalisation N_7		
Comparaison	Friedman	p-value
E_2C_4 vs. E_2C_2	15.07135	0
E_2C_4 vs. E_4C_2	15.07135	0
E_2C_4 vs. E_2C_3	15.047465	0
E_2C_4 vs. E_4C_3	14.163725	0
E_3C_1 vs. E_2C_2	14.147802	0
E_3C_1 vs. E_4C_2	14.147802	0
E_3C_1 vs. E_2C_3	14.123917	0
E_2C_4 vs. E_3C_3	14.036339	0
E_1C_1 vs. E_2C_2	13.797491	0
E_1C_1 vs. E_4C_2	13.797491	0
.	.	.
.	.	.
.	.	.
E_2C_3 vs. E_4C_2	0.023885	0.980944
E_2C_2 vs. E_4C_2	0	1

Table 46 – Comparaison multiple de la segmentation du *noyau* en fonction de la normalisation MV.

Normalisation N_7		
Comparaison	Friedman	p-value
E_3C_3 vs. E_3C_4	8.733899	0
E_3C_3 vs. E_2C_2	8.463204	0
E_3C_3 vs. E_4C_2	8.224355	0
E_3C_3 vs. E_1C_4	7.770543	0
E_4C_4 vs. E_3C_4	7.452078	0
E_4C_4 vs. E_2C_2	7.181383	0
E_4C_4 vs. E_4C_2	6.942534	0
E_3C_3 vs. E_4C_1	6.831072	0
E_3C_3 vs. E_3C_1	6.679801	0
E_3C_3 vs. E_1C_1	6.671839	0
.	.	.
.	.	.
.	.	.
E_1C_1 vs. E_2C_1	0.063693	0.949215
E_1C_1 vs. E_3C_1	0.007962	0.993648

Table 47 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la normalisation MV.

Normalisation N_8		
Comparaison	Friedman	p-value
E_2C_4 vs. E_4C_2	15.939167	0
E_3C_2 vs. E_4C_2	14.689192	0
E_2C_4 vs. E_1C_1	14.171687	0
E_3C_4 vs. E_4C_2	13.447179	0
E_2C_3 vs. E_4C_2	13.367563	0
E_2C_4 vs. E_3C_1	13.359602	0
E_3C_2 vs. E_1C_1	12.921712	0
E_4C_3 vs. E_4C_2	12.555478	0
E_1C_3 vs. E_4C_2	12.491785	0
E_3C_2 vs. E_3C_1	12.109627	0
.	.	.
.	.	.
.	.	.
E_2C_3 vs. E_3C_4	0.079616	0.936542
E_1C_3 vs. E_4C_3	0.063693	0.949215

Table 48 – Comparaison multiple de la segmentation du *noyau* en fonction de la normalisation RGBcb.

Normalisation N_8		
Comparaison	Friedman	p-value
E_3C_2 vs. E_1C_1	14.203534	0
E_3C_2 vs. E_1C_2	14.203534	0
E_3C_2 vs. E_3C_1	14.203534	0
E_3C_2 vs. E_4C_2	14.203534	0
E_3C_2 vs. E_4C_4	13.558642	0
E_3C_3 vs. E_1C_1	12.268859	0
E_3C_3 vs. E_1C_2	12.268859	0
E_3C_3 vs. E_3C_1	12.268859	0
E_3C_3 vs. E_4C_2	12.268859	0
E_3C_3 vs. E_4C_4	11.623968	0
.	.	.
.	.	.
.	.	.
E_1C_2 vs. E_4C_2	0	1
E_3C_1 vs. E_4C_2	0	1

Table 49 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la normalisation RGBcb.

Normalisation N_9		
Comparaison	Friedman	p-value
E_1C_2 vs. E_2C_2	14.784732	0
E_1C_2 vs. E_2C_3	14.784732	0
E_2C_4 vs. E_2C_2	14.020416	0
E_2C_4 vs. E_2C_3	14.020416	0
E_4C_2 vs. E_2C_2	14.012455	0
E_4C_2 vs. E_2C_3	14.012455	0
E_4C_3 vs. E_2C_2	14.004493	0
E_4C_3 vs. E_2C_3	14.004493	0
E_4C_1 vs. E_2C_2	13.996531	0
E_4C_1 vs. E_2C_3	13.996531	0
.	.	.
.	.	.
.	.	.
E_4C_2 vs. E_4C_3	0.007962	0.993648
E_2C_2 vs. E_2C_3	0	1

Table 50 – Comparaison multiple de la segmentation du *noyau* en fonction de la normalisation Sans Normalisation.

Normalisation N_9		
Comparaison	Friedman	p-value
$E1C2$ vs. E_2C_2	15.684395	0
$E1C2$ vs. E_2C_3	15.405738	0
$E1C2$ vs. E_1C_1	15.206698	0
$E1C2$ vs. E_3C_4	15.047465	0
$E3C2$ vs. E_2C_2	14.585691	0
$E4C2$ vs. E_2C_2	14.338881	0
$E3C2$ vs. E_2C_3	14.307035	0
$E3C2$ vs. E_1C_1	14.107994	0
$E4C2$ vs. E_2C_3	14.060224	0
$E3C2$ vs. E_3C_4	13.948762	0
.	.	.
.	.	.
.	.	.
E_1C_1 vs. E_2C_3	0.199041	0.842231
E_1C_1 vs. E_3C_4	0.159232	0.873486

Table 51 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la normalisation Sans Normalisation.

Annexe B : Résultat non-paramétrique de la composante couleur

Dans cette annexe, nous présentons le détail de l'analyse non-paramétrique concernant la composante de couleur. Les tableaux 52, 54, 56 et 58 affichent les mesures de Friedman de chaque combinaison entre N_i et C_k pour la segmentation du *noyau*, et les tableaux 53, 55, 57 et 59 pour la segmentation du *cytoplasme*.

Espace couleur E_1		
Comparaison	Friedman	p-value
N_1C_2 vs. N_3C_4	14.363146	0
N_2C_2 vs. N_3C_4	13.977521	0
N_1C_2 vs. N_9C_4	13.966344	0
N_2C_2 vs. N_9C_4	13.580718	0
N_1C_2 vs. N_2C_1	13.536008	0
N_2C_2 vs. N_2C_1	13.150383	0
N_1C_2 vs. N_6C_3	13.128028	0
N_2C_2 vs. N_6C_3	12.742402	0
N_1C_2 vs. N_4C_3	12.317656	0
N_8C_3 vs. N_3C_4	12.317656	0
.	.	.
.	.	.
.	.	.
N_1C_1 vs. N_4C_1	0.078243	0.937635
N_3C_1 vs. N_6C_4	0.050299	0.959884

Table 52 – Comparaison multiple de la segmentation du *noyau* en fonction de l'espace couleur HSV.

Espace couleur E_1		
Comparaison	Friedman	p-value
N_6C_2 vs. N_9C_4	14.529289	0
N_9C_2 vs. N_9C_4	14.259853	0
N_1C_2 vs. N_9C_4	13.990418	0
N_6C_2 vs. N_1C_4	13.828757	0
N_2C_2 vs. N_9C_4	13.700776	0
N_9C_2 vs. N_1C_4	13.559322	0
N_1C_2 vs. N_1C_4	13.289887	0
N_6C_2 vs. N_3C_4	13.141698	0
N_2C_2 vs. N_1C_4	13.000244	0
N_9C_2 vs. N_3C_4	12.872263	0
.	.	.
.	.	.
.	.	.
N_1C_3 vs. N_9C_3	0.087566	0.930221
N_4C_3 vs. N_8C_3	0.087566	0.930221

Table 53 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de l'espace couleur HSV.

Espace couleur E_2		
Comparaison	Friedman	p-value
N_8C_4 vs. N_1C_1	16.7656	0
N_8C_4 vs. N_7C_4	15.721539	0
N_2C_1 vs. N_1C_1	15.007536	0
N_8C_4 vs. N_3C_4	14.65727	0
N_2C_4 vs. N_1C_1	14.603383	0
N_8C_3 vs. N_1C_1	14.266589	0
N_2C_2 vs. N_1C_1	14.091457	0
N_2C_1 vs. N_7C_4	13.963475	0
N_2C_4 vs. N_7C_4	13.559322	0
N_4C_4 vs. N_1C_1	13.485228	0
.	.	.
.	.	.
.	.	.
N_2C_2 vs. N_8C_3	0.175133	0.860975
N_4C_4 vs. N_9C_4	0.033679	0.973133

Table 54 – Comparaison multiple de la segmentation du *noyau* en fonction de l'espace couleur I1I2I3.

Espace couleur E_2		
Comparaison	Friedman	p-value
N_2C_2 vs. N_9C_1	8.31438	0
N_2C_2 vs. N_2C_4	8.271778	0
N_6C_4 vs. N_9C_1	8.158174	0
N_6C_4 vs. N_2C_4	8.115573	0
N_2C_2 vs. N_7C_3	7.810263	0
N_6C_4 vs. N_7C_3	7.654057	0
N_2C_2 vs. N_2C_1	7.391349	0
N_6C_4 vs. N_2C_1	7.235143	0
N_2C_2 vs. N_1C_1	7.078938	0
N_2C_2 vs. N_3C_1	7.071838	0
.	.	.
.	.	.
.	.	.
N_2C_4 vs. N_9C_1	0.042601	0.966019
N_1C_1 vs. N_3C_1	0.0071	0.994335

Table 55 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de l'espace couleur I1I2I3.

Espace couleur E_3		
Comparaison	Friedman	p-value
N_8C_2 vs. N_9C_4	15.819037	0
N_1C_2 vs. N_9C_4	15.643918	0
N_2C_2 vs. N_9C_4	14.943444	0
N_8C_4 vs. N_9C_4	13.635893	0
N_4C_2 vs. N_9C_4	13.128049	0
N_8C_2 vs. N_2C_4	13.063839	0
N_1C_2 vs. N_2C_4	12.888721	0
N_2C_2 vs. N_2C_4	12.188247	0
N_9C_2 vs. N_9C_4	12.164898	0
N_9C_1 vs. N_9C_4	11.978105	0
.	.	.
.	.	.
.	.	.
N_3C_2 vs. N_6C_3	0.023349	0.981372
N_6C_4 vs. N_9C_3	0	1

Table 56 – Comparaison multiple de la segmentation du *noyau* en fonction de l'espace couleur Lab.

Espace couleur E_3		
Comparaison	Friedman	p-value
N_6C_2 vs. N_9C_1	14.097038	0
N_6C_2 vs. N_3C_1	14.097038	0
N_6C_2 vs. N_8C_4	13.781825	0
N_1C_2 vs. N_9C_1	13.583357	0
N_1C_2 vs. N_3C_1	13.583357	0
N_6C_2 vs. N_2C_4	13.402402	0
N_1C_2 vs. N_8C_4	13.268144	0
N_6C_2 vs. N_2C_1	13.244795	0
N_4C_2 vs. N_9C_1	13.011304	0
N_4C_2 vs. N_3C_1	13.011304	0
.	.	.
.	.	.
.	.	.
N_3C_2 vs. N_4C_3	0.06421	0.948803
N_3C_1 vs. N_9C_1	0	1

Table 57 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de l'espace couleur Lab.

Espace couleur E_4		
Comparaison	Friedman	p-value
N_1C_2 vs. N_3C_1	11.191556	0
N_1C_2 vs. N_8C_1	10.934981	0
N_2C_2 vs. N_3C_1	10.049187	0
N_2C_2 vs. N_8C_1	9.792612	0
N_1C_1 vs. N_3C_1	9.426076	0
N_2C_1 vs. N_3C_1	9.413858	0
N_1C_1 vs. N_8C_1	9.169501	0
N_2C_1 vs. N_8C_1	9.157283	0
N_1C_2 vs. N_4C_1	8.124874	0
N_1C_2 vs. N_6C_1	7.92328	0
.	.	.
.	.	.
.	.	.
N_6C_3 vs. N_9C_3	0	1
N_4C_2 vs. N_6C_2	0	1

Table 58 – Comparaison multiple de la segmentation du *noyau* en fonction de l'espace couleur RGB.

Espace couleur E_4		
Comparaison	Friedman	p-value
N_1C_2 vs. N_6C_4	13.989445	0
N_1C_2 vs. N_8C_1	13.653454	0
N_1C_2 vs. N_4C_4	13.519058	0
N_2C_2 vs. N_6C_4	12.926492	0
N_6C_2 vs. N_6C_4	12.755442	0
N_1C_2 vs. N_6C_1	12.755442	0
N_2C_2 vs. N_8C_1	12.590501	0
N_2C_2 vs. N_4C_4	12.456104	0
N_6C_2 vs. N_8C_1	12.419451	0
N_6C_2 vs. N_4C_4	12.285054	0
.	.	.
.	.	.
.	.	.
N_1C_3 vs. N_6C_3	0.128287	0.897921
N_4C_3 vs. N_9C_3	0.103852	0.917287

Table 59 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de l'espace couleur RGB.

Annexe C : Résultat non-paramétrique de la caractérisation super-pixellique

Dans cette annexe, nous présentons le détail de l'analyse non-paramétrique concernant la composante de couleur. Les tableaux 60, 62, 64 et 66 affichent les mesures de Friedman de chaque combinaison entre N_i et E_j pour la segmentation du *noyau*, et les tableaux 61, 63, 65 et 67 pour la segmentation du *cytoplasme*.

Caractérisation C_1		
Comparaison	Friedman	p-value
N_2E_4 vs. N_4E_1	15.351406	0
N_1E_4 vs. N_4E_1	15.082308	0
N_2E_4 vs. N_1E_1	14.979795	0
N_1E_4 vs. N_1E_1	14.710697	0
N_2E_2 vs. N_4E_1	14.012323	0
N_2E_4 vs. N_3E_1	13.877774	0
N_2E_2 vs. N_1E_1	13.640711	0
N_1E_4 vs. N_3E_1	13.608676	0
N_2E_2 vs. N_3E_1	12.538691	0
N_2E_4 vs. N_4E_2	12.308035	0
.	.	.
.	.	.
.	.	.
N_2E_3 vs. N_8E_4	0.04485	0.964227
N_6E_4 vs. N_9E_3	0.038443	0.969335

Table 60 – Comparaison multiple de la segmentation du *noyau* en fonction de la caractérisation CM.

Caractérisation C_1		
Comparaison	Friedman	p-value
N_4E_4 vs. N_1E_3	12.095678	0
N_4E_3 vs. N_1E_3	11.765796	0
N_1E_4 vs. N_1E_3	11.539765	0
N_2E_4 vs. N_1E_3	11.509221	0
N_9E_4 vs. N_1E_3	10.904437	0
N_4E_4 vs. N_9E_3	10.305762	0
N_4E_3 vs. N_9E_3	9.97588	0
N_4E_4 vs. N_9E_2	9.902572	0
N_1E_4 vs. N_9E_3	9.749849	0
N_4E_4 vs. N_3E_3	9.731523	0
.	.	.
.	.	.
.	.	.
N_3E_1 vs. N_3E_3	0.067198	0.946424
N_1E_4 vs. N_2E_4	0.030545	0.975633

Table 61 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la caractérisation CM.

Caractérisation C_2		
Comparaison	Friedman	p-value
N_1E_4 vs. N_4E_1	13.354117	0
N_2E_4 vs. N_4E_1	12.889838	0
N_1E_4 vs. N_9E_1	12.682135	0
N_8E_2 vs. N_4E_1	12.65159	0
N_9E_4 vs. N_4E_1	12.639372	0
N_6E_4 vs. N_4E_1	12.443887	0
N_8E_1 vs. N_4E_1	12.297272	0
N_2E_4 vs. N_9E_1	12.217856	0
N_4E_4 vs. N_4E_1	12.132331	0
N_8E_2 vs. N_9E_1	11.979608	0
.	.	.
.	.	.
.	.	.
N_2E_1 vs. N_4E_3	0.122179	0.902758
N_8E_2 vs. N_9E_4	0.012218	0.990252

Table 62 – Comparaison multiple de la segmentation du *noyau* en fonction de la caractérisation FOS.

Caractérisation C_2		
Comparaison	Friedman	p-value
N_1E_4 vs. N_3E_1	8.914093	0
N_6E_3 vs. N_3E_1	8.908504	0
N_6E_3 vs. N_3E_3	8.774373	0
N_1E_4 vs. N_3E_3	8.768785	0
N_6E_3 vs. N_4E_2	8.640243	0
N_1E_4 vs. N_4E_2	8.634654	0
N_6E_4 vs. N_3E_1	8.595533	0
N_6E_3 vs. N_7E_3	8.595533	0
N_1E_4 vs. N_7E_3	8.589944	0
N_6E_4 vs. N_3E_3	8.455813	0
.	.	.
.	.	.
.	.	.
N_1E_3 vs. N_6E_1	0.022355	0.982165
N_1E_4 vs. N_6E_3	0.005589	0.995541

Table 63 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la caractérisation FOS.

Caractérisation C_3		
Comparaison	Friedman	p-value
N_1E_4 vs. N_4E_1	13.354117	0
N_2E_4 vs. N_4E_1	12.889838	0
N_1E_4 vs. N_9E_1	12.682135	0
N_8E_2 vs. N_4E_1	12.65159	0
N_9E_4 vs. N_4E_1	12.639372	0
N_6E_4 vs. N_4E_1	12.443887	0
N_8E_1 vs. N_4E_1	12.297272	0
N_2E_4 vs. N_9E_1	12.217856	0
N_4E_4 vs. N_4E_1	12.132331	0
N_8E_2 vs. N_9E_1	11.979608	0
.	.	.
.	.	.
.	.	.
N_2E_1 vs. N_4E_3	0.122179	0.902758
N_8E_2 vs. N_9E_4	0.012218	0.990252

Table 64 – Comparaison multiple de la segmentation du *noyau* en fonction de la caractérisation Contraste.

Caractérisation C_3		
Comparaison	Friedman	p-value
N_6E_3 vs. N_3E_1	8.914093	0
N_1E_4 vs. N_3E_1	8.908504	0
N_6E_3 vs. N_3E_3	8.774373	0
N_1E_4 vs. N_3E_3	8.768785	0
N_6E_3 vs. N_4E_2	8.640243	0
N_1E_4 vs. N_4E_2	8.634654	0
N_6E_4 vs. N_3E_1	8.595533	0
N_6E_3 vs. N_7E_3	8.595533	0
N_1E_4 vs. N_7E_3	8.589944	0
N_6E_4 vs. N_3E_3	8.455813	0
.	.	.
.	.	.
.	.	.
N_1E_3 vs. N_6E_1	0.022355	0.982165
N_1E_4 vs. N_6E_3	0.005589	0.995541

Table 65 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la caractérisation Contraste.

Caractérisation C_4		
Comparaison	Friedman	p-value
N_8E_2 vs. N_4E_1	13.10976	0
N_8E_2 vs. N_6E_1	12.352253	0
N_8E_2 vs. N_1E_3	12.211747	0
N_8E_2 vs. N_2E_3	11.765796	0
N_8E_2 vs. N_1E_1	11.387042	0
N_8E_2 vs. N_8E_1	11.246537	0
N_2E_2 vs. N_4E_1	11.05716	0
N_8E_3 vs. N_4E_1	10.513465	0
N_2E_2 vs. N_6E_1	10.299653	0
N_8E_2 vs. N_8E_4	10.165256	0
.	.	.
.	.	.
.	.	.
N_3E_4 vs. N_8E_4	0.024436	0.980505
N_1E_2 vs. N_3E_3	0.012218	0.990252

Table 66 – Comparaison multiple de la segmentation du *noyau* en fonction de la caractérisation Hu.

Caractérisation C_4		
Comparaison	Friedman	p-value
N_6E_2 vs. N_7E_4	11.551983	0
N_6E_2 vs. N_9E_1	11.051051	0
N_6E_2 vs. N_6E_3	10.678406	0
N_6E_2 vs. N_2E_4	10.342415	0
N_9E_2 vs. N_7E_4	10.049187	0
N_6E_2 vs. N_1E_1	9.945335	0
N_9E_2 vs. N_9E_1	9.548255	0
N_9E_2 vs. N_6E_3	9.17561	0
N_6E_1 vs. N_7E_4	9.145065	0
N_3E_2 vs. N_7E_4	9.065649	0
.	.	.
.	.	.
.	.	.
N_3E_2 vs. N_6E_1	0.079416	0.936702
N_6E_4 vs. N_8E_1	0.05498	0.956154

Table 67 – Comparaison multiple de la segmentation du *cytoplasme* en fonction de la caractérisation Hu.

Bibliographie

[1] Nicolas Vandenbroucke, Laurent Busin, and Ludovic Macaire, ''Unsupervised color-image segmentation by multicolor space iterative pixel classification,'' *Journal of Electronic Imaging*, vol. 24, no. 2, pp. 023032–023032, 2015.

[2] Michael Goebel and Le Gruenwald, ''A survey of data mining and knowledge discovery software tools,'' *SIGKDD Explor. Newsl.*, vol. 1, no. 1, pp. 20–33, June 1999.

[3] Anil Jain and Douglas Zongker, ''Feature selection : Evaluation, application, and small sample performance,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[4] Fabrice Muhlenbach, Stane Lallich, and Djamel A. Zighed, ''Identifying and handling mislabelled instances.,'' *J. Intell. Inf. Syst.*, vol. 22, no. 1, pp. 89–109, 2004.

[5] Isaac Triguero, José A. Sáez, Julián Luengo, Salvador García, and Francisco Herrera, ''On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification,'' *Neurocomputing*, vol. 132, pp. 30–41, 2014.

[6] Ming Li and Zhi-Hua Zhou, ''Setred : Self-training with editing.,'' in *PAKDD*, Tu Bao Ho, David Wai-Lok Cheung, and Huan Liu, Eds. 2005, vol. 3518 of *Lecture Notes in Computer Science*, pp. 611–621, Springer.

[7] Yu Wang, Xiaoyan Xu, Haifeng Zhao, and Zhongsheng Hua, ''Semi-supervised learning based on nearest neighbor rule and cut edges,'' *Know.-Based Syst.*, vol. 23, no. 6, pp. 547–554, Aug. 2010.

[8] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, ''Slic superpixels compared to state-of-the-art superpixel methods,'' *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[9] Leo Breiman, ''Random forests,'' *Machine Learning*, vol. 45, pp. 5–32, 2001.

[10] Isaac Triguero, Salvador García, and Francisco Herrera, ''Self-labeled techniques for semi-supervised learning : taxonomy, software and empirical study,'' *Knowledge and Information Systems, (in press)*, pp. 1–40, 2014.

[11] E Cernadas, M Fernández-Delgado, E González-Rufino, and P Carrión, ''Influence of normalization and color space to color texture classification,'' *Pattern Recognition*, vol. 61, pp. 120–138, 2017.

[12] Encarnación González-Rufino, Pilar Carrión, Eva Cernadas, Manuel Fernández-Delgado, and Rosario Domínguez-Petit, ''Exhaustive comparison of colour texture features and classification methods to discriminate cells categories in histological images of fish ovary,'' *Pattern Recognition*, vol. 46, no. 9, pp. 2391–2407, 2013.

[13] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard : Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. IEEE, 1992, vol. 1, pp. 517–520.

[14] David Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 1995, ACL '95, pp. 189–196, Association for Computational Linguistics.

[15] David McClosky, Eugene Charniak, and Mark Johnson, "Effective self-training for parsing," in *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 152–159.

[16] Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, New York, NY, USA, 1998, COLT' 98, pp. 92–100.

[17] Leslie G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, Nov. 1984.

[18] Sally A. Goldman and Yan Zhou, "Enhancing supervised learning with unlabeled data," in *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, CA, USA, 2000, ICML '00, pp. 327–334, Morgan Kaufmann Publishers Inc.

[19] Zhi-Hua Zhou and Ming Li, "Tri-training : Exploiting unlabeled data using three classifiers," *IEEE Trans. on Knowl. and Data Eng.*, vol. 17, no. 11, pp. 1529–1541, Nov. 2005.

[20] Ming Li and Zhi-Hua Zhou, "Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples," *Trans. Sys. Man Cyber. Part A*, vol. 37, no. 6, pp. 1088–1098, Nov. 2007.

[21] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine learning*, vol. 39, no. 2-3, pp. 103–134, 2000.

[22] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell, "Text classification from labeled and unlabeled documents using em," *Mach. Learn.*, vol. 39, no. 2-3, pp. 103–134, May 2000.

[23] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[24] Behzad M Shahshahani and David A Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon," *IEEE Transactions on Geoscience and remote sensing*, vol. 32, no. 5, pp. 1087–1095, 1994.

[25] Masashi Inoue and Naonori Ueda, "Exploitation of unlabeled sequences in hidden markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1570–1581, 2003.

[26] Fabio Gagliardi Cozman, Ira Cohen, and M Cirelo, "Unlabeled data can degrade classification performance of generative classifiers.," in *Flairs conference*, 2002, pp. 327–331.

[27] Xiaojin Zhu, "Semi-supervised learning literature survey," 2005.

[28] Avrim Blum and Shuchi Chawla, "Learning from labeled and unlabeled data using graph mincuts," *In : Proc. of the 18th International Conference on Machine Learning (ICML 2001)*, p. 1926, 2001.

[29] Avrim Blum, John Lafferty, Mugizi Robert Rwebangira, and Rajashekar Reddy, "Semi-supervised learning using randomized mincuts," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 13.

[30] Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al., "Semi-supervised learning using gaussian fields and harmonic functions," in *In : Proc. of the 20th International Conference on Machine Learning (ICML 2003)*, 2003, vol. 3, pp. 912–919.

[31] Kristin P. Bennett and Ayhan Demiriz, "Semi-supervised support vector machines," in *Advances in Neural Information Processing Systems*. 1998, pp. 368–374, MIT Press.

[32] Yuanqing Li, Cuntai Guan, Huiqi Li, and Zhengyang Chin, "A self-training semi-supervised svm algorithm and its application in an eeg-based brain computer interface speller system," *Pattern Recogn. Lett.*, vol. 29, no. 9, pp. 1285–1294, July 2008.

[33] Thorsten Joachims, "Transductive inference for text classification using support vector machines," in *In : Proc. of the 16th International Conference on Machine Learning*, 1999, vol. 99, pp. 200–209.

[34] Mohamed Farouk Abdel Hady and Friedhelm Schwenker, "Combining committee-based semi-supervised learning and active learning," *Journal of Computer Science and Technology*, vol. 25, no. 4, pp. 681–698, 2010.

[35] Mohamed Farouk Abdel Hady, Friedhelm Schwenker, and Günther Palm, "Semi-supervised learning for tree-structured ensembles of rbf networks with co-training," *Neural Networks*, vol. 23, no. 4, pp. 497–509, 2010.

[36] Olivier Chapelle and Alexander Zien, "Semi-supervised classification by low density separation.," in *In : Proc. of the 10th International Workshop on Artificial Intelligence and Statistics*, 2005, pp. 57–64.

[37] Olivier Chapelle, Mingmin Chi, and Alexander Zien, "A continuation method for semi-supervised svms," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 185–192.

[38] Tijl De Bie and Nello Cristianini, "Semi-supervised learning using semi-definite programming," *Semi-supervised learning. MIT Press, Cambridge-Massachussets*, vol. 32, 2006.

[39] Mathias M Adankon and Mohamed Cheriet, "Genetic algorithm–based training for semi-supervised svm," *Neural Computing and Applications*, vol. 19, no. 8, pp. 1197–1206, 2010.

[40] Ian H. Witten, Eibe Frank, and Mark A. Hall, *Data Mining : Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[41] Ming Li and Zhi-Hua Zhou, "Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples," *IEEE Transactions on Systems, Man, and Cybernetics-Part A : Systems and Humans*, vol. 37, no. 6, pp. 1088–1098, 2007.

[42] Thomas Kemp and Alex Waibel, "Unsupervised training of a speech recognizer : recent experiments.," in *Eurospeech*, 1999.

[43] Svetlana Kiritchenko and Stan Matwin, "Email classification with co-training," in *Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research*. IBM Corp., 2011, pp. 301–312.

[44] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien, "Semi-supervised learning (chapelle, o. et al., eds. ; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[45] Weizhong Zhao, Qing He, Huifang Ma, and Zhongzhi Shi, "Effective semi-supervised document clustering via active learning with instance-level constraints," *Knowledge and information systems*, vol. 30, no. 3, pp. 569–587, 2012.

[46] M. A. BECHAR, N. SETTOUTI, and MA. CHIKH, "L'impact de la mesure de similarit auto-apprentissage," in *Colloque sur l'Optimisation et les Systs d'Information COSI'2015*, 1 au 3 juin 2015.

[47] M. A. BECHAR, N. SETTOUTI, MA. CHIKH, and M. ADEL, ''Reinforced con ?dence in self-training (r-coset) for a semi-supervised medical data classi ?cation,'' *Int. J. Applied Pattern Recognition*, vol. 4, pp. pp–pp, 2017.

[48] Zhiliang Liu, Xiaomin Zhao, Jianxiao Zou, and Hongbing Xu, ''A semi-supervised approach based on k-nearest neighbor.,'' *JSW*, vol. 8, no. 4, pp. 768–775, 2013.

[49] S Cheng, Q. Huang, J. Liu, and X. Tang, ''Costra : Confidence-based self-training,'' *Journal of Computational Information Systems*, vol. 9 : 24, pp. 9761–9769, 2013.

[50] Yun Jin, Chengwei Huang, and Li Zhao, ''A semi-supervised learning algorithm based on modified self-training svm,'' *Journal of Computers*, vol. 6, no. 7, 2011.

[51] Anindya Halder, Susmita Ghosh, and Ashish Ghosh, ''Ant based semi-supervised classification,'' in *Swarm Intelligence : Lecture Notes in Computer Science : 7th International Conference, ANTS 2010, Brussels, Belgium, September 8-10, 2010. Proceedings, Volume 6234, pp 376-383*, 2010.

[52] Mohammed Abdul Wajeed and T. Adilakshmi., ''Different similarity measures in semi-supervised text classification,'' in *India Conference (INDICON), 2011 Annual IEEE*, 16-18 Dec. 2011.

[53] G. Penney, J. Weese, J. Little, P. Desmedt, D. Hill, and D. Hawkes, ''A comparison of similarity measures for use in 2-d-3-d medical image registration,'' *Medical Imaging, IEEE Transactions on*, vol. 17, pp. 586–595, 1998.

[54] K. Chang, R. Lee, C. Wen, and M. Yeh, ''Comparison of similarity measures for clustering electrocardiogram complexes,'' *Computers in Cardiology*, pp. 759–762, 2005.

[55] D. Yong, S. WenKang, Z. ZhenFu, and L & Qi, ''Combining belief functions based on distance of evidence,'' *Decision Support Systems*, vol. 38, pp. 489–493, 2004.

[56] Maria Rifqi, ''Mesures de similarité, raisonnement et modélisation de l'utilisateur,'' Tech. Rep., HABILITATION A DIRIGER DES RECHERCHES DE L'UNIVERSITÉ PIERRE ET MARIE CURIE, 2010.

[57] Evelyn Fix and Jr, ''Discriminatory analysis : Nonparametric discrimination : Consistency properties,'' Tech. Rep. Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, Randolf Field, Texas, 1951.

[58] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

[59] Claude Grasland, *IINITIATION AUX METHODES STATISTIQUES EN SCIENCES SOCIALES*, Université Paris VII / UFR GHSS /, 2000.

[60] A. Dragut, *Data mining : streams et similarités*, Université Aix-Marseille, I.U.T.d'Aix en Provence - Département Informatique, créé le 28/10/2011 -derniére mise á jour : 7/11/2012 edition, 2012.

[61] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, ''Uci repository of machine learning databases,'' 1998.

[62] Reza Zafarani and Huan Liu, ''Asu repository of social computing databases,'' 1998.

[63] Bin Wang, Bruce Spencer, Charles X Ling, and Harry Zhang, ''Semi-supervised self-training for sentence subjectivity classification,'' in *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 2008, pp. 344–355.

[64] D. Randall Wilson and Tony R. Martinez, ''Improved heterogeneous distance functions,'' *Journal of artificial intelligence research*, vol. 6, pp. 1–34, 1997.

[65] Nikos Fazakis, Stamatis Karlos, Sotiris Kotsiantis, and Kyriakos Sgarbas, ''Self-trained lmt for semisupervised learning,'' *Computational intelligence and neuroscience*, vol. 2016, pp. 10, 2016.

[66] Niels Landwehr, Mark Hall, and Eibe Frank, ''Logistic model trees,'' *Machine Learning*, vol. 59, no. 1-2, pp. 161–205, 2005.

[67] Marc Sumner, Eibe Frank, and Mark Hall, ''Speeding up logistic model tree induction,'' in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2005, pp. 675–683.

[68] E. Fix and J. L. Hodges, ''Discriminatory analysis. nonparametric discrimination : consistency properties,'' *Jr. U.S. Air Force Sch. Aviation Medicine, Randolf Field*, vol. Rep.4, ., pp. Project 21–49–004, Contract AF 41 (128)–31, 1951.

[69] Godfried T. Toussaint, ''The relative neighbourhood graph of a finite planar set,'' *Pattern Recognition*, vol. 12, pp. 261–268, 1980.

[70] Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera, ''A study of statistical techniques and performance measures for genetics-based machine learning : accuracy and interpretability,'' *Soft Comput.*, vol. 13, no. 10, pp. 959–977, 2009.

[71] Janez Demsar, ''Statistical comparisons of classifiers over multiple data sets,'' *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

[72] Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera, ''Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining : Experimental analysis of power,'' *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, May 2010.

[73] Xingquan Zhu, Xindong Wu, and Qijun Chen, ''Eliminating class noise in large datasets,'' in *ICML*, 2003, vol. 3, pp. 920–927.

[74] Humayun Irshad, Antoine Veillard, Ludovic Roux, and Daniel Racoceanu, ''Methods for nuclei detection, segmentation, and classification in digital histopathology : a reviewcurrent status and future potential,'' *IEEE reviews in biomedical engineering*, vol. 7, pp. 97–114, 2014.

[75] Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis, ''A survey of skin-color modeling and detection methods,'' *Pattern recognition*, vol. 40, no. 3, pp. 1106–1122, 2007.

[76] Lili Zhao, Kuan Li, Mao Wang, Jianping Yin, En Zhu, Chengkun Wu, Siqi Wang, and Chengzhang Zhu, ''Automatic cytoplasm and nuclei segmentation for color cervical smear image using an efficient gap-search mrf,'' *Computers in biology and medicine*, vol. 71, pp. 46–56, 2016.

[77] Anant Madabhushi, ''Digital pathology image analysis : opportunities and challenges,'' *Imaging in Medicine*, vol. 1, no. 1, pp. 7–10, 2009.

[78] Jadwiga Rogowska, ''Overview and fundamentals of medical image segmentation,'' in *Handbook of medical imaging*. Academic Press, Inc., 2000, pp. 69–85.

[79] Mehmet Sezgin et al., ''Survey over image thresholding techniques and quantitative performance evaluation,'' *Journal of Electronic imaging*, vol. 13, no. 1, pp. 146–168, 2004.

[80] PC Chen and T Pavlidis, ''Image segmentation as an estimation problem,'' *Computer Graphics and Image Processing*, vol. 12, no. 2, pp. 153–172, 1980.

[81] Yu Xiaohan, Juha Yla-Jaaski, Olli Huttunen, T Vehkomaki, Outi Sipila, and Toivo Katila, ''Image segmentation combining region growing and edge detection,'' in *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol. III. Conference C : Image, Speech and Signal Analysis,*. IEEE, 1992, pp. 481–484.

[82] Luc Vincent and Pierre Soille, ''Watersheds in digital spaces : an efficient algorithm based on immersion simulations,'' *IEEE transactions on pattern analysis and machine intelligence*, vol. 13, no. 6, pp. 583–598, 1991.

[83] Rolf Adams and Leanne Bischof, ''Seeded region growing,'' *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[84] Laurent Najman and Michel Schmitt, ''Geodesic saliency of watershed contours and hierarchical segmentation,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 12, pp. 1163–1173, 1996.

[85] Ravi Malladi, James A Sethian, and Baba C Vemuri, ''Shape modeling with front propagation : A level set approach,'' *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 2, pp. 158–175, 1995.

[86] Zhenyu Wu and Richard Leahy, ''An optimal graph theoretic approach to data clustering : Theory and its application to image segmentation,'' *IEEE transactions on pattern analysis and machine intelligence*, vol. 15, no. 11, pp. 1101–1113, 1993.

[87] Jianbo Shi and Jitendra Malik, ''Normalized cuts and image segmentation,'' *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[88] Sang Ho Park, Il Dong Yun, and Sang Uk Lee, ''Color image segmentation based on 3-d clustering : morphological approachfn1fn1the authors are with the signal processing lab., school of electrical eng., seoul national university, shinlim-dong, kwanak-gu, seoul 151-742, south korea, and are also affiliated with the automation systems research institute (asri) of seoul national university, seoul 151-742, south korea.,'' *Pattern Recognition*, vol. 31, no. 8, pp. 1061–1076, 1998.

[89] Thrasyvoulos N Pappas, ''An adaptive clustering algorithm for image segmentation,'' *IEEE Transactions on signal processing*, vol. 40, no. 4, pp. 901–914, 1992.

[90] Nicolas Vandenbroucke, Ludovic Macaire, and Jack-Gérard Postaire, ''Color image segmentation by pixel classification in an adapted hybrid color space. application to soccer image analysis,'' *Computer Vision and Image Understanding*, vol. 90, no. 2, pp. 190–216, 2003.

[91] Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai, ''Skin segmentation using color pixel classification : analysis and comparison,'' *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 1, pp. 148–154, 2005.

[92] Durga P. Panda and Azriel Rosenfeld, ''Image segmentation by pixel classification in (gray level, edge value) space,'' *IEEE transactions on computers*, vol. 100, no. 9, pp. 875–879, 1978.

[93] Raimondo Schettini, ''A segmentation algorithm for color images,'' *Pattern Recognition Letters*, vol. 14, no. 6, pp. 499 – 506, 1993.

[94] Patrick Lambert and L Macaire, ''Filtering and segmentation : the specificity of color images,'' *International Conference on Color in Graphics and Image Processing*, vol. 1, pp. 57–71, 2000.

[95] James. MacQueen, ''Some methods for classification and analysis of multivariate observations,'' in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. Le Cam and J. Neyman, Eds. 1967, vol. 1, pp. 281–297, University of California Press.

[96] Young Won Lim and Sang Uk Lee, ''On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques.,'' *Pattern Recognition*, vol. 23, no. 9, pp. 935–952, 1990.

[97] Jean-Pierre Cocquerez and Sylvie Philipp, *Analyse d'images : filtrage et segmentation*, Masson, 1997.

[98] Walter D. Fisher, ''On grouping for maximum homogeneity,'' *Journal of the American Statistical Association*, vol. 53, no. 284, 1958.

[99] Ronald A. Fisher, ''The use of multiple measurements in taxonomic problems,'' *Annals of Eugenics*, vol. 7, no. 7, pp. 179–188, 1936.

[100] Xiaofeng Ren and Jitendra Malik, ''Learning a classification model for segmentation,'' in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 10–17.

[101] Derek Hoiem, Alexei A Efros, and Martial Hebert, ''Automatic photo pop-up,'' *ACM transactions on graphics (TOG)*, vol. 24, no. 3, pp. 577–584, 2005.

[102] Yin Li, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum, ''Lazy snapping,'' in *ACM Transactions on Graphics (ToG)*. ACM, 2004, vol. 23, pp. 303–308.

[103] Xuming He, Richard S Zemel, and Debajyoti Ray, "Learning and incorporating top-down cues in image segmentation," in *European conference on computer vision*. Springer, 2006, pp. 338–351.

[104] Alex Levinshtein, Sven J Dickinson, and Cristian Sminchisescu, "Multiscale symmetric part detection and grouping.," in *ICCV*, 2009, pp. 2162–2169.

[105] Brian Fulkerson, Andrea Vedaldi, Stefano Soatto, et al., "Class segmentation and object localization with superpixel neighborhoods.," in *ICCV*. Citeseer, 2009, vol. 9, pp. 670–677.

[106] Pedro F Felzenszwalb and Daniel P Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[107] Alex Levinshtein, Adrian Stere, Kiriakos N Kutulakos, David J Fleet, Sven J Dickinson, and Kaleem Siddiqi, "Turbopixels : Fast superpixels using geometric flows," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2290–2297, 2009.

[108] Andrea Vedaldi and Stefano Soatto, "Quick shift and kernel methods for mode seeking," in *European Conference on Computer Vision*. Springer, 2008, pp. 705–718.

[109] Olga Veksler, Yuri Boykov, and Paria Mehrani, "Superpixels and supervoxels in an energy optimization framework," in *European conference on Computer vision*. Springer, 2010, pp. 211–224.

[110] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[111] Pedro F Felzenszwalb and Daniel P Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[112] Marc Ebner, *Color constancy*, vol. 6, John Wiley & Sons, 2007.

[113] Graham D Finlayson, Bernt Schiele, and James L Crowley, "Comprehensive colour image normalization," in *European conference on computer vision*. Springer, 1998, pp. 475–490.

[114] Nicolas Limare, Jose-Luis Lisani, Jean-Michel Morel, Ana Belén Petro, and Catalina Sbert, "Simplest color balance," *Image Processing On Line*, vol. 1, pp. 297–315, 2011.

[115] Alice Porebski, Nicolas Vandenbroucke, and Ludovic Macaire, "Selection of color texture features from reduced size chromatic co-occurrence matrices," in *Signal and Image Processing Applications (ICSIPA), 2009 IEEE International Conference on*. IEEE, 2009, pp. 273–278.

[116] Xiaoqing Liu, *Analyse d'images couleur en composantes indépendantes par réseau de neurones*, Ph.D. thesis, 1991.

[117] David T Berry, "Colour recognition using spectral signatures," *Pattern Recognition Letters*, vol. 6, no. 1, pp. 69–75, 1987.

[118] JOEL POKORNY and VIVIANNEC SMITH, "Colorimetry and color discrimination," *Handbook of perception and human performance.*, vol. 1, pp. 8–1, 1986.

[119] Choquet C. Garbay C, Brugal G, "Application of colored image analysis to bone marrow cell recognition," *Anal. Quant. Cytol.*, vol. 3(4), pp. 272–280, 1981.

[120] Michael J Swain and Dana H Ballard, "Color indexing," *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991.

[121] RL Swenson and Keith R Dimond, "A universal colour transformation architecture," *Pattern Recognition Letters*, vol. 19, no. 9, pp. 805–813, 1998.

[122] Patrick Lambert and Thierry Carron, "Symbolic fusion of luminance-hue-chroma features for region segmentation," *Pattern recognition*, vol. 32, no. 11, pp. 1857–1872, 1999.

[123] Yu-Ichi Ohta, Takeo Kanade, and Toshiyuki Sakai, "Color information for region segmentation," *Computer graphics and image processing*, vol. 13, no. 3, pp. 222–241, 1980.

[124] Tian-Yuan Shih, "The reversibility of six geometric color spaces," *Photogrammetric Engineering and Remote Sensing*, vol. 61, no. 10, pp. 1223–1232, 1995.

[125] Allan Hanbury, "A 3d-polar coordinate colour representation well adapted to image analysis," in *Scandinavian Conference on Image Analysis*. Springer, 2003, pp. 804–811.

[126] Young Won Lim and Sang Uk Lee, "On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques," *Pattern recognition*, vol. 23, no. 9, pp. 935–952, 1990.

[127] Jianqing Liu and Yee-Hong Yang, "Multiresolution color image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 7, pp. 689–700, 1994.

[128] J-H Lee, B-H Chang, and S-D Kim, "Comparison of colour transformations for image segmentation," *Electronics Letters*, vol. 30, no. 20, pp. 1660–1661, 1994.

[129] Enno Littmann and Helge Ritter, "Adaptive color segmentation-a comparison of neural and statistical methods," *IEEE Transactions on neural networks*, vol. 8, no. 1, pp. 175–185, 1997.

[130] Jayanta Mukherjee, "Mrf clustering for segmentation of color images," *Pattern Recognition Letters*, vol. 23, no. 8, pp. 917–929, 2002.

[131] Sokratis Makrogiannis, George Economou, and Spiros Fotopoulos, "A region dissimilarity relation that combines feature-space and spatial information for color image segmentation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 35, no. 1, pp. 44–53, 2005.

[132] Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai, "Skin segmentation using color pixel classification : analysis and comparison," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 1, pp. 148–154, 2005.

[133] Olivier Lezoray, "Supervised automatic histogram clustering and watershed segmentation. application to microscopic medical color images," *Image Analysis & Stereology*, vol. 22, no. 2, pp. 113–120, 2011.

[134] Dina Khattab, Hala Mousher Ebied, Ashraf Saad Hussein, and Mohamed Fahmy Tolba, "Color image segmentation based on different color space models using automatic grabcut," *The Scientific World Journal*, vol. 2014, 2014.

[135] Rocio A Lizarraga-Morales, Raul E Sanchez-Yanez, Victor Ayala-Ramirez, and Alberto J Patlan-Rosales, "Improving a rough set theory-based segmentation approach using adaptable threshold selection and perceptual color spaces," *Journal of Electronic Imaging*, vol. 23, no. 1, pp. 013024–013024, 2014.

[136] Domingo Mery and Dieter Filbert, "Classification of potential defects in automated inspection of aluminium castings using statistical pattern recognition," in *Proceedings of 8th European Conference on Non-Destructive Testing (ECNDT 2002), Jun*, 2002, pp. 17–21.

[137] Domingo Mery, *Automated Flaw Detection in Castings from Digital Radioscopic Image Sequences*, Ph.D. thesis, Dr. Kster Verlag, Berlin, (Ph.D. Thesis in German),, 2001.

[138] Ming-Kuei Hu, "Visual pattern recognition by moment invariants," *IRE transactions on information theory*, vol. 8, no. 2, pp. 179–187, 1962.

[139] Michael Reed Teague, "Image analysis via the general theory of moments," *JOSA*, vol. 70, no. 8, pp. 920–930, 1980.

[140] James F Boyce and WJ Hossack, ''Moment invariants for pattern recognition,'' *Pattern Recognition Letters*, vol. 1, no. 5-6, pp. 451–456, 1983.

[141] Yaser S Abu-Mostafa and Demetri Psaltis, ''Recognitive aspects of moment invariants,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, , no. 6, pp. 698–706, 1984.

[142] Richard Szeliski, *Computer vision : algorithms and applications*, Springer Science & Business Media, 2010.

[143] Davide Boschetto, Hadis Mirzaei, Rupert WL Leong, and Enrico Grisan, ''Superpixel-based automatic segmentation of villi in confocal endomicroscopy,'' in *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on*. IEEE, 2016, pp. 168–171.

[144] Marina E Plissiti, Michalis Vrigkas, and Christophoros Nikou, ''Segmentation of cell clusters in pap smear images using intensity variation between superpixels,'' in *Systems, Signals and Image Processing (IWSSIP), 2015 International Conference on*. IEEE, 2015, pp. 184–187.

[145] Pierre Buyssens, Isabelle Gardin, and Su Ruan, ''Eikonal based region growing for superpixels generation : Application to semi-supervised real time organ segmentation in ct images,'' *IRBM*, vol. 35, no. 1, pp. 20–26, 2014.

[146] Azadeh Kianisarkaleh and Hassan Ghassemian, ''Nonparametric feature extraction for classification of hyperspectral images with limited training samples,'' *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 119, pp. 64–78, 2016.

[147] Shuyuan Yang, Yuan Lv, Yu Ren, and Licheng Jiao, ''Superpixel-wise semi-supervised structural sparse coding classifier for image segmentation,'' *Engineering Applications of Artificial Intelligence*, vol. 26, no. 10, pp. 2608–2612, 2013.

[148] Hang Su, Zhaozheng Yin, Seungil Huh, Takeo Kanade, and Jun Zhu, ''Interactive cell segmentation based on active and semi-supervised learning,'' *IEEE transactions on medical imaging*, vol. 35, no. 3, pp. 762–777, 2016.

[149] C Dupont, N Betrouni, N Reyns, and M Vermandel, ''On image segmentation methods applied to glioblastoma : state of art and new trends,'' *IRBM*, vol. 37, no. 3, pp. 131–143, 2016.

[150] Fuyong Xing and Lin Yang, ''Robust nucleus/cell detection and segmentation in digital pathology and microscopy images : a comprehensive review,'' *IEEE reviews in biomedical engineering*, vol. 9, pp. 234–263, 2016.

[151] Mourtada Benazzouz, Ismahan Baghli, and Med Amine Chikh, ''Microscopic image segmentation based on pixel classification and dimensionality reduction.,'' *Int. J. Imaging Systems and Technology*, vol. 23, no. 1, pp. 22–28, 2013.

[152] Ismahan Baghli, Amir Nakib, Elie Sellam, Mourtada Benazzouz, Amine Chikh, and Eric Petit, ''Hybrid framework based on evidence theory for blood cell image segmentation,'' in *Published in SPIE Proceedings Vol. 9038 : Medical Imaging 2014 : Biomedical Applications in Molecular, Structural, and Functional Imaging Robert C. Molthen ; John B. Weaver, Editor(s)*, 3 March 2014.

[153] Hui Kong, Metin Gurcan, and Kamel Belkacem-Boussaid, ''Partitioning histopathological images : an integrated framework for supervised color-texture segmentation and cell splitting,'' *IEEE transactions on medical imaging*, vol. 30, no. 9, pp. 1661–1677, 2011.

[154] Feng Zhou, Ju Fu Feng, and Qing Yun Shi, ''Texture feature based on local fourier transform,'' in *Image Processing, 2001. Proceedings. 2001 International Conference on*. IEEE, 2001, vol. 2, pp. 610–613.

[155] Yin Zhou, Hang Chang, Kenneth E Barner, and Bahram Parvin, ''Nuclei segmentation via sparsity constrained convolutional regression,'' in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 2015, pp. 1284–1287.

[156] Zhaozheng Yin, Ryoma Bise, Mei Chen, and Takeo Kanade, ''Cell segmentation in microscopy imagery using a bag of local bayesian classifiers,'' in *Biomedical Imaging : From Nano to Macro, 2010 IEEE International Symposium on*. IEEE, 2010, pp. 125–128.

[157] Thomas Janssens, Laura Antanas, Sarah Derde, Ilse Vanhorebeek, Greet Van den Berghe, and Fabian Güiza Grandas, ''Charisma : An integrated approach to automatic h&e-stained skeletal muscle cell segmentation using supervised learning and novel robust clump splitting,'' *Medical image analysis*, vol. 17, no. 8, pp. 1206–1219, 2013.

[158] Li Cheng, Ning Ye, Weimiao Yu, and Andre Cheah, ''Discriminative segmentation of microscopic cellular images,'' in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2011, pp. 637–644.

[159] Alex Levinshtein, Adrian Stere, Kiriakos N Kutulakos, David J Fleet, Sven J Dickinson, and Kaleem Siddiqi, ''Turbopixels : Fast superpixels using geometric flows,'' *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2290–2297, 2009.

[160] Shiyong Ji, Benzheng Wei, Zhen Yu, Gongping Yang, and Yilong Yin, ''A new multistage medical segmentation method based on superpixel and fuzzy clustering,'' *Computational and mathematical methods in medicine*, vol. 2014, 2014.

[161] Wei Wu, Albert YC Chen, Liang Zhao, and Jason J Corso, ''Brain tumor detection and segmentation in a crf (conditional random fields) framework with pixel-pairwise affinity and superpixel-level features,'' *International journal of computer assisted radiology and surgery*, vol. 9, no. 2, pp. 241–253, 2014.

[162] Tai Sing Lee, ''Image representation using 2d gabor wavelets,'' *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[163] E. Niaf, R. Flamary, A. Rakotomamonjy, O. Rouvi, and C. Lartizien, ''Svm with feature selection and smooth prediction in images : application to cad of prostate cancer,'' in *IEEE International Conference on Image Processing (ICIP)*, 2014.

[164] O. Cuisenaire N. Houhou S. Nedevschi J.P. Thiran A. Ciurte, X. Bresson and M. Bach Cuadra, ''Semi-supervised segmentation of ultrasound images based on patch representation and continuous min cut,'' *PLoS ONE*, vol. 9(7), 2014.

[165] Reza Azmi, Narges Norozi, Robab Anbiaee, Leila Salehi, and Azardokht Amirzadi, ''Impst : A new interactive self-training approach to segmentation suspicious lesions in breast mri,'' *Journal of Medical Signals and Sensors*, vol. 1 :2, pp. 138–148, 2011.

[166] David Yarowsky, ''Unsupervised word sense disambiguation rivaling supervised methods,'' in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 1995, ACL '95, pp. 189–196, Association for Computational Linguistics.

[167] Reza Azmi, Boshra Pishgoo, Narges Norozi, and Samira Yeganeh, ''Ensemble semi-supervised frame-work for brain magnetic resonance imaging tissue segmentation,'' *Journal of Medical Signals and Sensors*, vol. 3(2), pp. 94–106, 2013.

[168] Antoine Cornuéjols and Laurent Miclet, *Apprentissage artificiel : Concepts et algorithmes*, Eyrolles, June 2010.

[169] Leo Breiman, ''Bagging predictors,'' *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[170] Yoav Freund and Robert E. Schapire, ''Experiments with a new boosting algorithm,'' in *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, Lorenza Saitta, Ed. 1996, pp. 148–156, Morgan Kaufmann.

[171] Tin Kam Ho, ''The random subspace method for constructing decision forests,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[172] Thomas G. Dietterich and Ghulum Bakiri, ''Error-correcting output codes : A general method for improving multiclass inductive learning programs,'' in *IN PROCEEDINGS OF AAAI-91*. 1991, pp. 572–577, AAAI Press.

[173] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth Publishing Company, 1984.

[174] J. Ross Quinlan, ''Induction of decision trees,'' *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[175] J. Ross Quinlan, *C4.5 : Programs for Machine Learning*, Morgan Kaufmann, 1993.

[176] Yali Amit and Donald Geman, ''Shape quantization and recognition with randomized trees,'' *Neural Computation*, vol. 9, no. 7, pp. 1545–1588, 1997.

[177] N. Sirikulviriya and S. Sinthupinyo, ''Integration of rules from a random forest.,'' in *International Conference on Information and Electronics Engineering IPCSIT vol.6 (2011) IACSIT Press, Singapore*, 2011.

[178] Kamal Nigam and Rayid Ghani, ''Analyzing the effectiveness and applicability of co-training,'' in *Proceedings of the Ninth International Conference on Information and Knowledge Management*, New York, NY, USA, 2000, CIKM '00, pp. 86–93, ACM.

[179] Chao Deng and Maozu Guo, ''A new co-training-style random forest for computer aided diagnosis.,'' *J. Intell. Inf. Syst.*, vol. 36, no. 3, pp. 253–281, 2011.

[180] G. J. Mclachlan, ''Estimating the linear discriminant function from initial samples containing a small number of unclassified observations,'' *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 403–406, 1977.

[181] J. G. Fryer and C. A. Robertson, ''A comparison of some methods for estimating mixed normal distributions,'' *Biometrika*, vol. 59, pp. 639–648, 1972.

[182] A. P. Dempster, N. M. Laird, and D. B. Rubin, ''Maximum likelihood from incomplete data via the em algorithm,'' *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[183] Li Deng and Dong Yu, ''Deep learning : Methods and applications,'' Tech. Rep., May 2014.

[184] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, ''Semi-supervised learning with deep generative models,'' in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 3581–3589. Curran Associates, Inc., 2014.

[185] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert, ''Deep learning via semi-supervised embedding,'' in *Neural Networks : Tricks of the Trade*, pp. 639–655. Springer, 2012.

[186] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han, ''Decoupled deep neural network for semi-supervised semantic segmentation,'' in *Advances in Neural Information Processing Systems*, 2015, pp. 1495–1503.

[187] Dan Klein, Sepandar D. Kamvar, and Christopher D. Manning, ''From instance-level constraints to space-level constraints : Making the most of prior knowledge in data clustering,'' Technical Report 2002-10, Stanford InfoLab, February 2002.

[188] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall, ''Learning distance functions using equivalence relations,'' *ICML*, vol. 3, pp. 11–18, 2003.

[189] Mikhail Bilenko and Raymond J. Mooney, ''Adaptive duplicate detection using learnable string similarity measures,'' in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2003, KDD '03, pp. 39–48, ACM.

[190] Kristin Bennett and Ayhan Demiriz, ''Semi-supervised support vector machines,'' *Advances in Neural Information processing systems*, pp. 368–374, 1999.

[191] Janne Sinkkonen and Samuel Kaski, ''Clustering by similarity in an auxiliary space,'' in *Intelligent Data Engineering and Automated Learning - IDEAL 2000, Data Mining, Financial Engineering, and Intelligent Agents, Second International Conference, Shatin, N.T. Hong Kong, China, December 13-15, 2000, Proceedings*, 2000, pp. 3–8.

[192] Sugato Basu, Arindam Banerjee, and Raymond J. Mooney, ''Semi-supervised clustering by seeding,'' in *Proceedings of the Nineteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2002, ICML '02, pp. 27–34, Morgan Kaufmann Publishers Inc.

Résumé : Les données partiellement supervisées, c'est un effet qui reflète une véritable problématique concernant la difficulté d'étiquetage manuel des données. En classification supervisée des données médicales, l'hypothèse d'apprentissage nécessite une connaissance apriori sur les données où le médecin a apporté l'étiquette nécessaire. Néanmoins, face aux volumes de données disponibles actuellement, la supervision des données médicales est devenue une tâche fastidieuse pour le médecin et parfois même coûteuse dans certaines applications. De ce fait, les données non étiquetées sont plus nombreuses et disponibles par rapport aux données étiquetées. Cependant, sachant que la performance d'un classifieur est liée au nombre de données d'apprentissage, la principale question qui ressort est comment améliorer l'apprentissage d'un classifieur en intégrant des données non étiquetées à l'ensemble d'apprentissage. La technique d'apprentissage issue de la réponse à cette question est appelée l'apprentissage semi-supervisé.

Dans cette thèse, nous détaillons notre problématique majeure à savoir l'étiquetage automatique par apprentissage semi-supervisé en se basant sur le principe « d'auto-apprentissage ». L'auto-apprentissage est un algorithme de référence en classification semi supervisée, son usage est fondamental dans plusieurs applications. Dans l'auto-apprentissage (self-training), nous entrainons un classifieur supervisé avec les données étiquetées. Ensuite ce classifieur est utilisé pour prédire les étiquettes manquantes des données non étiquetées. Les données nouvellement étiquetées avec un haut degré de confiance sont ajoutées à la base étiquetée. Le classifieur est ré-entrainé sur les nouvelles données et cette procédure est répétée jusqu'à satisfaire un critère d'arrêt (convergence). Nous introduisons de manière progressive le concept d'auto-apprentissage dans des applications médicales. Une première partie dans cette thèse a été réservée pour la compréhension du principe d'auto-apprentissage par l'étude de l'algorithme ***SNNRCE***. Par la suite, nous détaillons notre contribution proposée au problème d'annotation des données médiales qui est portée sous le nom de ***R-COSET***.

Dans la dernière partie de cette thèse, nous nous intéressons plus particulièrement à la segmentation des images médicales utilisant les procédés de classification. La classification super-pixellique est devenue une méthode fréquente et importante dans la segmentation automatique. Une étude expérimentale est proposée dans cette thèse, nous mettons en discussion de manière empirique les considérations requises dans la classification super-pixellique à savoir l'information couleur de l'image et la caractérisation super-pixellique. La classification est effectuée par un apprentissage supervisé et semi-supervisé afin de mettre en évidence l'importance du semi supervisé dans la segmentation des images médicales.

Mots clés : Apprentissage semi-supervisé; classification; segmentation; auto-apprentissage; SETRED; SNNRCE; données médicales.

Abstract : The semi-supervised data is a context that reflects a real problem concerning the difficulty of manual data labeling. In supervised classification of medical data, the learning hypothesis requires the supervised information where the doctor has brought the necessary label. However, in view of the currently available data, the supervision of medical data has become a tedious task for the doctor. Therefore, the unlabeled data is more numerous and available than the labeled data. However, the classifier performance is related to the number of learning data, the main question that emerges is how to improve the learning by including unlabeled data at the learning set. The paradigm using both labeled and unlabeled data is called semi-supervised learning.

In this thesis, we focus on understanding the semi supervised learning, and we detail our main problem about automatic labeling by semi-supervised learning based on the "self-training" principle. Self-training is a reference method in semi supervised classification, its use is fundamental in several applications. The self-training method is an iterative mechanism. First, it trains a supervised classifier on labeled data to predict the labels of unlabeled data. Second, it iteratively enriches the labeled set by adding newly labeled examples with high confident prediction from the unlabeled data (confidence data). We introduce progressively the concept of self-training pcosses in medical applications. A first part of this thesis has been reserved for the understanding of the self-learning principle starting with the study of the ***SNNRCE*** algorithm. Next, we detail our proposed contribution to the annotation problem of the medial data, which is called ***R-COSET***.

In the last part of this thesis, we are particularly interested in the segmentation of medical images using the classification processes. The super-pixel classification has become a frequent and important method for the automatic segmentation. An experimental study is proposed in this thesis. We discuss empirically the considerations required in the super-pixel classification, namely the space color information and the super-pixel characterization. The classification is performed by supervised and semi-supervised learning in order to highlight the importance of semi-supervised in the medical image segmentation.

Keywords : Semi-supervised learning; classification; segmentation; self-training; SETRED; SNNRCE; medical dataset

ملخص : البيانات شبه الخاضعة للإشراف هي سياق يعكس مشكلة حقيقية تتعلق بصعوبة وضع البيانات يدويا. فيما يتعلق عن الإشراف البيانات الطبية، تتطلب فرضية التعلم المعلومات الخاضعة تحت إشراف الطبيب. ومع ذلك، وبالنظر إلى البيانات المتاحة حاليا، أصبح الإشراف على البيانات الطبية مهمة شاقة للطبيب. ولذلك، فإن البيانات غير المعينة هي أكثر عددا ومتوفرة من البيانات المعينة. ومع ذلك، يرتبط أداء المصنف بعدد بيانات التعلم، والسؤال الرئيسي الذي يبرز هو كيفية تحسين التعلم من خلال تضمين البيانات غير المعينة في مجموعة التعلم. ويطلق على النموذج باستخدام كل من البيانات المصنفة وغير المسماة التعلم شبه الخاضع للإشراف.

في هذه الأطروحة، نركز على فهم التعلم شبه الخاضع للإشراف، ونفصل بالتفصيل عن مشكلتنا الرئيسية حول وضع العلامات التلقائية عن طريق التعلم شبه الخاضع للإشراف على أساس مبدأ "التدريب الذاتي". التدريب الذاتي هو طريقة مرجعية في تصنيف تحت شبه الإشراف، واستخدامه أمر أساسي في العديد من التطبيقات. طريقة التدريب الذاتي هي آلية تكرارية. أولا، تقوم بتدريب مصنف تحت الإشراف على البيانات المصنفة للتنبؤ بتسميات البيانات غير المصنفة. ثانيا، فإنه يثري مرارا وتكرارا وضعت المسمى بإضافة أمثلة وصفت حديثا مع التنبؤ عالية ثقة من البيانات غير المسماة (بيانات الثقة). نحن نقدم تدريجيا مفهوم التدريب الذاتي بوسوس في التطبيقات الطبية. وقد تم تخصيص الجزء الأول من هذه الرسالة لفهم مبدأ التعلم الذاتي بدءا بدراسة خوارزمية SNNRCE. بعد ذلك، نحن بالتفصيل المساهمة المقترحة لمشكلة الشرح من البيانات وسطي، وهو ما يسمى R-COSET.

في الجزء الأخير من هذه الرسالة، نحن مهتمون بشكل خاص في تجزئة الصور الطبية باستخدام عمليات التصنيف. أصبح تصنيف فائقة الصعوبة. تم اقتراح دراسة تجريبية في هذه الرسالة. نناقش تجريبيا الاعتبارات المطلوبة في تصنيف فائقة الصعوبة. يتم إجراء التصنيف من خلال الإشراف تحت إشراف وشبه الإشراف من أجل تسليط الضوء على أهمية شبه إشراف في تجزئة الصورة الطبية.

كلمات البحث : التعلم شبه إشراف. تصنيف؛ تجزئة. تدريب ذاتي؛ SETRED. SNNRCE.