

# **RAPPORT PROJET DE FIN D'ÉTUDE**

Réalisation d'un programme d'aide au diagnostic, application sur les pathologies cérébrales.

Écrit par:

Philippe Chuzel

Promotion 2016, option SPID Robotique

Sous la supervision de :

Ali Mansour

ENSTA Bretagne



Option Systèmes Perception Information Décision



# Remerciement

Je tiens à remercier ici M. Ali Mansour et M. Cexus pour m'avoir aidé à trouver ce PFE en rapport avec les problématiques de la santé. Ce stage fut riche en rencontre très intéressante et ce fut un plaisir de travailler avec les contacts que j'ai pu avoir dont notamment les chercheurs de l'INSERM et les médecins du CHRU de Brest.

Je remercie donc M. Nacim Betrouni et M. Denis Hamad pour leurs conseils et leur accueil au sein de l'INSERM de Lille. Je remercie également les Médecins M. Jean Christophe Gentrrix et M. Julien Ognard pour leurs aide et conseils sur l'analyse des images médicales. Je leur dois aussi toute la base de données d'image à laquelle j'ai pu avoir accès.

Ce fut un projet avec peu de rapport avec les thématiques que j'ai pu étudier à l'ENSTA mais je me rends compte à quel point les compétences et expériences que j'ai eu tant pendant mon année militaire que pendant ma formation de trois ans à l'ENSTA m'ont permis d'aboutir à ce résultat. Je tiens donc à remercier tout les professeurs de l'ENSTA pour les cours qu'ils nous ont fait.

J'ai remercie encore une fois M. Ali Mansour pour m'avoir permis de publier, sa patience pour les relectures, le stylo rouge qu'il a sacrifié pour cette tache et la liberté de mouvement qu'il m'a laissé pendant ce PFE afin que je puisse également me former sur d'autres thématiques.

# Sommaire

<b>I</b>	<b>Présentation de la problématique du projet de fin d'étude.</b>	<b>7</b>
<b>1</b>	<b>Quelles sont les sources d'informations exploitables?</b>	<b>8</b>
1.1	Elastométrie . . . . .	8
1.2	Échographie . . . . .	9
1.3	Computed tomography angiography / Angiographie . . . . .	9
1.4	Scanner . . . . .	10
1.5	IRM . . . . .	10
<b>2</b>	<b>Premier sujet: Caractérisation de la thrombose par les IRM</b>	<b>12</b>
2.1	Première piste, utilisation des IRM de perfusion. . . . .	12
2.2	Qu'est ce que la thrombose veineuse profonde . . . . .	13
2.3	Problème rencontré vis à vis des IRM de thrombus. . . . .	15
<b>3</b>	<b>Changement de sujet: Réalisation d'un programme d'aide au diagnostic pour les pathologies cérébrales.</b>	<b>16</b>
<b>II</b>	<b>Choix de la chaine de traitement.</b>	<b>17</b>
<b>4</b>	<b>Réalisation de la chaine de traitement.</b>	<b>18</b>
<b>5</b>	<b>Intérêt de la classification spectrale.</b>	<b>20</b>
<b>6</b>	<b>Choix réalisé pour l'implémentation de la classification spectrale.</b>	<b>22</b>
6.1	Choix dans la construction du graphe et de la matrice de similitude. . . . .	22
6.2	Choix de l'algorithme de classification spectrale. . . . .	24
6.3	Choix du nombre de classe. . . . .	24
<b>7</b>	<b>Résultats de l'algorithme et interprétations.</b>	<b>25</b>
<b>III</b>	<b>Résultats.</b>	<b>29</b>
<b>8</b>	<b>IMRs T1 de la prostate de l'INSERM de Lille.</b>	<b>30</b>
8.1	Types de courbe rencontrées dans le cadre du cancer de la prostate . . . . .	30
8.2	Résultats de l'algorithme . . . . .	31
<b>9</b>	<b>IMRs T2 du crane CHRU Brest Cavale Blanche.</b>	<b>33</b>
9.1	Patient 1: Cancer identifiable par l'IRM de perfusion. . . . .	34
9.2	Patient 2: Cancer non-identifiable par l'IRM de perfusion. . . . .	35

<b>IV</b>	<b>Annexes</b>	<b>39</b>
<b>A</b>	<b>Anonymisation des images DICOM</b>	<b>40</b>
<b>B</b>	<b>Visualisation des données en 3D et 4D</b>	<b>41</b>
<b>C</b>	<b>Choix automatique du nombre de classe.</b>	<b>42</b>
<b>D</b>	<b>Réalisation de programme pour le milieu médical en stand-alone.</b>	<b>45</b>
<b>E</b>	<b>Rappel sur la théorie des graphes.</b>	<b>47</b>
	E.1 Définitions de base. . . . .	47
	E.2 Choix du laplacien de la matrice. . . . .	47
<b>F</b>	<b>Construction de la matrice de voisinage.</b>	<b>48</b>
	F.1 Choix du type de voisinage. . . . .	48
	F.2 Choix de la distance et de la similitude. . . . .	48
	F.3 Choix du paramètre de dispersion. . . . .	49
<b>G</b>	<b>Quels algorithmes de classification spectrale?</b>	<b>50</b>

# Introduction

Le sujet d'étude de ce PFE a changé plusieurs fois pendant les 6 mois de travail. Au départ, M. Ali Mansour m'a proposé de m'incorporer à son projet d'étude sur la détection et la caractérisation des caillots sanguins, sujet qui était également étudié par Thibaud Berthomier. M. Berthomier cherchait à réaliser cette étude en utilisant des images issues de l'écho-graphiques et de l'élastométrie. Ce dernier est en thèse à l'ENSTA Bretagne et plusieurs bases de données ont déjà été mises au point pour son étude.

Pour ma part, je devais réaliser cette tâche en étudiant des images IRM de thrombose et en utilisant une méthode mise au point par l'INSERM de Lille [1]. J'ai rapidement développé les algorithmes qui étaient implémentés dans cette thèse et j'ai pu les tester sur des données simples. Néanmoins, la récupération d'image fut beaucoup plus ardue. Dans un premier temps, j'ai pu rencontrer plusieurs médecins du CHRU de Brest qui m'ont beaucoup orienté et qui m'ont permis de délimiter un sujet clair d'étude pour mon projet de fin d'étude.

Après plusieurs entretiens avec deux médecins du CHRU, M. Jean Christophe Gentrix et Pierre-Yves Salaun, et un interne actuellement au CHRU de la Cavale Blanche, M. Julien Ognard, nous sommes arrivés à la conclusion que tenter de caractériser les caillots sanguins par les IRM n'était pas réalisable pendant la période du PFE. M. Gentrix et M. Ognard étant principalement spécialisés sur les pathologies du cerveau, ces derniers m'ont donc proposé de réaliser un programme d'aide au diagnostic pour les pathologies cérébrales, chose qui était possible avec les algorithmes que j'ai mis au point au début du stage.

Pendant ces 6 mois, j'ai pu rencontrer une partie du personnel de l'INSERM de Lille qui avait mis au point un premier programme d'aide au diagnostic pour le cancer de la prostate et j'ai pu avoir 3 bases de données:

- Une base de données simulées pour le tissu cancéreux de la prostate,
- Une base de données réelle d'image IRM de perfusion de prostate présentant des cancers qui a été utilisée pour valider dans un premier temps les résultats de l'algorithme,
- Une base de données d'image IRM de perfusion de cerveau présentant des pathologies cérébrales qui ont permis de voir si le programme d'aide au diagnostic mis au point proposait des résultats pertinents.

Au final, les résultats obtenus ont été satisfaisants et j'ai eu la chance de pouvoir publier les résultats dans la revue International Conference on Frontiers of Signal Processing (ICFSP) grâce à l'aide et aux recommandations de M. Ali Mansour.

Ce rapport va dans un premier temps montrer l'évolution du projet et les différents travaux bibliographiques qui ont été réalisés pendant cette étude. La seconde partie va présenter comment est architecturée notre chaîne de traitement pour notre programme d'aide au diagnostic, l'intérêt de la classification spectrale pour notre chaîne de traitement et les données sur lesquelles nous avons travaillé. La dernière partie va récapituler les grands résultats qui ont été obtenus.

## **Part I**

### **Présentation de la problématique du projet de fin d'étude.**

# Chapter 1

## Quelles sont les sources d'informations exploitables?

Le sujet de départ de l'étude, qui est mené actuellement à l'ENSTA Bretagne, est de proposer des méthodes d'identification et de caractérisation de la thrombose et du thrombose veineuse profonde ou DVT. Cette étude est menée en partenariat avec plusieurs medecins du CHRU de Brest, M. Bressollette et M. Moittier qui fournissent actuellement la majorité des images qui sont utilisées pour cette étude. Dans un premier temps, nous nous sommes intéressés aux différents types d'image que nous pouvions avoir accès.

### 1.1 Elastométrie

L'élastométrie est une méthode d'acquisition d'image médicale non invasive et très pratique. En effet, très facile à mettre en œuvre et peu couteuse en temps et en argent. Cette méthode consiste à créer une représentation d'une zone du corps à partir de l'élasticité du tissu biologique [2].

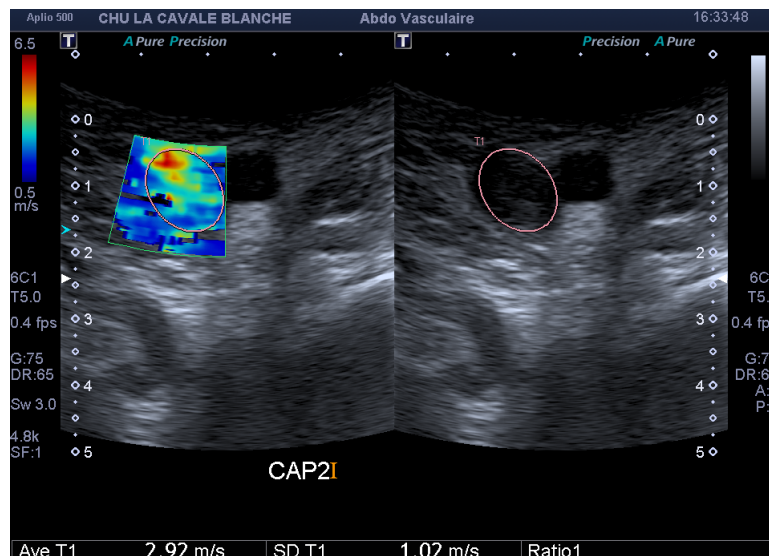


Figure 1.1: Exemple d'image d'élastometrie.

Cette méthode présente plusieurs particularités pour cette étude:

- Elle est très facile à utiliser et se fait en un temps très faible.
- C'est une méthode qui n'est pas invasive.



- Bien que la mise en œuvre soit facile, les images obtenues sont de médiocre qualité en terme de résolution.

Néanmoins, il y a un problème à cette méthode. Il est nécessaire de définir un protocole unique à tous les patients pour faire en sorte que les images obtenues par ce procédé soient comparables.

## 1.2 Échographie

L'échographie est une technique d'imagerie qui consiste à utiliser des ondes ultrasons. En fonction des propriétés du milieu, certains tissus ou partie du corps apparaitront en noir, gris ou blanc.



Figure 1.2: Exemple d'image échographique pour l'obstétrique.

Cette méthode présente les mêmes avantages que l'élastométrie. Elle est peu chère, facile à mettre en œuvre et non invasive pour le patient. Néanmoins, elle présente également des images de très faible qualité en terme de résolution.

## 1.3 Computed tomography angiography / Angiographie

L'angiographie est une technique d'imagerie qui est invasive. Bien qu'elle présente plusieurs intérêts, notamment car elle consiste à faire une imagerie des vaisseaux sanguins par rayon X (veines ou artères), il y a plusieurs effets secondaires (vomissement, vertiges, nausées, réactions allergiques, saignement important...) ou contre-indications (grossesse, tension artérielle trop basse...). Cette piste ne sera pas étudiée ici.

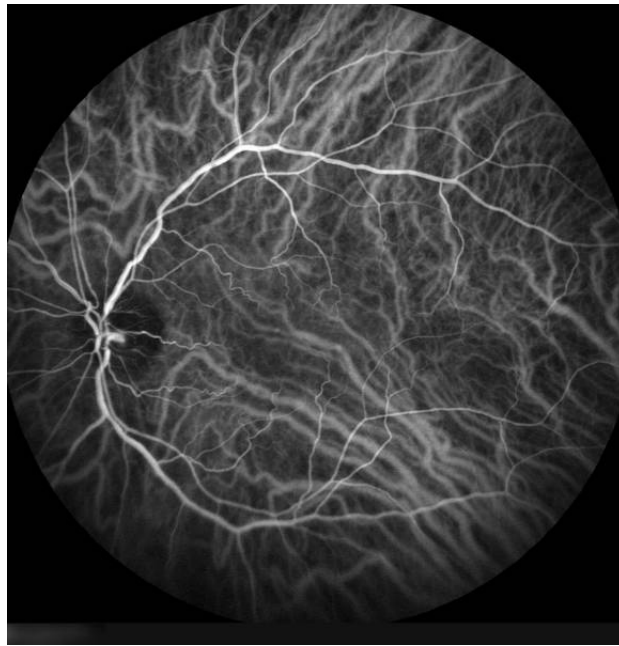


Figure 1.3: Exemple d'angiographie pour l'œil.

## 1.4 Scanner

Le scanner est un outil qui propose de reconstituer des images des tissus par ordinateur à partir des différences d'absorption des rayons X. Ce procédé permet d'obtenir des coupes du corps du patient très rapidement et présente certains avantages et inconvénients. C'est une méthode qui permet d'obtenir très rapidement des images médicales exploitable pour faire un diagnostic. Néanmoins, c'est un procédé qui impose l'utilisation de rayon X donc irradiant. C'est donc un procédé invasif.

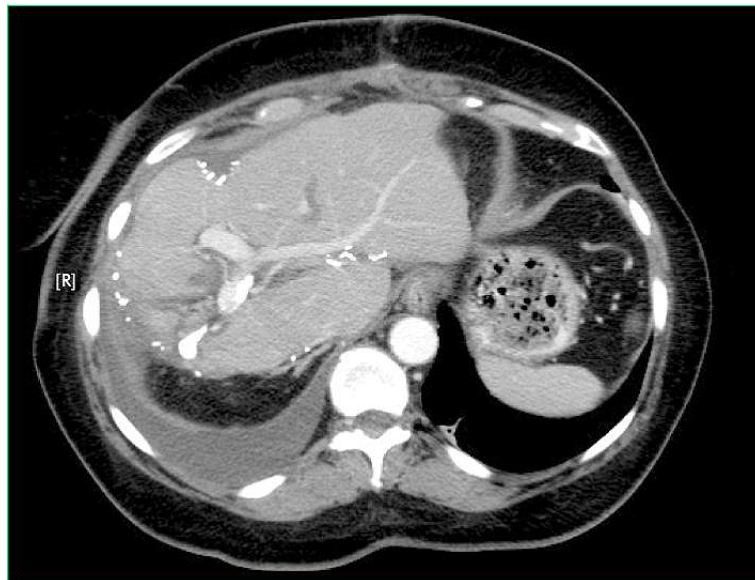


Figure 1.4: Exemple d'image issue d'un scanner.

## 1.5 IRM

L'IRM ou l'imagerie par résonance magnétique sera ici la source d'image médicale qui sera étudiée. Elle produit des images avec une haute résolution et présent un énorme avantage pour nous. C'est

un procédé entièrement non invasif et, à la différence du scanner, elle ne nécessite aucun produit irradiant.

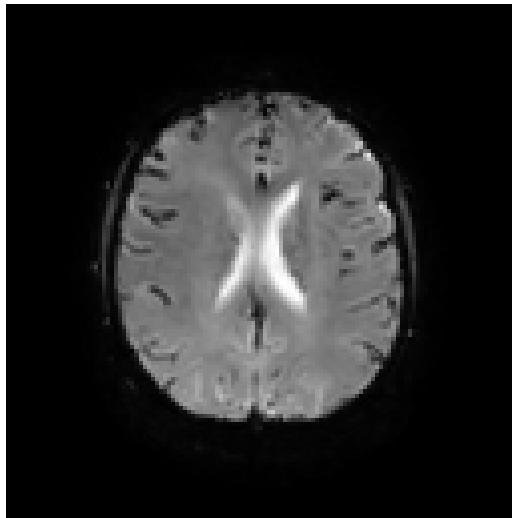


Figure 1.5: Exemple d'image IRM.

Le principe de base de l'IRM est de faire une imagerie des atomes d'hydrogène présent dans le tissu que nous souhaitons étudier. Nous considérons ici que le noyau de l'atome d'hydrogène qui est juste un proton qui tourne sur lui-même. L'idée est d'utiliser un puissant aimant qui va générer un champs et une onde radio excitatrice de fréquence identique à la fréquence de rotation des protons. Ceci a pour conséquence que tous les protons tournent dans un même axe qui est légèrement différent de l'axe du champs de l'aimant et qu'ils soient en phase. Ce léger écart dans l'orientation de l'axe est dû à l'excitation par l'onde radio.

A partir de ce point, on coupe l'onde excitatrice et deux phénomènes se présentent:

- Les protons vont revenir à leur état chaotique de départ et vont tous se déphaser,
- Les protons vont tous s'orienter par rapport à l'axe du champs de l'aimant.

Ce changement d'état se traduit par la diffusion d'un signal qui est capté et permet la réalisation de la cartographie des protons et donc de notre image IRM. On peut paramétrer la machine pour mesurer deux paramètres:

- Soit on mesure le temps que mettent les protons pour se remettre dans l'axe de l'aimant ce qui est appelé temps de relaxation T1 ou pondération T1.
- Soit on mesure le temps que mettent les protons pour se désynchroniser ce qui est appelé temps de relaxation T2 ou pondération T2.

## Chapter 2

### Premier sujet: Caractérisation de la thrombose par les IRM

Pendant la durée du stage, nous étions deux à travailler sur la problématique de la détection et de la caractérisation de la thrombose. Thibaud Berthomier qui est actuellement en thèse à l'ENSTA Bretagne a pu récupérer de la part du CHRU plusieurs images de caillot sanguin. L'étude qu'il mène se concentre sur les caillots situés au niveau de la cuisse. Je devais, pour ma part, explorer une autre piste qui consistait à utiliser des images IRM. Je devais également utiliser le résultat d'une thèse pour voir si la méthode qui y est développée pouvait être utilisée [1], [3]. Il nous a fallu donc dans un premier temps faire un état de l'art sur les éléments développés dans cette thèse et sur le thrombose veineuse profonde (DVT).

#### 2.1 Première piste, utilisation des IRM de perfusion.

Le principe de l'IRM de perfusion est d'injecter un produit de contraste, ici le galonium, qui est visible par IRM et de voir comment ce produit se diffuse au cours du temps dans les différentes zones du corps. A terme, nous obtenons un ensemble d'image traduisant l'évolution de la concentration du produit dans la zone étudiée au cours du temps. A partir de ces images, nous pouvons déterminer la courbe d'intensité au cours du temps de chaque pixel de l'IRM. L'idée est donc de voir le comportement d'une zone du corps avant, pendant et après injection du produit de contraste et de voir si on peut caractériser ainsi une zone en particulier.

Dans un premier temps, nous souhaitions utiliser la méthodologie mise au point dans la thèse [1]. Cette dernière propose deux outils pour l'aide au diagnostic. Le premier est l'établissement de carte de paramètre pharmacocinétique. Le second est l'utilisation d'algorithme de classification non supervisé pour réaliser des clusters à partir différentes courbes d'intensité obtenues afin d'isoler les courbes présentant le même comportement entre elles. Ainsi, on obtient une nouvelle carte segmentée qui indique les régions qui présentent les mêmes caractéristiques.

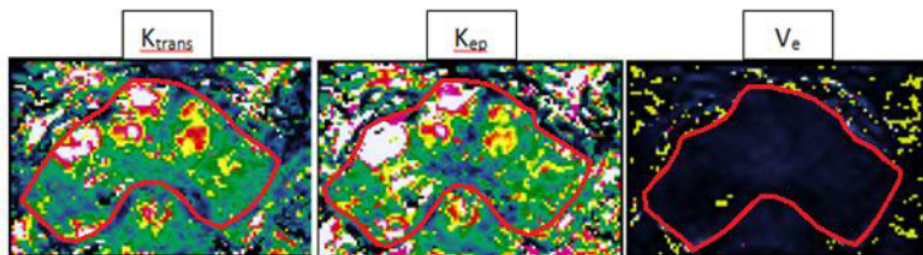


Figure 2.1: Cartes de parametre pharmacocinetique  $K_p$ ,  $K_{trans}$  et  $V_e$  de la prostate.

La figure qui suit montre la méthodologie appliquée ici 2.2.

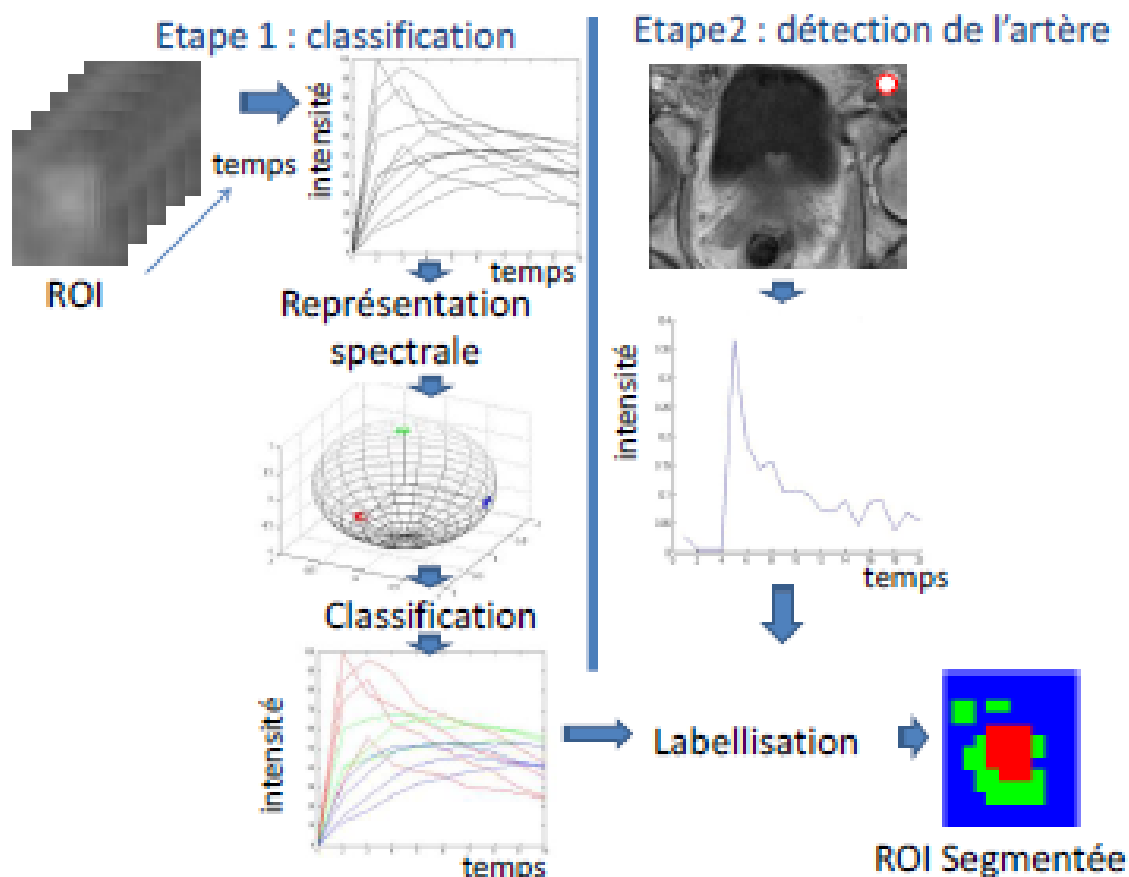


Figure 2.2: Méthodologie suivie pour la classification.

L'idée est d'avoir au final des images qui segmentent directement les tissus sains et les tissus pathologiques.

## 2.2 Qu'est ce que la thrombose veineuse profonde

On appelle thrombose le fait qu'un caillot sanguin ou thrombus se développe à l'intérieur d'un vaisseau sanguin. Il peut être soit artériel, soit veineux. Dans les deux cas, ce caillot peut obstruer le vaisseau et donc empêcher la circulation du sang.

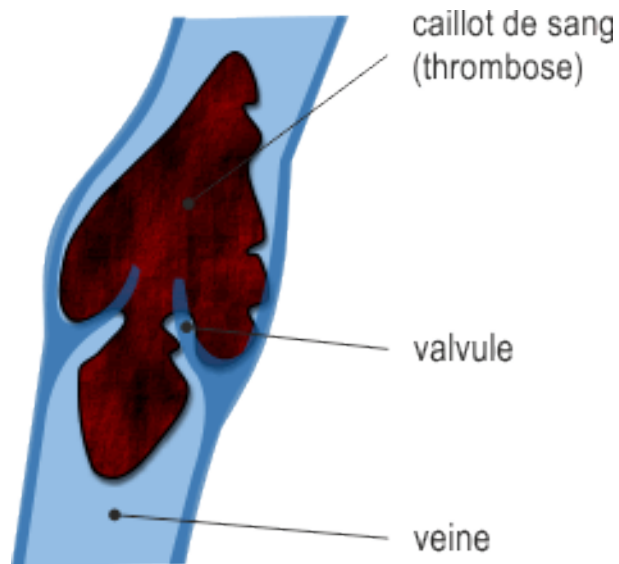


Figure 2.3: Exemple illustrant la thrombose.

La phlébite est un cas de thrombose veineuse pour une veine profonde et est consécutif à l'inflammation de la veine et au risque de déplacement du caillot. D'où le terme de thrombose veineuse profonde DVT (deep venous thrombosis).

Parmi les principales causes, nous pouvons énumérer:

- Une lésion dans une veine qui entraîne une coagulation au niveau du tissu sanguin (blessure...),
- Certains paramètres génétiques qui peuvent favoriser le processus de thrombose,
- Les périodes prolongées dans une même position (alitement, accouchement, voyage long en avion...),
- Des précédentes expériences traumatisantes pour le corps (cancer, opération...),
- Influence de certaines substances sur la paroi des artères (tabac, médicaments...),
- Facteurs favorisant la formation de thrombus (obésité, diabète, hypercholestérolémie...).

Une fois le caillot formé, il peut se produire différents phénomènes:

- Il peut de lui même se désagréger et disparaître naturellement,
- Il peut se calcifier ou changer de structure ce qui peut amener à la formation de pelote fibreuse,
- Il peut se détacher et atteindre un des vaisseaux pulmonaires. Ceci se traduit par une embolie pulmonaire qui s'avère très dangereux.

Dans tous les cas, la présence de ce caillot dans le système circulatoire sanguin d'un patient peut mener à des perturbations graves du flux sanguin avec l'apparition d'œdème ou d'ulcère pour la zone inférieure du corps que nous allons étudier.

## **2.3 Problème rencontré vis à vis des IRM de thrombus.**

Les premières semaines du stage ont donc consisté à tenter de récupérer des images IRM de thrombus afin de voir ce qui était possible de faire. Très rapidement, plusieurs problèmes se sont présentés:

- Le fait d'utiliser des IRM pour des caillots sanguins est encore très récent et nécessite la création de nouveaux protocoles qui doivent être validés par des comités d'éthique. Cette démarche prend beaucoup de temps et cette utilisation des IRM reste encore au niveau de la recherche expérimentale.
- Il n'y a pour l'instant aucune image IRM de caillot situé au niveau de la cuisse, caillot qui est pour l'instant principalement étudié.
- Pour les thrombus cérébraux, il existe des IRM mais plusieurs problèmes se sont rapidement manifestés. La signature IRM du caillot change assez rapidement en fonction de son âge. De plus, sa composition et sa position ont un lourd impact sur comment il sera vu sur l'IRM.

Au final, si nous devons poursuivre sur cette voie, nous aurions dû nous concentrer sur les thromboses cérébraux. De plus, après plusieurs discussions avec les médecins du CHRU, cela aurait nécessité un travail qui n'était pas réalisable en 6 mois.

## Chapter 3

### Changement de sujet: Réalisation d'un programme d'aide au diagnostic pour les pathologies cérébrales.

Après une période d'investigation qui a démontré que le premier objectif était irréalisable dans le temps imparti, les médecins du CHRU m'ont proposé de réaliser un programme d'aide au diagnostic pour les pathologies cérébrales. Ces derniers étaient justement spécialisés sur cette partie du corps humain et pouvaient me fournir les images nécessaires pour que je puisse travailler. De plus, j'ai pu prendre contact avec les personnes qui ont encadré la thèse sur laquelle je m'appuyais pour récupérer leurs données expérimentales et simulées. J'ai donc pu déjà profiter de leur expertise et d'autre part récupérer plusieurs données simulées de prostate présentant des tumeurs et plusieurs IRM de perfusion de prostate présentant cette pathologie.

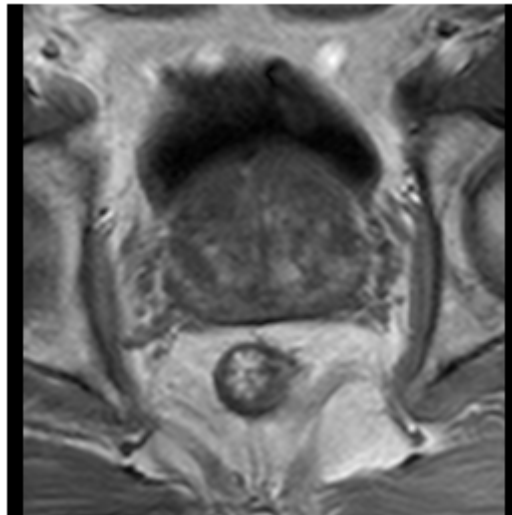


Figure 3.1: Image IRM de la prostate.

Au final, j'ai pu récupérer avec le CHRU les IRM de perfusion de 13 patients présentant des pathologies assez visibles pour que je puisse tester mes algorithmes. Parmi les pathologies qui ont été recensées, il y a plusieurs formes de cancer et des œdèmes assez significatifs. Le travail qui va suivre a donc pour but de voir quelles sont les pathologies qui peuvent être identifiées par cette méthode, généraliser et proposer des améliorations au niveau de la chaîne de traitement et donc réaliser un programme d'aide au diagnostic pour ces pathologies.



## **Part II**

### **Choix de la chaine de traitement.**

# Chapter 4

## Réalisation de la chaine de traitement.

Tout le travail qui va être présenté a été fait sur Matlab. Il existe plusieurs bibliothèques qui sont spécialement dédiées à l'utilisation d'image IRM. Toutes les images qui ont été utilisées sont au format Dicom. Ce format contient d'une part l'image du patient et d'autre part un vaste ensemble de méta-données qui donne des informations sur la machine qui a pris l'image, sur le patient et son état de santé, le lieu de l'établissement qui a fait l'image et les nom des praticiens qui ont fait la manipulation.

Il y a un certain travail de pré-traitement des données à faire avant de pouvoir utiliser les IRM de perfusion. Grâce aux méta-données, on peut récupérer très facilement le nombre de coupe et le nombre de séquence pour une coupe donnée. La Figure 4.1 montre comment évolue l'intensité de l'IRM d'une coupe au cours du temps pendant la diffusion du produit de contraste.

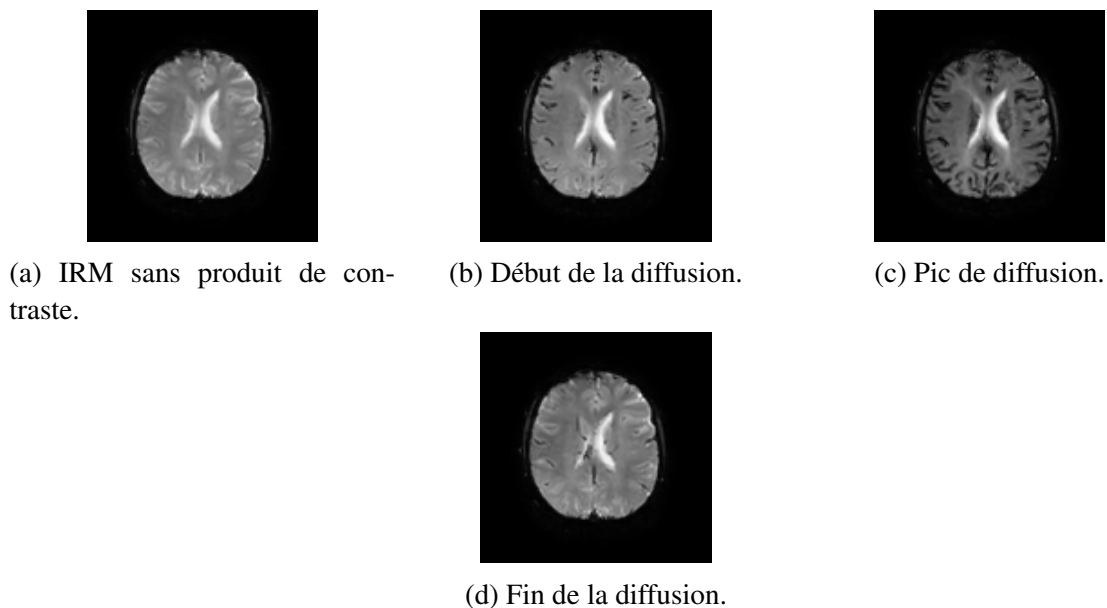


Figure 4.1: Diffusion du produit de contraste dans une IRM de perfusion.

Pour chacune des IRM de cerveau que nous avons récupéré, nous avons donc obtenu un ensemble d'images IRM qui sont réparties selon 24 ou 26 coupes avec pour chaque coupe 40 images traduisant la diffusion du produit de contraste et chaque image est de dimension 256 sur 256 pixels. On a donc pour chaque coupe un tenseur de dimension  $256 \times 256 \times 40$  qui contient toutes les informations sur la diffusion du produit de contraste.

Quelque soit l'IRM considérée, la chaine de traitement restera la même. Il faudra choisir l'IRM d'un patient, stocker les images IRMs par coupe, choisir une coupe puis une région d'intérêt et enfin appliquer l'algorithme sur les éléments sélectionnés comme le montre la figure 4.2.

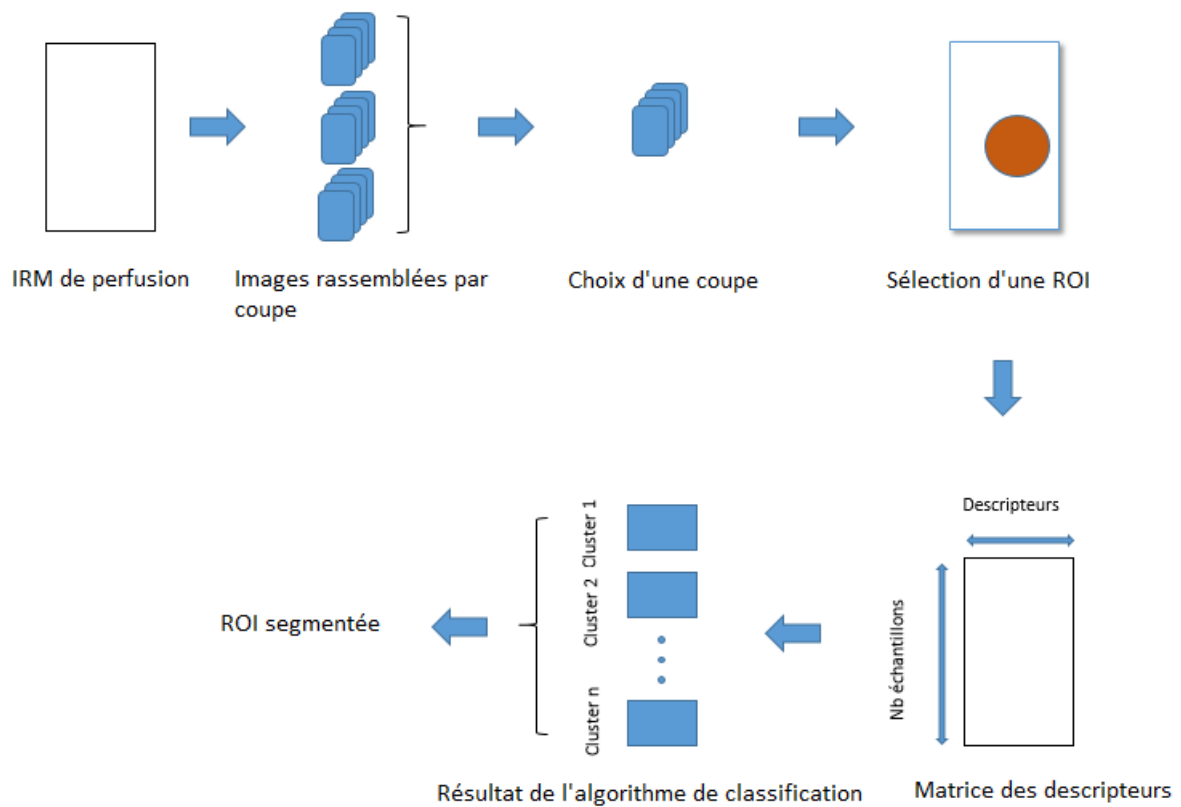


Figure 4.2: Chaîne de traitement développée dans le cadre du projet.

La sélection de la région d'intérêt se fait assez facilement avec Matlab comme le montre la figure suivante. L'idée est de prendre une zone et une coupe qui contiennent la zone présentant une pathologie. La partie qui suit va donc consister à expliquer comment on peut, à partir des signaux d'intensité de chacun des pixels, réaliser des clusters qui représentent des tissus homogènes et qui ont donc le même comportement pendant la diffusion du produit de contraste.

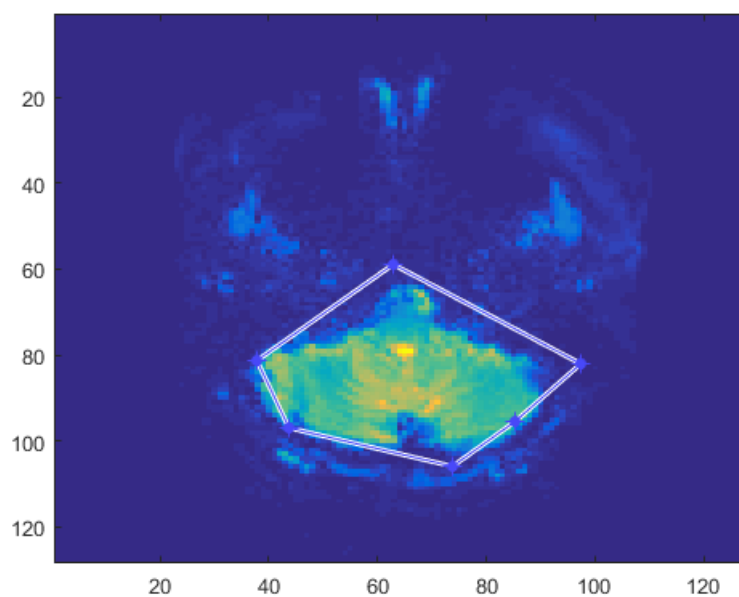
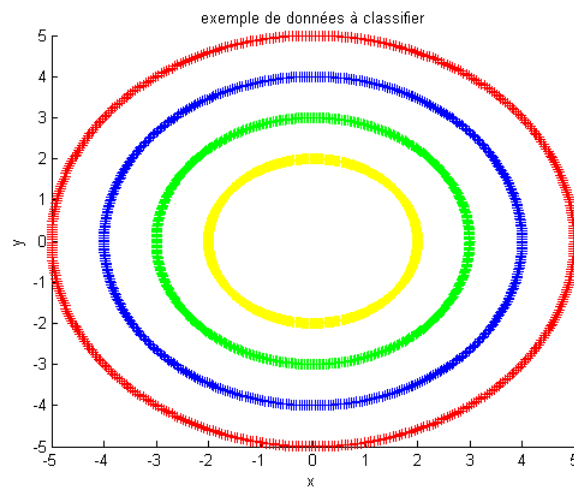


Figure 4.3: Choix d'une région d'intérêt avec Matlab.

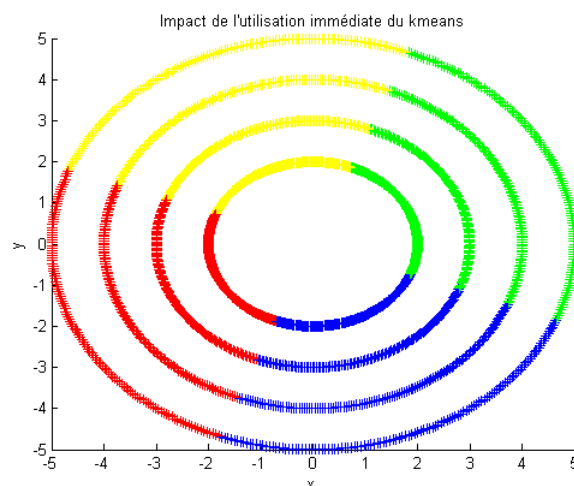
# Chapter 5

## Intérêt de la classification spectrale.

Il existe une très large bibliographie sur la classification non supervisée. Néanmoins, comme le montre la figure suivante, l'algorithme du K-means ne parvient pas à aboutir au résultat escompté.



(a) Ensemble de données de départ. Les données sont générées sur un cercle dont le rayon varie selon la classe



(b) Application du Kmeans sur cet ensemble de données. Les données ne sont pas correctement classifiées.

Ici, tous les nuages de points présentent un même centre et cela fausse le résultat bien que des zones de frontières sont facilement identifiables. L'intérêt de la classification spectrale est de construire un graphe par l'intermédiaire d'une matrice de voisinage qui va traduire les similitudes entre les

entrées. Ensuite, il faut travailler sur le spectre de cette matrice afin de pouvoir faire une classification.

Dans notre cas d'étude, nous devons étudier des signaux temporels dont on ne connaît a priori aucuns paramètres discriminants. Il est donc nécessaire de parvenir à trouver un espace dans lequel la classification puisse aboutir à des résultats pertinents. La classification spectrale est donc ici particulièrement adaptée par rapport aux autres algorithmes de classification non supervisé.

# Chapter 6

## Choix réalisé pour l'implémentation de la classification spectrale.

Nous devons implémenter un algorithme qui puisse être assez robuste pour traiter différents types d'informations, que ce soit les intensités des pixels des images IRM que nous étudions que des données simulées quelconques. C'est pour cela que certains choix dans la mise en place de l'algorithme ont été fait. Le lecteur pourra trouver en annexe des informations sur la théorie des graphes et également une partie des algorithmes qui ont été développés mais qui n'ont pas servis dans la chaîne de traitement finale.

Une fois que l'opérateur a sélectionné une région d'intérêt, on génère une matrice qui contient les courbes d'intensité des pixels sélectionnés. Cette matrice est donnée en entrée de notre chaîne de traitement comme le montre la figure 4.2.

### 6.1 Choix dans la construction du graphe et de la matrice de similitude.

Nous voulions avoir une chaîne de traitement qui soit la plus robuste et la plus automatisée possible. Nous n'avons aucune connaissance a priori sur les données en entrée donc il faut calculer un coefficient de similitude entre toutes les données entre elles. Par conséquent, on a choisi de construire un graphe entièrement connecté ce qui se traduit par l'attribution d'un coefficient de similitude  $W_{ij} = W_{ji}$  pour les signaux du pixel  $i$  et du pixel  $j$ .

On va donc représenter notre graphe par une matrice de similitude  $W$  qui aura comme propriété d'être symétrique. Chaque nœud du graphe correspondra à une ligne ou une colonne de la matrice et le coefficient  $W_{ij}$  correspondra au lien reliant les nœuds  $i$  et  $j$ . Il faut après définir le coefficient de similitude  $W_{ij}$ . L'annexe E contient les différentes définitions qui peuvent être choisies. Celle qui donne le plus satisfaction est la similitude euclidienne qui est définie par:

$$W_{ij} = \exp\left(-\frac{d^2(x_i, x_j)}{2\sigma_i\sigma_j}\right) \quad (6.1)$$

Où  $x_i$  est le signal du pixel  $i$ ,  $\sigma_i$  est un paramètre qui traduit la dispersion des entrées autour de  $x_i$  et  $d(x_i, x_j)$  est la distance euclidienne entre les signaux  $x_i$  et  $x_j$ . Le document [4] donne plusieurs définitions pour ce paramètre et la définition qui donne le plus satisfaction est de prendre  $\sigma_i = d(x_i, x_r)$  où  $x_r$  correspond au septième voisin le plus proche de  $x_i$ .

Au final, nous arrivons à une matrice de similitude comme celle ci-dessous pour des données réelles. Cette dernière est affichée en fausse couleur. Les données qui ont été utilisées sont issues

d'une IRM de cerveau où on a demandé une classification en 4 classes. La première figure montre la matrice d'affinité une fois tous les calculs réalisés et la seconde montre la matrice d'affinité après l'utilisation de notre algorithme de classification spectrale. On voit nettement 4 blocs au niveau de la diagonal de la matrice. Ces blocs traduisent que les données constitue une classe homogène. Les blocs hors-diagonaux traduisent la ressemblance interclasse.

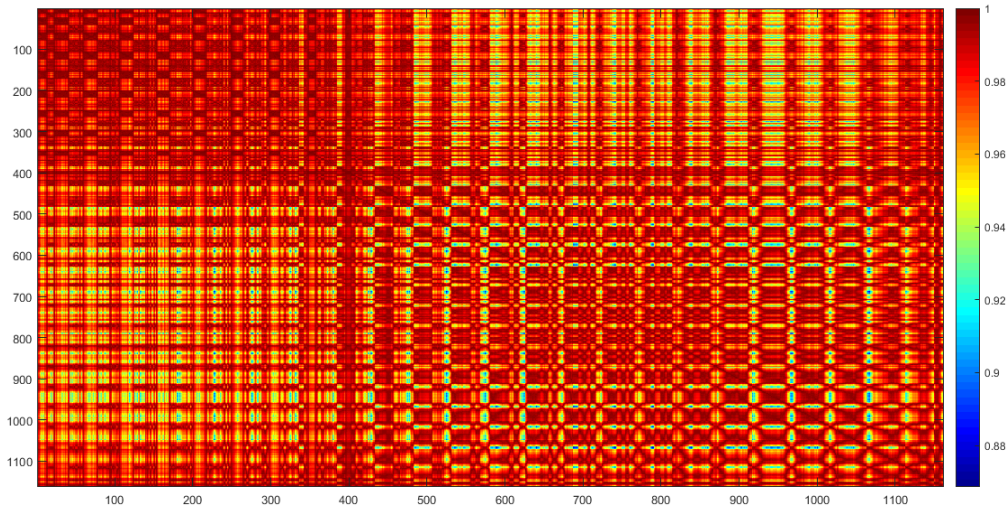


Figure 6.1: Exemple de matrice de similitude.

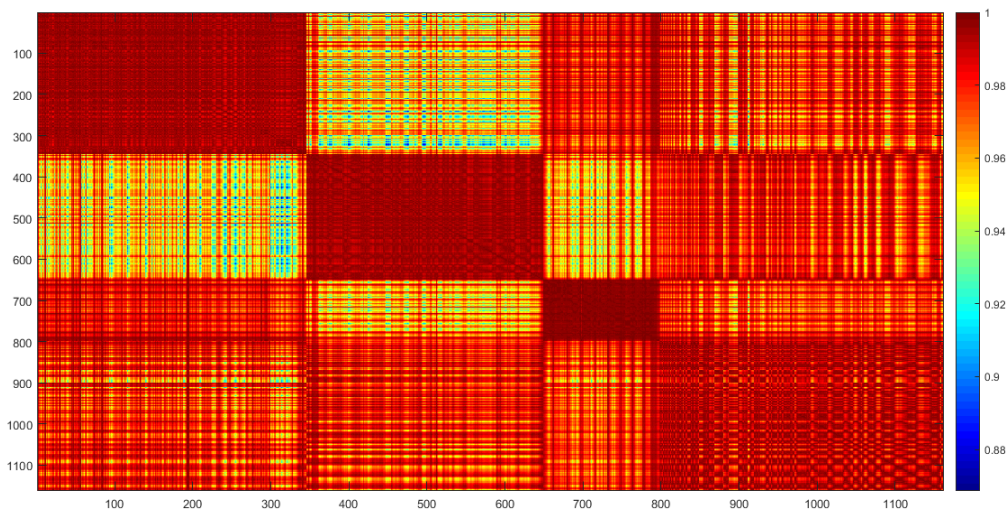


Figure 6.2: Matrice de similitude après utilisation de l'algorithme de classification spectrale.

Le dernière chose à faire est de définir le laplacien  $L$  de la matrice  $W$ . Pour notre algorithme, nous utilisons le laplacien normalisée symétrique  $L$  défini par:

$$L = I - D^{-1/2} W D^{-1/2} \quad (6.2)$$

Où la matrice  $D$ , matrice diagonale, contient les degrés des nœuds du graphe. Le degré d'un nœud est défini par  $d_i = \sum_{v_j \in V} W_{ij}$ . Et  $I$  est la matrice identité.

## 6.2 Choix de l'algorithme de classification spectrale.

L'annexe G contient les algorithmes qui ont été implémentés durant le PFE. Pour chacun d'entre eux le principe reste le même. Il faut extraire des informations du spectre de la matrice Laplacienne du graphe qui a été construit avec les remarques ci-dessus. Dans un premier temps, on extrait les plus petites valeurs propres de la matrice  $L$  et les vecteurs associés. Ensuite, on applique un algorithme des  $k$ -means sur les données générées par la matrice qui contient les précédents vecteurs propres. Il y a plusieurs algorithmes de classification spectrale dans la littérature [4] mais l'algorithme de Jordan et Weiss est celui qui nous a donné les meilleurs résultats. Les principales étapes sont données ci-dessous.

---

**Algorithm 1** Normalized spectral clustering, Jordan and Weiss

---

**Entrées**

- 1: Matrice  $W$  matrice de voisinage
- 2: Entier  $k$  nombre de classe

**Sorties**

- 1: Vecteur  $Ver$  table de vérité calculée.

**Algorithme**

- 1:  $L \leftarrow$  Laplacien symétrique de la matrice  $W$
  - 2:  $VectP \leftarrow$  matrice contenant les vecteurs propres des  $k$  plus petites valeurs propres non nulles de  $L$
  - 3: normaliser toutes les lignes de la matrice  $VectP$
  - 4:  $Ver \leftarrow$  résultat de l'algorithme du Kmeans sur la matrice  $VectP$  en  $k$  clusters.
  - 5: **return**  $Ver$
- 

## 6.3 Choix du nombre de classe.

Pendant cette étude, plusieurs méthodes ont été implémentées pour trouver automatiquement le nombre optimal de classe en fonction des données d'entrée. Le lecteur trouvera dans l'annexe C quelques algorithmes qui peuvent être utilisés. Sur les données réelles, aucuns d'entre eux ont donné de résultat. Par conséquent, l'opérateur qui utilisera notre algorithme et notre chaine de traitement devra tout de même donner le nombre de classe qu'il souhaite générer.

Au final, notre chaine de traitement a besoin de deux éléments:

- Une zone d'intérêt ou ROI qui doit être sélectionnée par l'opérateur. Ce qui suppose néanmoins une connaissance a priori sur la zone à étudier.
- Le nombre de cluster souhaité car il ne peut être trouver de manière automatique.



# Chapter 7

## Résultats de l'algorithme et interprétations.

Nous avons dans un premier temps utilisé notre algorithme sur des données simples comme celle qui est utilisée dans le chapitre 5. Voici le résultat que nous obtenons après l'utilisation de l'algorithme que nous avons expliqué dans la partie précédente.

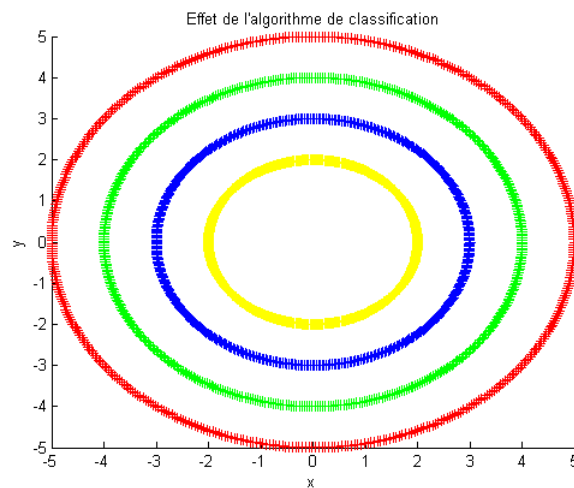


Figure 7.1: Application du premier algorithme sur les premières données. Les données sont correctement classifiées.

On peut voir que le résultat obtenu est bien celui espéré et que c'est une solution parfaitement viable pour agréger des données dans des clusters bien que les liens entre elles ne soient pas identifiable immédiatement. Les valeurs propres extraites du laplacien et qui sont utilisées pour la classification sont données ci-dessous.

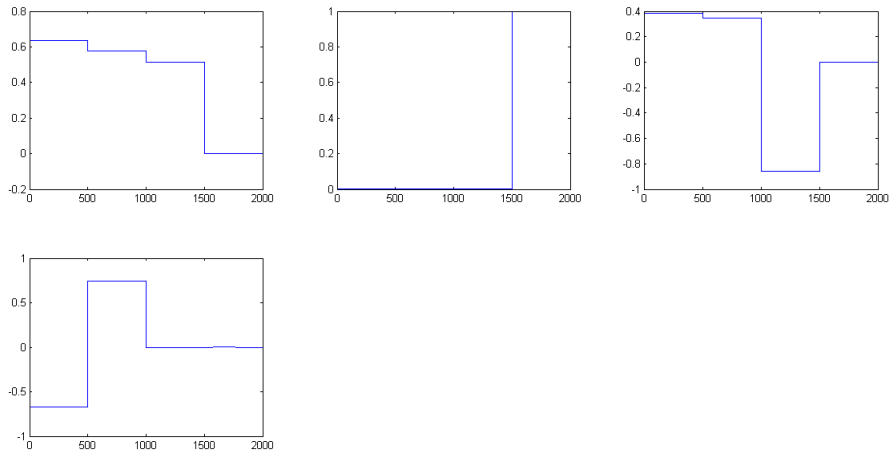


Figure 7.2: Vecteurs propres extraits du laplacien.

On peut clairement voir comment les K-means, sur ces données extraites, permettent de discriminer les quatre classes. Nous l'avons testé sur un ensemble de test et les figures suivantes montrent les résultats que nous avons obtenu.

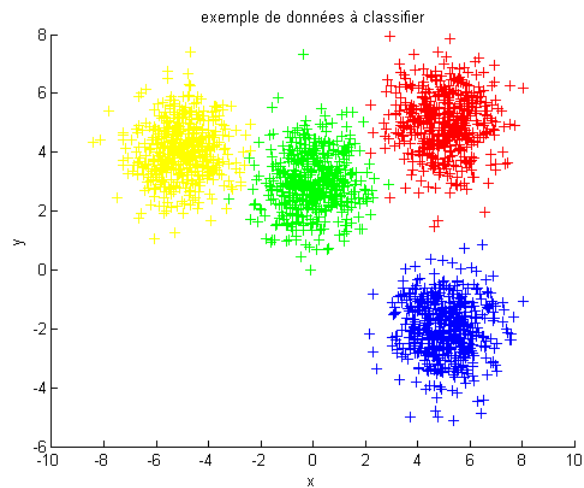


Figure 7.3: Second ensemble de données.

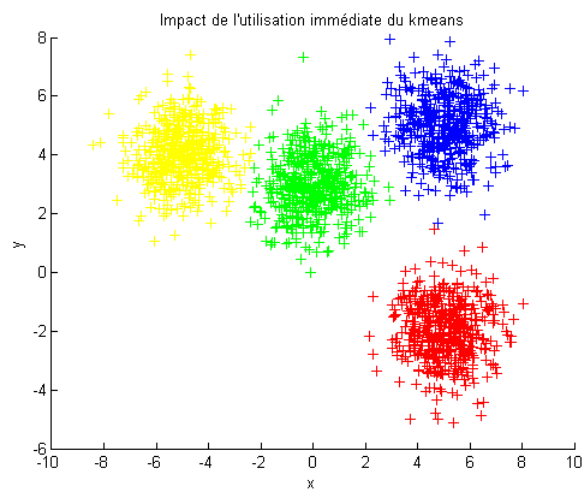


Figure 7.4: Application du Kmeans sur le second ensemble de données.

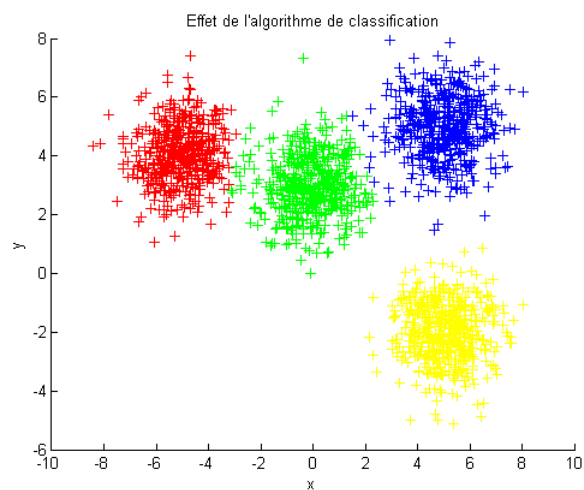


Figure 7.5: Application du premier algorithme sur le second ensemble de données.

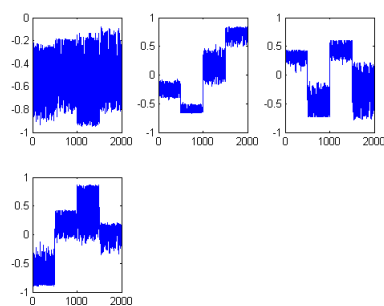


Figure 7.6: Vecteurs propres extraits du laplacien.

Là encore, le résultat est très satisfaisant.

# Conclusion

Au terme de cette étude, nous pouvons extraire ces différents résultats:

- Tout d'abord, la classification spectrale est particulièrement adaptée pour les problèmes que les algorithmes traditionnels comme le K-means ne peuvent résoudre. On arrive à se placer facilement dans un nouvel espace qui permet d'identifier clairement les éléments appartenant à une même classe. Il est donc tout à fait pertinent pour les signaux que nous devons traiter.
- La mise en place des différents algorithmes est au final relativement simple et peut couteuse en temps de calcul. Néanmoins, si le nombre de point devient trop grand, il peut y avoir un problème sur le temps d'exécution.
- La chaine de traitement qui a été développé est très robuste et peut fonctionner dans beaucoup de cas étant donné que nous ne faisons pas d'a priori sur les données de départ et que nous calculons les similitudes de toutes les entrées entre elles.

Néanmoins, il y a plusieurs contraintes qui faut prendre en compte:

- Le résultat final est fortement impacté par l'utilisation du K-means à la fin de l'algorithme de classification spectrale. Il arrive que l'algorithme ne converge pas vers les solutions souhaitées en fonction de l'initialisation.
- Il y a un besoin de connaissance a priori sur le nombre de classe que nous souhaitons. Cela veut dire que notre chaine de traitement ne pourra pas être entièrement autonome.

# **Part III**

## **Résultats.**

# Chapter 8

## IMRs T1 de la prostate de l'INSERM de Lille.

Tous les résultats, codes et bases de données d'images IRM peuvent être trouvés sur le git suivant:  
<https://github.com/chuzelph-ENSTA-Bretagne/ThromboseIRM.git>

### 8.1 Types de courbe rencontrées dans le cadre du cancer de la prostate

Le document [5] présente les différentes courbes caractéristiques de la prostate. Les courbes d'intensité présentent toutes un hyper signal en IRM T1. Néanmoins, les courbes issus de tissus cancéreux présentent un pic d'absorption beaucoup plus marqué avec un temps d'assimilation très caractéristique alors que les tissus sain présentent une courbe d'absorption beaucoup plus lissée. La figure suivante issue de [5] donne une illustration de ces courbes caractéristiques.

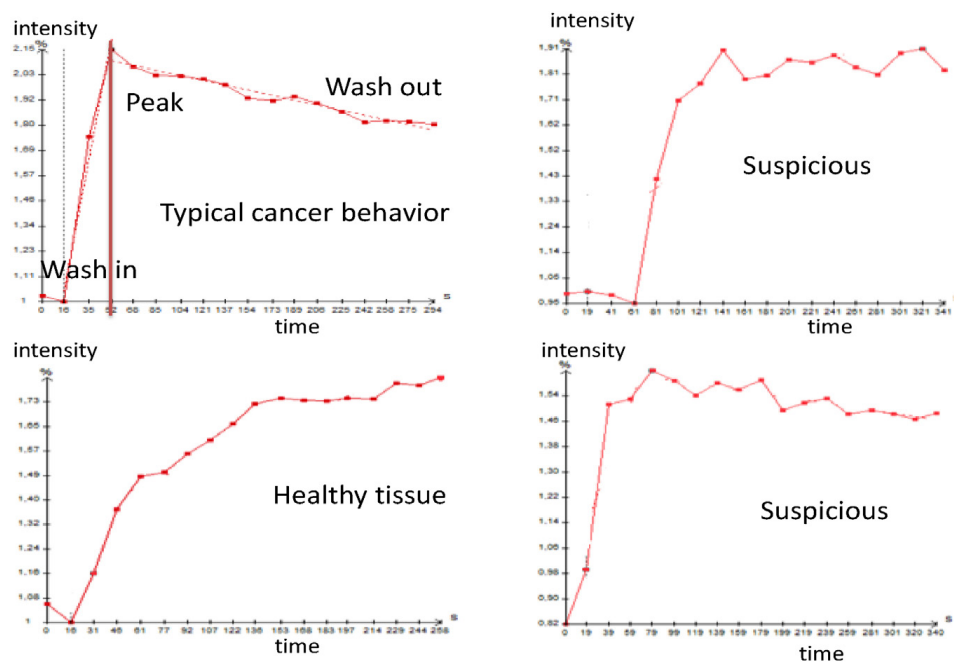


Figure 8.1: Courbes caractéristiques des tissus sain et cancéreux de la prostate.

Néanmoins, certaines courbes sont difficilement à classifier. L'intérêt de notre algorithme va donc

être de pouvoir distinguer aisément les tissus sains et les tissus cancéreux et d'estimer avec une autre classe supplémentaire les cas d'incertitude.

## 8.2 Résultats de l'algorithme

Pour appliquer nos algorithmes, il y a un petit pré-traitement à faire. Pour discriminer les différentes courbes, il faut surtout voir le phénomène de diffusion au sein du tissu et normaliser les courbes afin de faire ressortir le comportement du tissu.

Tous les courbes sont donc remises à l'échelle en appliquant la formule suivante:

$$x(t) \leftarrow \frac{x(t) - x_{min}}{x_{max} - x_{min}} \quad (8.1)$$

Tous les signaux temporels sont donc placés dans une matrice, comme expliqué dans la figure 4.2. Au final, en appliquant nos algorithmes à une IRM de perfusion de la prostate, nous arrivons au résultat suivant :

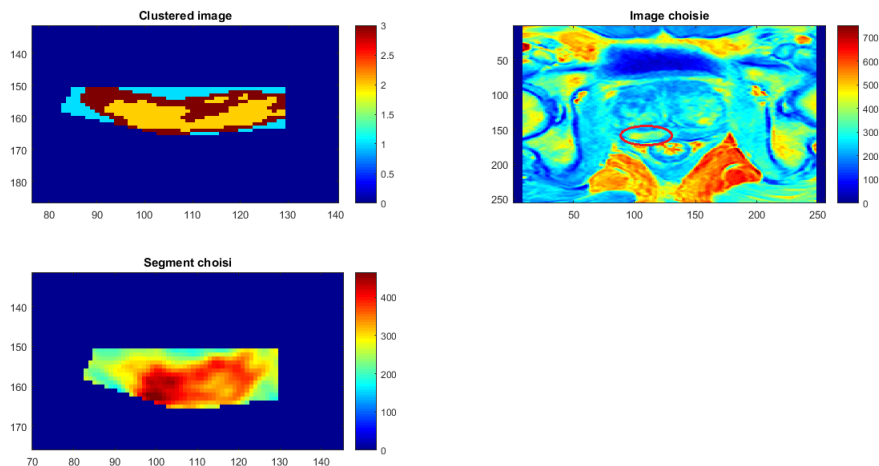


Figure 8.2: Coupes sélectionnées. En haut à gauche le résultat de l'algorithme de classification. En haut à droite, image de la coupe sélectionnée avec la ROI représentée par l'ellipse rouge. En bas à gauche, ROI en fausses couleurs.

La figure suivante montre comment les courbes sont réparties selon les différents clusters.

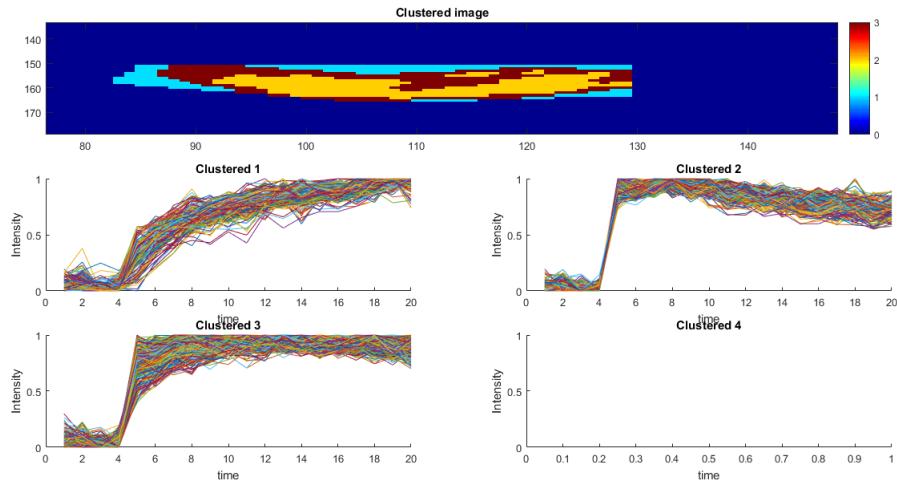


Figure 8.3: Résultat de l'algorithme sur l'IRM avec les courbes associées aux clusters.

Nous avons ici lancé une simulation avec 3 classes. Nous pouvons interpréter ces résultats de la façon suivante:

- Le cluster 1 rassemble les courbes qui sont saines en accord avec les informations issues de la figure 8.1,
- Le cluster 2 lui rassemble les courbes des tissus tumoraux qui présentent un temps d'absorption très court et un pic très marqué,
- Le cluster 3 lui représentent des courbes qui sont suspectes. Ici, l'algorithme comme l'humain va avoir du mal à trancher sur l'appartenance de ces tissus au sain ou tumoral.

Le résultat obtenu est très satisfaisant et on arrive à un résultat semblable à ceux obtenus dans [5] bien que la chaîne de traitement ne soit pas la même.

Néanmoins, il y a quelques remarques à faire:

- Il faut parfois choisir une bonne représentation des données de départ afin que l'algorithme converge vers une solution satisfaisante.
- Il est nécessaire de choisir une ROI qui est composée d'une zone de taille pas trop importante. En effet, l'algorithme ne converge pas si il y a trop de points sélectionnés.



## Chapter 9

### IMRs T2 du crane CHRU Brest Cavale Blanche.

Le CHRU de Brest nous a permis de constituer une base de données de 13 patients présentant des lésions cérébrales caractéristiques sur lesquels nous avons pu tester notre chaîne de traitement. Les deux exemples qui vont être présentés ici sont particulièrement pertinents sur les capacités et limites de la méthode pour cette partie du corps. Les IRM que nous utilisons ici sont des IRM T2 ce qui implique que nous aurons la présence d'un hypo-signal pendant la diffusion du produit de contraste.

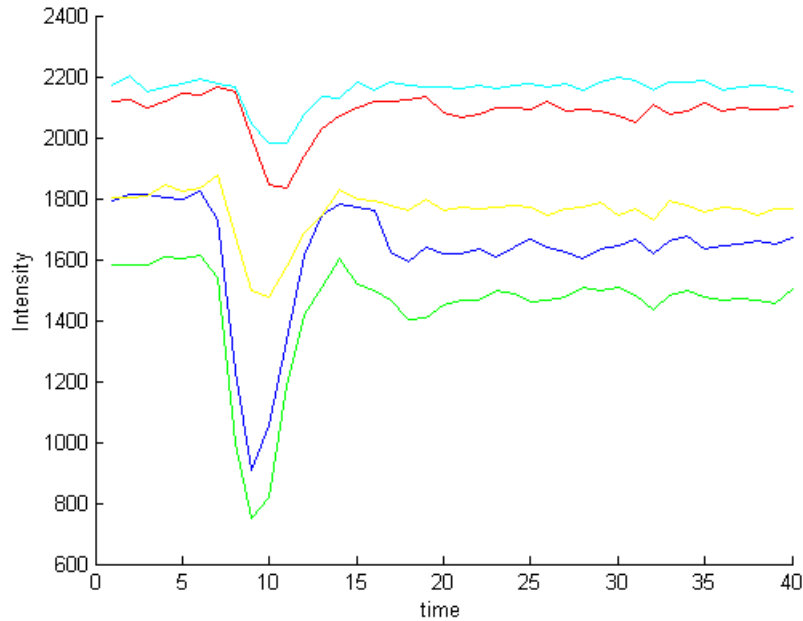


Figure 9.1: Exemples de courbes obtenues par les IRM du CHRU.

Nous pouvons créer deux cartes paramétriques à partir des IRM de perfusion qui peuvent aider pour faire un diagnostic. Le premier paramètre est le CBV ou *cerebral brain volume*. Ce paramètre permet de traduire l'importance d'absorption de produit de contraste par le tissu considéré. Le document [6] définit le CBV par la méthodologie suivante. Le temps de relaxation transversal  $\Delta R2^*$  est donné par la formule:

$$\Delta R2^*(t) = \frac{-\ln(S(t)/S_0)}{TE} \quad (9.1)$$

où  $TE$  est le temps de pulsation d'écho de l'IRM. Il est récupérable sur les méta-données de

l'IRM. Le CBV est donc défini comme l'intégrale de  $R2^*(t)$

$$CBV = \int_{t_0}^{t_f} R2^*(t) dt \quad (9.2)$$

Le second paramètre est le CBF ou *cerebral brain flow*. Il n'est pas pertinent pour l'étude que nous avons menée mais il a été implémenté sur Matlab.

## 9.1 Patient 1: Cancer identifiable par l'IRM de perfusion.

L'IRM qui est ici présentée provient d'un patient qui a un cancer bénin situé à la frontière du cerveau et de la boîte crânienne. La tumeur est visible sur la figure suivante sur la partie gauche de l'image. A sa droite, on peut voir la présence d'un assez gros œdème résultant de la présence de la tumeur. Le reste du cerveau est sain et il y a au centre de l'image le ventricule cérébrale qui produit le liquide céphalo-rachidien. Ce dernier apparaît ici avec un hyper-signal caractéristique.

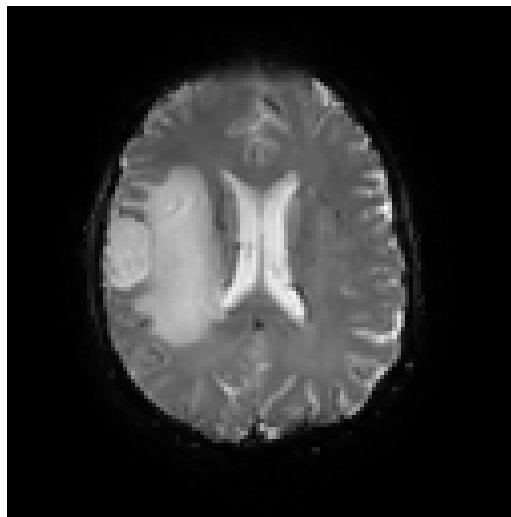


Figure 9.2: IRM du patient présent une tumeur bénigne et un œdème.

Nous avons donc appliqué notre algorithme en sélectionnant une ROI qui ne contient pas le ventricule cérébral afin qu'il ne vienne pas parasiter le résultat de l'algorithme. Nous avons également demandé qu'il ressorte quatre classes. La figure qui suit montre dans l'ordre, l'IRM du patient en fausse couleur, la carte paramétrique du CBV et le résultat de notre algorithme.

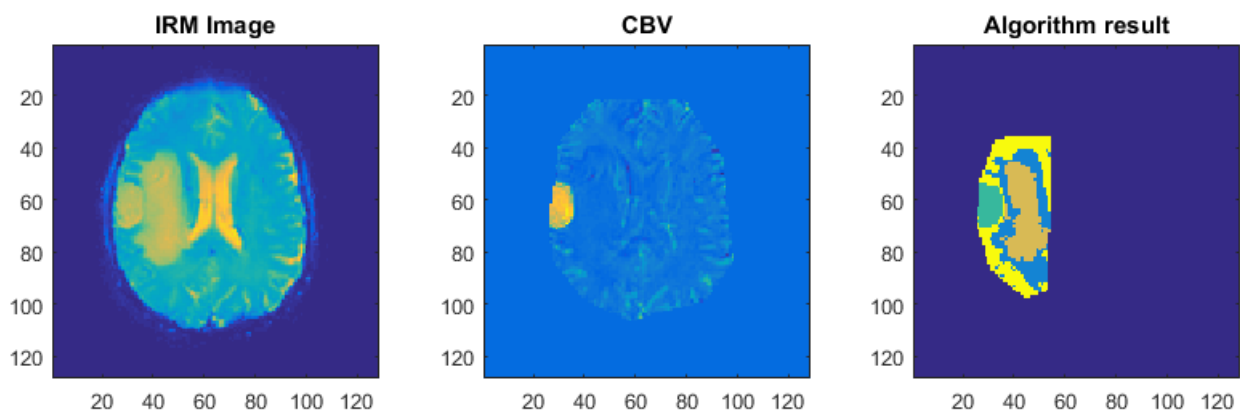


Figure 9.3: Résultat de l'algorithme.

La carte de CBV est ici particulièrement adaptée car elle fait ressortir tout tissu qui a tendance à absorber un grand volume de produit de contraste par rapport au reste du cerveau. Sur cette carte, on voit que la tumeur est immédiatement isolée.

On peut remarquer que l'algorithme que nous avons mis en place permet d'isoler parfaitement la tumeur mais également l'œdème et permet d'établir une zone de frontière entre cette œdème et les tissus sains.

## Conclusion:

On peut voir par cet exemple que notre algorithme est particulièrement adapté pour cette pathologie car cette tumeur impacte beaucoup la diffusion du produit de contraste. Néanmoins, on peut regretter que nous ne pouvons détecter qu'un unique œdème avec notre algorithme.

## 9.2 Patient 2: Cancer non-identifiable par l'IRM de perfusion.

L'IRM qui suit provient d'un patient qui a un cancer situé dans la région en haut à gauche du cerveau. On peut voir un très gros œdème qui est situé tout autour de la tumeur mais cette dernière n'implique pas de modification notable de la diffusion du produit de contraste ce qui se traduit par un CBV qui n'est pas particulièrement modifié.

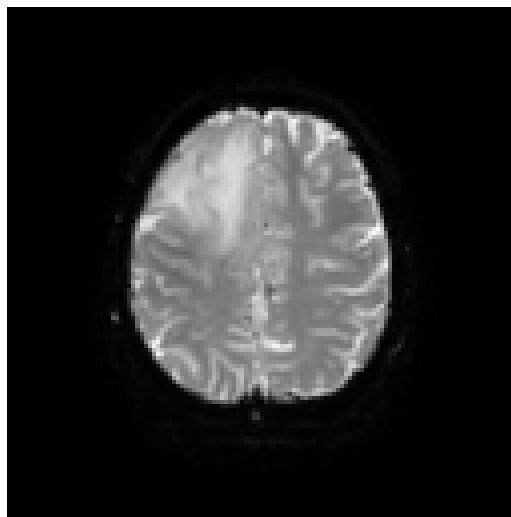


Figure 9.4: IRM du patient présentant une tumeur et un œdème. Ici, seul l'œdème est identifié par notre algorithme.

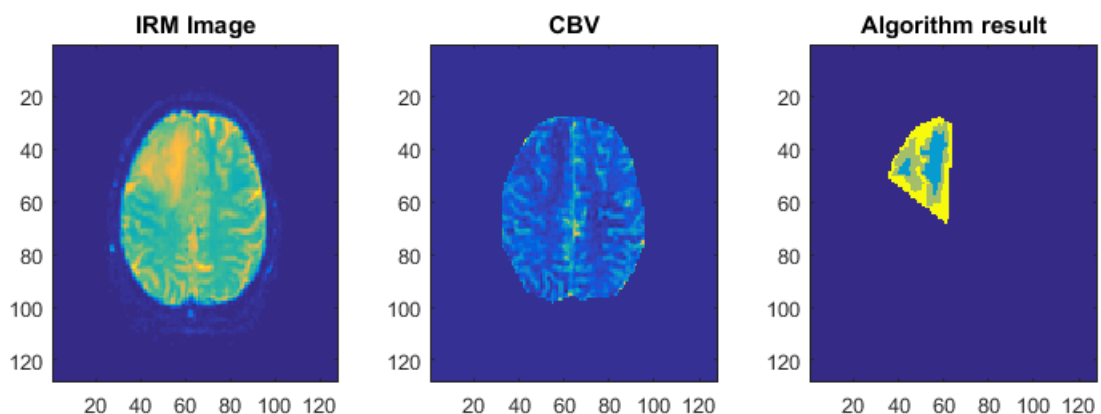
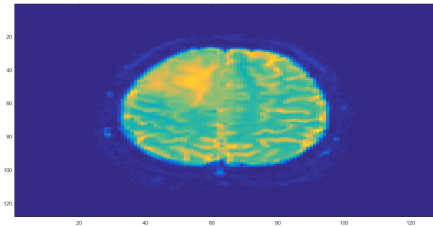
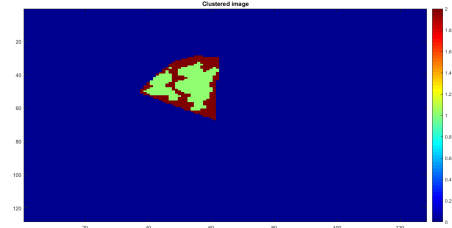


Figure 9.5: Résultat de l'algorithme.

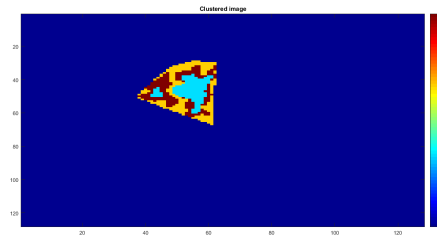
Nous avons testé notre algorithme en demandant la création de 2 et 3 classes et on retrouve un résultat similaire à celui ci-dessus. On observe que dans le cas où on demande deux classes, on a l'œdème qui est identifié et pour 3 classes, on a une zone frontière qui se crée entre le tissu sain et l'œdème. Néanmoins, la tumeur ne peut être identifiée.



(a) Image originale.



(b) Résultat de l'algorithme avec 2 classes.



(c) Résultat de l'algorithme avec 3 classes.

Figure 9.6: Résultats de notre algorithme sur le patient 2 avec 2 et 3 classes.

## Conclusion:

Tant notre algorithme que la carte de CBV sont ici inadapté pour cette pathologie. Bien que on identifie clairement l'œdème, le principal but de notre algorithme est de pouvoir identifier des pathologies plus graves comme les tumeurs. Mais dans certains cas comme celui ci-dessus, il est nécessaire de trouver d'autres sources d'information plus pertinentes pour proposer une segmentation de la zone pathologique.

# **Conclusion**

Ce stage m'a permis de me former sur beaucoup de problématiques en rapport avec le milieu médical et rencontrer plusieurs médecins qui m'ont appuyé. Cela m'a demandé de développer un esprit d'initiative afin de voir et résoudre les problèmes qui sont présentés devant moi durant ces 6 mois. J'ai pu approfondir mes connaissances sur des domaines scientifiques comme la classification non supervisée et j'ai pu voir comment il faut analyser des images médicales. Ce stage fut très intéressant et épanouissant.

L'objectif que nous nous sommes fixé au départ était de réaliser un programme d'aide au diagnostic pour les pathologies cérébrales. Au final, la chaîne de traitement que nous avons mis en place présente plusieurs caractéristiques:

- elle est très fortement automatisée. L'opérateur qui souhaiterait l'utiliser n'aura qu'à sélectionner une zone d'intérêt suspecte sur l'image IRM et le nombre de tissu qu'il semble voir à l'image.
- Pour certaines pathologies, l'algorithme développé permet d'isoler la pathologie de la zone présentant des tissus sains.
- Néanmoins, pour d'autres pathologies, les IRM de perfusion que nous avons utilisé ici ne permettent pas d'aboutir à un résultat.

L'objectif que l'on m'a posé est donc partiellement atteint. Ce résultat est intéressant mais une piste semble prometteuse pour résoudre ce problème. Nous nous sommes ici principalement focalisés sur l'étude d'un seul type d'IRM alors que les médecins se servent de plusieurs modalités d'image IRM pour faire leur diagnostic. L'analyse multimodale semble donc tout indiquée cependant le temps imparti ne m'a pas permis d'implémenter cette solution.

# **Part IV**

## **Annexes**

# Appendix A

## Anonymisation des images DICOM

Pour anonymiser les données, j'utilise le programme "Santé Dicom Editor". Il suffit d'ouvrir le dossier contenant les images et aller dans l'onglet Edit → Anonymise current file/series. Il existe des fonctions Matlab pour faire le même travail mais certains champs sensibles étaient malgré tous intacts. Voir [http://www.santesoft.com/win/sante\\_dicom\\_editor/sante\\_dicom\\_editor.html](http://www.santesoft.com/win/sante_dicom_editor/sante_dicom_editor.html).

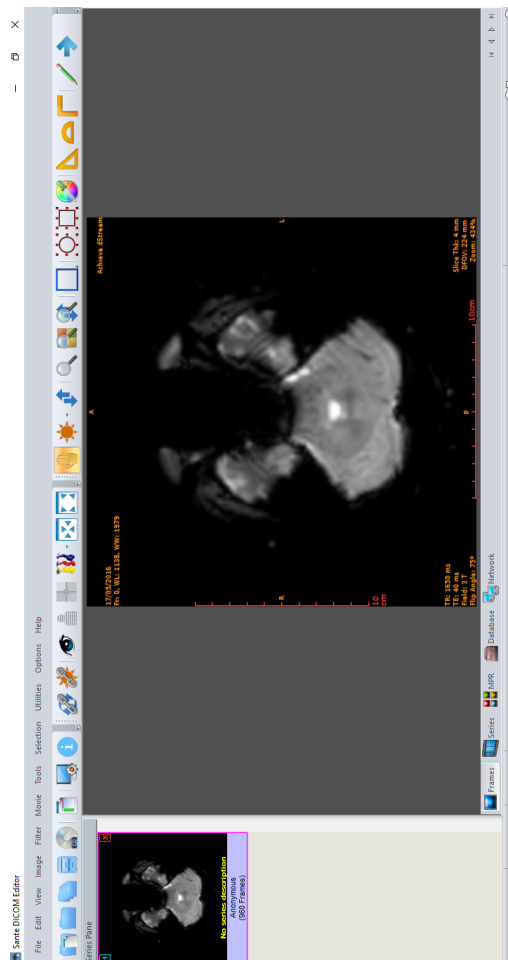


Figure A.1: Programme pour l'anonymisation des données.



# Appendix B

## Visualisation des données en 3D et 4D

Un gros problème lorsque nous recevons les images Dicom est que nous ne pouvons voir de façon très claire la localisation de la tumeur. J'ai donc réutilisé et modifié des codes Matlab disponibles sur internet afin d'avoir une représentation 3D et 4D. Ils sont disponibles sur les liens suivant:

<http://fr.mathworks.com/matlabcentral/fileexchange/41465-4d-volume-visualization>

<http://www.mathworks.com/matlabcentral/fileexchange/37268-3d-volume-visualization>

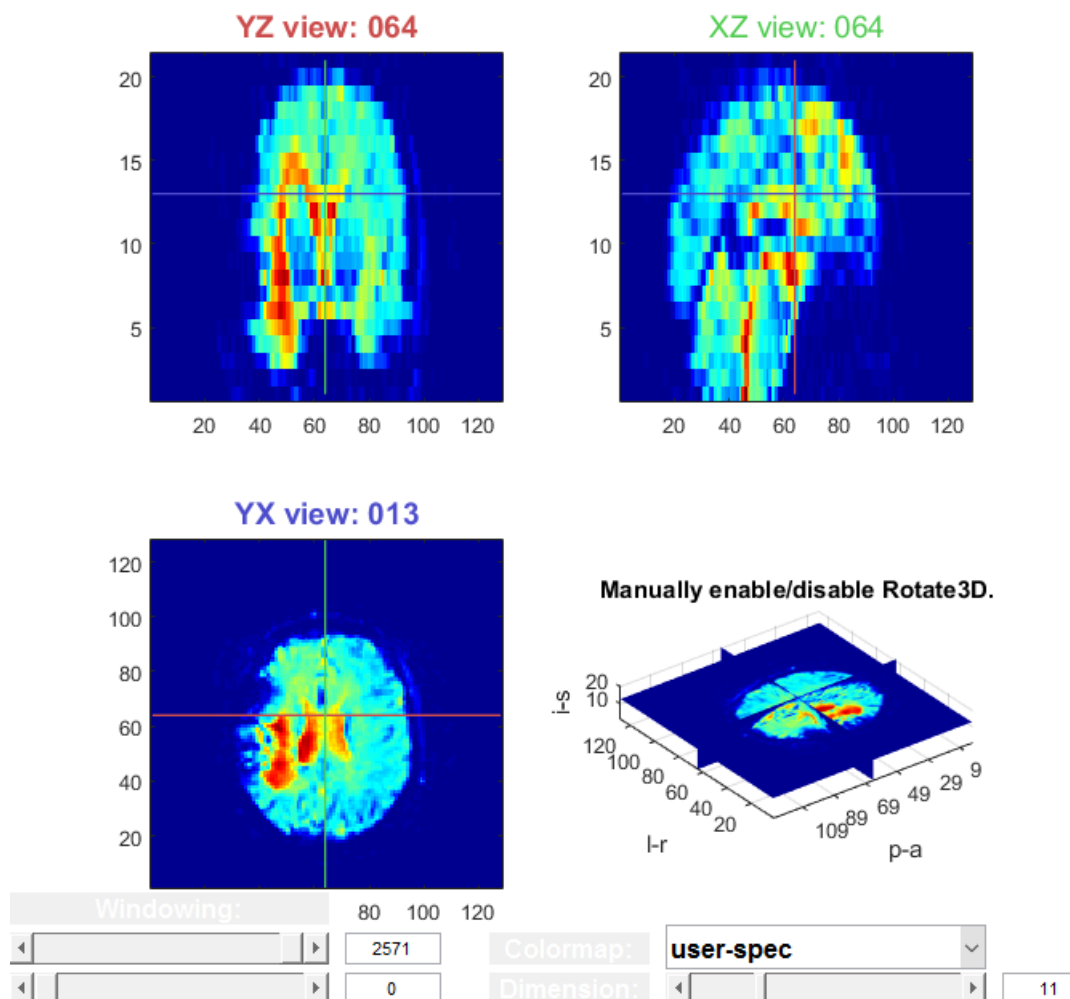


Figure B.1: Programme Matlab pour la visualisation de données en 4D.

# Appendix C

## Choix automatique du nombre de classe.

### Algorithme 1 : Analyse des valeurs propres et le n-gap

Dans le cas d'une classification idéale, la matrice laplacienne  $L$  peut être mise sous la d'une matrice par bloc où tout les blocs non-diagonaux sont nuls.

$$L = \begin{bmatrix} x^{(1)} & 0 & \dots & 0 \\ 0 & x^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & x^{(n)} \end{bmatrix}$$

Cela traduit le fait que les blocs ou clusters sont homogènes et ne présentent pas de similitude entre eux. Cependant, vu comment la matrice  $L$  est construite, la valeur propre 0 va avoir une multiplicité égale au nombre de bloc.

Une première méthode consiste donc à afficher les  $n$  premières valeurs propres de  $L$  et de "compter" le nombre de fois que la valeur propre 0 apparait. Dans le cas des données simulées ci-dessous:

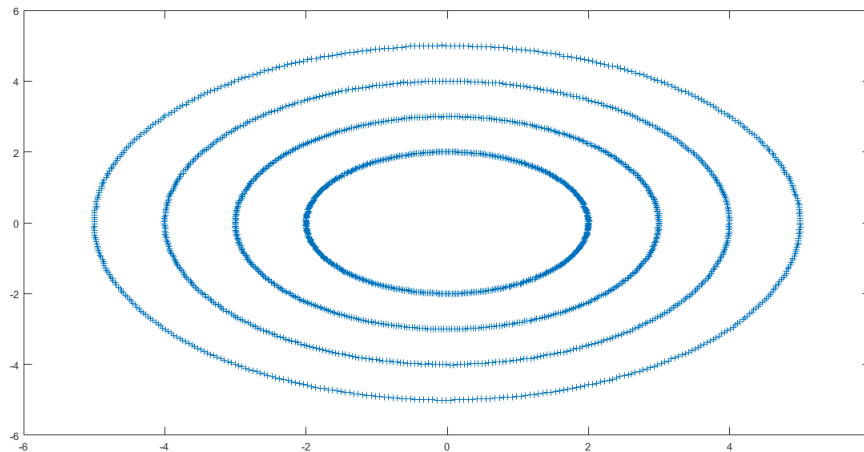


Figure C.1: Exemple simple de données à classifier. Le nombre de classe est 4.

Nous arrivons au valeurs propres suivantes:

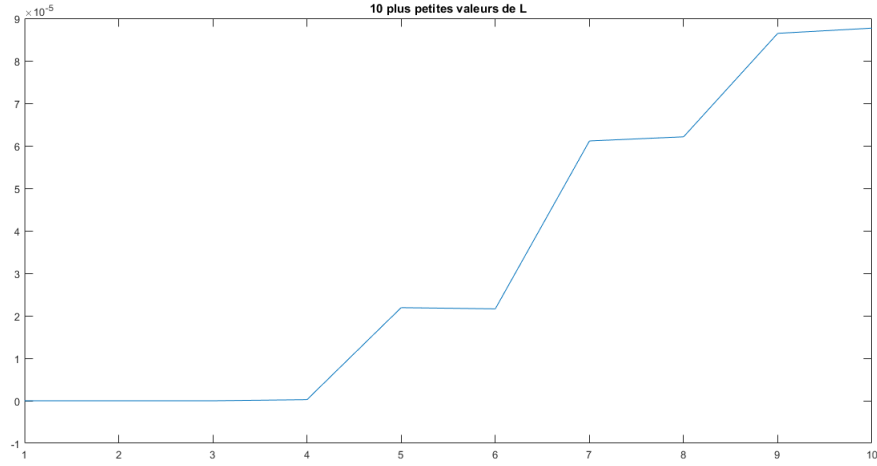


Figure C.2: 10 plus petites valeurs propres de L. On observe bien que on a un saut à partir de  $n=5$ . Le nombre optimal de classe est 4.

Une autre méthode consiste à calculer la différence entre deux valeurs propres successives et voir à partir de quand un gap significatif se produit afin d'identifier le nombre de classe optimale. Néanmoins, ces deux méthodes sont hautement discutables de par leur manque de rigueur.

## Algorithme 2 : Minimisation par la norme de Frobenius

Une autre méthode consiste à voir pour quel nombre de classe la similitude entre classe est la plus faible. On définit la norme de Frobenius d'une matrice par la formule suivante:

$$\|L\|_F = \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} |L_{ij}|$$

L'idée va donc être de calculer le volume des blocs diagonaux et de calculer une fonction coût qui dépend du nombre de classe.

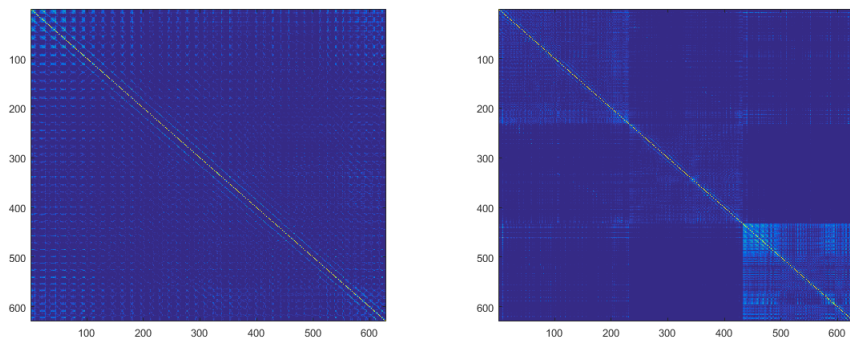


Figure C.3: Exemple de matrice laplacienne avant et après utilisation de l'algorithme de classification spectrale en trois classes et ré-agencement. On observe bien l'apparition de blocs.

On définit donc le ratio de Frobenius entre deux blocs de la matrice L par:

$$r_{ij} = \frac{\|L^{ij}\|_F}{\|L^{ii}\|_F}$$

On va donc regarder les valeurs de  $r_{ij}$  et voir pour quel valeur de  $n$  on atteint un minimum. Le but de cet algorithme va consister à trouver la valeur de  $n$  qui va minimiser la fonction :

$$\nu_k = \frac{2}{k(k-1)} \sum_{i=1, j=i+1}^k (r_{ij})$$

Il faut donc regarder pour plusieurs valeurs de  $n$  quel est le nombre de classe qui minimise cette fonction.

## **Application de ces algorithmes sur des données réelles.**

Sur des données simulées, les algorithmes précédents donnent de bons résultats mais cela n'est absolument pas le cas pour les données issues des IRM de cerveau. Les algorithmes ne semblent pas capables de déterminer de façon claire un nombre de classe optimal pour les signaux.

## Appendix D

### Réalisation de programme pour le milieu médical en stand-alone.

La société MathWorks propose un outil pour la réalisation de programme utilisant directement les fichiers Matlab. A partir d'une interface qui peut être très facilement créée par l'outil GUIDE de Matlab, on peut assez facilement proposer un programme qui peut être installé sur tout ordinateur ayant Windows comme système d'exploitation.

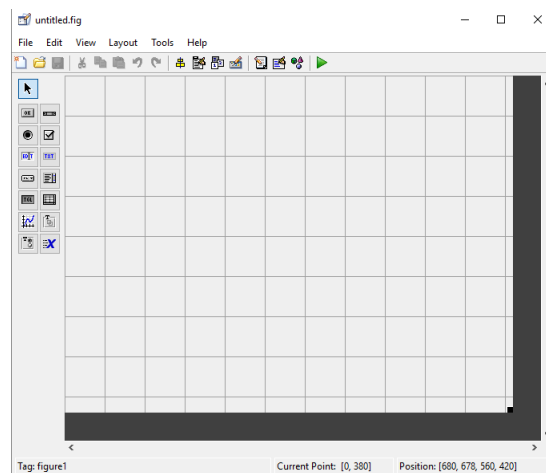


Figure D.1: Présentation de l'outil de Guide de Matlab.

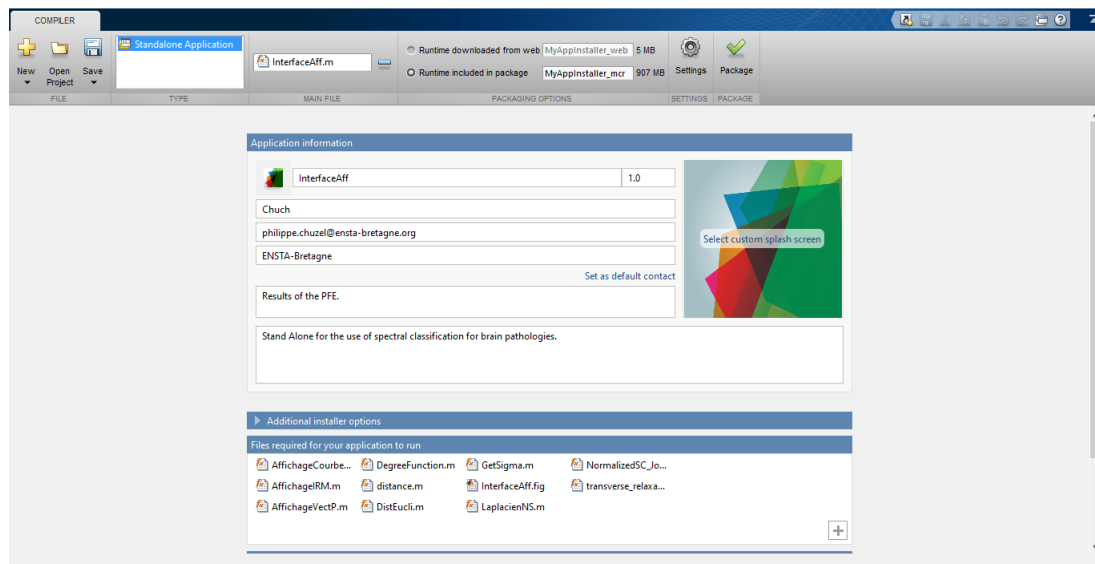


Figure D.2: Présentation de l'outil de Application Compiler de Matlab.

Néanmoins, l'exécutable qui est généré est très lourd, autour de 1GB ce qui est assez contraignant. Ce sont surtout les dépendance avec les bibliothèques en C et C++ qui impose un tel volume.

La meilleur solution si nous souhaitons avoir un programme entièrement portable est donc de passer le code en C++ et de réaliser une interface avec une bibliothèque adaptée du type QT.

# Appendix E

## Rappel sur la théorie des graphes.

### E.1 Définitions de base.

Un graphe est constitué du couple  $(V, E, W)$ , où  $V$  correspond à l'ensemble des nœuds du graphe,  $E$  correspond à l'ensemble des couple  $(v_i, v_j)$  symbolisant le lien entre deux nœuds et  $W$  correspond à la matrice de voisinage. Par définition, le coefficient  $W_{ij}$  prend comme valeur le poids liant le nœud  $v_i$  au nœud  $v_j$ . Il s'en suit que la matrice  $W$  est symétrique. Nous verrons plus tard comment nous pouvons construire une telle matrice.

Par la suite, nous définissons la matrice  $D$ , diagonale, qui contient les degrés des nœuds du graphe. Le degré d'un nœud est défini par  $d_i = \sum_{v_j \in V} W_{ij}$ .

### E.2 Choix du laplacien de la matrice.

Il existe trois définitions différentes de laplacien d'une matrice. Si on appelle  $D$  la matrice de degré du graphe et  $W$  la matrice de voisinage du graphe, on aboutit aux définitions suivantes.

**Matrice Laplacienne standard non normalisée** Ce laplacien est donné par la relation:

$$L = D - W \quad (\text{E.1})$$

**Matrice Laplacienne normalisée asymétrique** Ce laplacien est donné par la relation:

$$L = I - D^{-1}W \quad (\text{E.2})$$

**Matrice Laplacienne normalisée symétrique** Ce laplacien est donné par la relation:

$$L = I - D^{-1/2}WD^{-1/2} \quad (\text{E.3})$$

Ces trois matrices présentent les mêmes propriétés:

- Ces matrices présentent la valeur propre 0 une seule fois pour le vecteur propre  $I$  ou  $D^{1/2}I$ .
- Ces matrices sont réelles semi-définies positives.

Par conséquent, quand nous chercherons à trouver les  $k$  plus petites valeurs propres de cette matrice, il nous sera possible très facilement de les récupérer.

# Appendix F

## Construction de la matrice de voisinage.

Il y a trois choix à faire pour définir la matrice de voisinage d'un graphe.

### F.1 Choix du type de voisinage.

Pour un ensemble de nœud donné, il y a plusieurs manière de créer les liens qui vont compléter l'ensemble  $(E, V, W)$ .

**Graphe des epsilon voisins:** On considère ici que deux nœuds ont un lien si la distance entre eux est inférieur à  $\epsilon$ .

**Graphe des k-plus proche voisins:** On considère ici que tous les nœuds ont k voisins et qu'ils sont les k-plus proche du nœud considéré.

**Graphe entièrement connecté:** On considère que tous les sommets sont reliés entre eux.

Ici, nous avons implémenté le graphe entièrement connecté car il nous donne de très bons résultats et qu'il est très simple à mettre en place.

### F.2 Choix de la distance et de la similitude.

Dans la littérature de la classification, il est nécessaire de choisir une distance et une fonction de similarité.

Plusieurs choix s'offre à nous pour les distances:

- **Distance de Manhattan :**  $\sum_i |x_i - y_i|$
- **Distance d'euclidienne :**  $(\sum_i (x_i - y_i)^2)^{(1/2)}$
- **Distance de Minkowski :**  $(\sum_i |x_i - y_i|^p)^{(1/p)}$

La distance euclidienne a ici été choisie.

La fonction de similitude doit être une fonction qui va associer une valeur comprise entre 0 et 1 à deux sommet du graphe:

$$f(x_i, x_j) \Rightarrow [0; 1] \tag{F.1}$$

La littérature donne plusieurs types de fonction de similitude mais deux sont à retenir.



- La similitude euclidienne qui est définie par la fonction

$$f(x_i, x_j) = \frac{1}{1 + \frac{d^2(x_i, x_j)}{\sigma^2}} \quad (\text{F.2})$$

- La similitude gaussienne qui est définie par la fonction

$$f(x_i, x_j) = \exp\left(-\frac{d^2(x_i, x_j)}{2\sigma^2}\right) \quad (\text{F.3})$$

Dans les deux cas, le coefficient  $\sigma$  est un paramètre qui traduit la dispersion des données. Nous avons ici choisi la similitude gaussienne car elle permet de traduire de manière plus efficace la similitude des données.

### F.3 Choix du paramètre de dispersion.

Le gros problème pour les données que nous aurons à traiter est que le paramètre  $\sigma$  devra être déterminé de manière automatique en fonction des données d'entrées. Ce paramètre va traduire la similitude ou non des différents éléments entre eux et donc va conditionner tous les résultats de l'algorithme.

Pour ce faire, une solution assez élégante a été proposée [7] et consiste à définir le paramètre  $\sigma$  comme le produit  $\sigma_i \sigma_j$ , où ces deux coefficients dépendent du voisinage de chaque point.

Ainsi, on définit le coefficient par  $\sigma_i = d(x_i, x_r)$  où  $x_r$  correspond au  $r$ -ième voisin le plus proche de  $x_i$ . Après beaucoup de test, on constate comme dans le document [1] que les meilleurs résultats sont obtenus pour  $r = 7$  ou  $r = 9$ .

# Appendix G

## Quels algorithmes de classification spectrale?

Les algorithmes de classification spectrale sont répartis en deux classes:

- La première classe qui va faire une séparation en deux groupes de manière récursive à partir de la seconde plus petite valeur propre de la matrice laplacienne issue du graphe jusqu'à avoir  $k$  clusters.
- La seconde consiste à faire une classification sur les  $k$  plus petits vecteurs propres de la matrice laplacienne issue du graphe grâce à un algorithme plus simple comme le K-means.

Dans le cadre de ce projet, nous avons principalement étudié la seconde classe d'algorithme de classification et particulièrement celui développé par Jordan and Weiss. Au total, nous avons implémenté 4 algorithmes différents de classification spectrale. Les détails théoriques de ces algorithmes peuvent être trouver dans ces références [8] et [9].

Les trois premiers algorithmes correspondent à la seconde classe d'algorithme alors que le dernier algorithme appartient à la première classe d'algorithme:

---

### Algorithm 2 Unnormalized spectral clustering

---

#### Entrées

- 1: Matrice  $W$  matrice de voisinage
- 2: Entier  $k$  nombre de classe

#### Sorties

- 1: Vecteur  $Ver$  table de vérité calculée.

#### Algorithme

- 1:  $L \leftarrow$  Laplacien standard de la matrice  $W$
  - 2:  $VectP \leftarrow$  matrice contenant les vecteurs propres des  $k$  plus petites valeurs propres non nulles de  $L$
  - 3:  $Ver \leftarrow$  résultat de l'algorithme du Kmeans sur la matrice  $VectP$  en  $k$  clusters.
  - 4: **return**  $Ver$
-

---

**Algorithm 3** Normalized spectral clustering, Shi and Malik

---

**Entrées**

- 1: Matrice  $W$  matrice de voisinage
- 2: Entier  $k$  nombre de classe

**Sorties**

- 1: Vecteur  $Ver$  table de vérité calculée.

**Algorithme**

- 1:  $L \leftarrow$  Laplacien asymétrique de la matrice  $W$
  - 2:  $VectP \leftarrow$  matrice contenant les vecteurs propres des  $k$  plus petites valeurs propres non nulles de  $L$
  - 3:  $Ver \leftarrow$  résultat de l'algorithme du Kmeans sur la matrice  $VectP$  en  $k$  clusters.
  - 4: **return**  $Ver$
- 

---

**Algorithm 4** Normalized spectral clustering, Jordan and Weiss

---

**Entrées**

- 1: Matrice  $W$  matrice de voisinage
- 2: Entier  $k$  nombre de classe

**Sorties**

- 1: Vecteur  $Ver$  table de vérité calculée.

**Algorithme**

- 1:  $L \leftarrow$  Laplacien symétrique de la matrice  $W$
  - 2:  $VectP \leftarrow$  matrice contenant les vecteurs propres des  $k$  plus petites valeurs propres non nulles de  $L$
  - 3: normaliser toutes les lignes de la matrice  $VectP$
  - 4:  $Ver \leftarrow$  résultat de l'algorithme du Kmeans sur la matrice  $VectP$  en  $k$  clusters.
  - 5: **return**  $Ver$
- 

---

**Algorithm 5** Normalized spectral recursive clustering, Shi and Malik

---

**Entrées**

- 1: Matrice  $W$  matrice de voisinage
- 2: Entier  $k$  nombre de classe

**Sorties**

- 1: Vecteur  $Ver$  table de vérité calculée.

**Algorithme**

- 1:  $L \leftarrow$  Laplacien asymétrique de la matrice  $W$
  - 2:  $VectP \leftarrow$  seconde plus petite valeur propre de  $L$
  - 3: Partitionner le graphe en deux sous-ensemble selon les valeurs de  $VectP$
  - 4: Répéter l'algorithme jusqu'à avoir  $k$  clusters.
  - 5:  $Ver \leftarrow$  résultat de l'algorithme.
  - 6: **return**  $Ver$
-

# Bibliographie

- [1] G. Tartare, “Contribution à l’analyse de l’IRM dynamique pour l’aide au diagnostic du cancer de la prostate,” Ph.D. dissertation, Université du Littoral-Côte-d’Opale, Lille, France, 2014.
- [2] J. Ophir, I. Cespedes, H. Ponnekanti, Y. Yazdi, and X. Li, “Elastography: a quantitative method for imaging the elasticity of biological tissues,” *Ultrasonic imaging*, vol. 13, no. 2, pp. 111–134, 1991.
- [3] G. Tartare, D. Hamad, M. Azahaf, P. Puech, and N. Betrouni, “Spectral clustering applied for dynamic contrast-enhanced mr analysis of time–intensity curves,” *Computerized Medical Imaging and Graphics*, vol. 38, no. 8, pp. 702–713, 2014.
- [4] S. Mouysset, “Contributions à l’étude de la classification spectrale et applications,” Ph.D. dissertation, Institut National Polytechnique de Toulouse-INPT, France, 2010.
- [5] K. Doi, “Computer-aided diagnosis in medical imaging: historical review, current status and future potential,” *Computerized medical imaging and graphics*, vol. 31, no. 4, pp. 198–211, 2007.
- [6] E. R. Muir, L. T. Watts, Y. V. Tiwari, A. Bresnen, Q. Shen, and T. Q. Duong, “Quantitative cerebral blood flow measurements using MRI,” *Cerebral Angiogenesis: Methods and Protocols*, pp. 205–211, 2014.
- [7] L. Zelnik-Manor and P. Perona, “Self-tuning spectral clustering,” in *Advances in neural information processing systems*, 2004, pp. 1601–1608.
- [8] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.
- [9] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
- [10] S. Temme, C. Grapentin, C. Quast, C. Jacoby, M. Grandoch, Z. Ding, C. Owenier, F. Mayenfels, J. W. Fischer, and R. Schubert, “Noninvasive imaging of early venous thrombosis by 19f magnetic resonance imaging with targeted perfluorocarbon nanoemulsions,” *Circulation*, vol. 131, no. 16, pp. 1405–1414, 2015.
- [11] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.
- [12] A. Idbaih, M. Boukobza, I. Crassard, R. Porcher, M.-G. Bousser, and H. Chabriat, “MRI of clot in cerebral venous thrombosis high diagnostic value of susceptibility-weighted images,” *Stroke*, vol. 37, no. 4, pp. 991–995, 2006.
- [13] D. S. Liebeskind, N. Sanossian, W. H. Yong, S. Starkman, M. P. Tsang, A. L. Moya, D. D. Zheng, A. M. Abolian, D. Kim, and L. K. Ali, “CT and MRI early vessel signs reflect clot composition in acute stroke,” *Stroke*, vol. 42, no. 5, pp. 1237–1243, 2011.

- [14] C. Schmitt, A. H. Henni, and G. Cloutier, "Characterization of blood clot viscoelasticity by dynamic ultrasound elastography and modeling of the rheological behavior," *Journal of bio-mechanics*, vol. 44, no. 4, pp. 622–629, 2011.
- [15] C. Schmitt, E. Montagnon, A. H. Henni, S. Qi, and G. Cloutier, "Shear wave induced resonance elastography of venous thrombi: A proof-of-concept," *IEEE Trans. on Medical Imaging*, vol. 32, no. 3, pp. 565–577, 2013.
- [16] M. Meilă, "Comparing clusterings, an information based distance," *Journal of multivariate analysis*, vol. 98, no. 5, pp. 873–895, 2007.
- [17] A. Y. Yang, Z. Zhou, A. G. Balasubramanian, S. S. Sastry, and Y. Ma, "Fast-minimization algorithms for robust face recognition," *IEEE Trans. on Image Processing*, vol. 22, no. 8, pp. 3234–3246, 2013.
- [18] S. Wu, X. Feng, and W. Zhou, "Spectral clustering of high-dimensional data exploiting sparse representation vectors," *Neurocomputing*, vol. 135, pp. 229–239, 2014.
- [19] J. Jiao, X. Mo, and C. Shen, "Image clustering via sparse representation," in *Advances in Multimedia Modeling*. Germany: Springer, 2010, pp. 761–766.
- [20] S. Yan and H. Wang, "Semi-supervised learning by sparse representation." in *SDM*. SIAM, 2009, pp. 792–801.
- [21] D. L. Longo, W. Dastrù, L. Consolino, M. Espak, M. Arigoni, F. Cavallo, and S. Aime, "Cluster analysis of quantitative parametric maps from DCE-MRI: application in evaluating heterogeneity of tumor response to antiangiogenic treatment," *Magnetic resonance imaging*, vol. 33, no. 6, pp. 725–736, 2015.