Reconnaissance de types cellulaires à partir d'images hyperspectrales : Comparaison de modèles statistiques.

Laurent Chapillon

Université de Bourgogne Dijon

2019-2020

Rapport de Stage Master II
« Traitement du Signal et de l'Image »

Reconnaissance de types cellulaires à partir d'images hyperspectrales : Comparaison de modèles statistiques.

Laurent CHAPILLON

Sous la direction de Aurélien CHATEIGNER, Clément CUELLO et Julien MILLE

Remerciements

Je remercie dans un premier temps Messieurs Clément Cuello et Aurélien Chateigner pour leur disponibilité à mon égard, leurs judicieux conseils, leur qualité d'écoute et leur encadrement tout au long du stage.

Je remercie également Monsieur Julien Mille qui, tout au long du stage, a partagé avec moi ses précieuses connaissances vis-à-vis du traitement d'images et des modèles statistiques.

Je remercie également Monsieur Johel Miteran pour l'intérêt qu'il a accordé au bon déroulement du stage compte tenu de la crise sanitaire.

Enfin, je tiens à remercier l'ensemble des membres de l'équipe BioForA pour m'avoir aussi bien accueilli, ainsi que le RTR-DIAMS pour avoir financé ce stage.

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal

laurent.chapillon@wanadoo.fr

Bilan personnel

Selon moi, les stages de fin d'études sont la consécration de la formation de master où tout le savoir acquis est mis en œuvre afin d'atteindre un objectif défini. Durant celui-ci, j'ai eu l'occasion d'apprendre et de développer des connaissances qui peuvent être liées au domaine du traitement d'images, ce qui sera un plus pour moi dans le futur.

Ce stage de fin d'étude m'a fortement responsabilisé vis-à-vis du travail au niveau ingénieur, et m'ont fait comprendre l'importance de la communication afin de rapidement répondre à des problématiques de groupe ou individuelles.

Aussi, durant cette période, j'ai également appris l'importance d'effectuer des recherches bibliographiques afin de parfaire ses connaissances ou de donner des idées afin de mener le projet à bien.

Enfin, ce stage a été exceptionnel en raison de la crise sanitaire qui nous a mis dans une position d'isolement. C'est là que la communication est importante afin de garder la flamme et pouvoir avancer dans le projet.

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal

laurent.chapillon@wanadoo.fr

Liste des abréviations

ACP	Analyse en composantes principales
ATR-FTIR	attenuated total reflectance - Fourier transformed infrared
Couche G / g / FibG	couche gélatineuse
F / FibS	fibres
FN	faux négatif
FP	faux positif
INRA	institut national de la recherche agronomique
INRAE	institut national de recherche pour l'agriculture, l'alimentation et l'environnement
INSA	institut national des sciences appliquées
IRSTEA	institut de recherche en sciences et technologies pour l'agriculture et l'environnement
kNN	k-nearest neighbors
LDA	linear discriminant analysis
LICA	laboratoire d'ingénierie cellulaire de l'arbre
LIFAT	laboiratoire d'informatique fondamentale et appliquée de Tours
NW	bois normal
ONF	office nationale des forêts
OW	bois opposé
R / Ray	rayon
RTR-DIAMS	réseau thématique de recherche en données, intelligence artificielle, modélisation et simulation
RVB	rouge, vert, bleu
SVM	support vector machine
SVM_HP	support vector machine avec recherche d'hyperparamètre
SVM_HP_reeq	support vector machine avec recherche d'hyperparamètre et rééquilibrage des données
SVM_reeq	support vector machine avec rééquilibrage des données
TP	vrai positif
TW	bois de tension
UMR BioForA	unité mixte de recherche 'Biologie intégrée pour la valorisation de la diversité des arbres et de la forêt'
V / Ves	vaisseau

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal

laurent.chapillon@wanadoo.fr

SOMMAIRE

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal

laurent.chapillon@wanadoo.fr

Introduction

A. Organisation des structures d'accueil

1.L'institut National de recherche pour l'agriculture, l'alimentation et l'environnement

L'institut national de la recherche agronomique (INRA) fut créé en 1946. Ce nouvel institut a pour objectif de faire face à la pénurie alimentaire frappant la France au lendemain de la seconde guerre mondiale.

La mission de l'INRA est d'associer science et technologie afin d'améliorer les techniques d'agriculture et d'élevage. En 1960, l'objectif est atteint : l'agriculture française subvient aux besoins de la France. L'INRA est alors encouragé à se développer plus localement en créant différents pôles régionaux.

Dès 1970, la France devient excédentaire en matière de denrée alimentaire. L'INRA se fixe alors de nouveaux objectifs en matière de qualité et de valeur ajouté.

Suite à la fusion avec l'institut de recherche en sciences et technologies pour l'environnement et l'agriculture (IRSTEA), l'INRA devient, au 1er janvier 2020, l'institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE).

INRAE a pour mission de répondre à des questions concrètes de société dans le domaine de l'environnement, de l'agriculture et de l'alimentation en produisant des connaissances nouvelles et des innovations techniques utiles aux gestionnaires, aux décideurs et aux entreprises.

Ses thèmes de recherche sont centrés dans les domaines de l'agriculture, de l'alimentation, de la forêt, de l'environnement, de l'eau, de la biodiversité, de la bioéconomie, de l'économie circulaire et de la gestion durable des territoires.

2.Le centre INRAE Val De Loire

Ce centre est créé le 1er janvier 2013 par la fusion des centres INRA d'Orléans et de Tours, puis renforcé en 2020 par l'intégration du centre IRSTEA de Nogent-sur-Vernisson.

Le centre INRAE Val de Loire effectue des recherches pour obtenir une meilleure durabilité des ressources naturelles et des systèmes agricoles et forestiers ainsi que la biodiversité qui leur est associée. Grâce à plusieurs plateformes technologiques, des analyses peut être effectuées à différentes échelles : moléculaire, individuelle ou encore au niveau d'une population ou d'un écosystème.

Répartie sur environ 1 500 hectares, les différents dispositifs expérimentaux du centre permettent de développer de nouveaux modèles ainsi que d'étudier des ressources génétiques, animales ou végétales.

Ce centre, composé 980 agents, mène des recherches autour de quatre axes : (i) la dynamique des sols et la gestion de l'environnement, (ii) la biologie intégrative des arbres et de la biodiversité associée pour une gestion durable des écosystèmes forestiers, (iii) la biologie animale intégrative et la durabilité des systèmes d'élevage et (iv) l'infectiologie et « One Health ».

3.L'UMR BioForA

L'unité mixte de recherche « Biologie intégrée pour la valorisation de la diversité des arbres et de la forêt » (UMR BioForA) est située à l'interface entre recherche et gestion forestière intégrant du personnel INRAE et ONF (Office Nationale des Forêts). Elle dispose de diverses compétences en génétique, génomique et physiologie appliquée à l'étude des arbres forestiers. Les recherches conduites par l'unité visent à une valorisation des ressources génétiques forestières ayant pour but une production durable de bois d'œuvre et de biomasse. Ces recherches prennent également en compte l'impact écologique que les populations domestiquées ont sur l'écosystème dans un contexte climatique changeant.

L'UMR BioForA dirige plusieurs programmes d'amélioration génétique sur six espèces forestières et investit dans des programmes innovant en sélection et diffusion du progrès génétique. Elle est également impliquée dans l'évaluation et la gestion de la

diversité génétique et l'étude des interactions entre les variétés améliorées et leurs pendants sauvages.

Dans l'intention de remplir ces objectifs, une approche de biologie intégrative est développée notamment sur le peuplier afin de mieux comprendre la génétique participant au développement de l'arbre ainsi que son adaptation aux contraintes liées à son environnement.

4. Le laboratoire d'ingénierie cellulaire de l'arbre

Entré en service en janvier 2018, le laboratoire d'ingénierie cellulaire de l'arbre (LICA) est un laboratoire-serre de confinement niveau 2, cogéré par l'UMR BioForA et l'unité expérimentale 'Génétique et Biomasse Forestière d'Orléans'. Il est dédié à la production et la caractérisation d'arbres transformés par génie génétique. Il offre un appui sur tous les projets liés à la biologie intégrative sur le fonctionnement des arbres.

5. L'institut national des sciences appliqué

L'institut national des sciences appliqué (INSA) est un groupe d'écoles d'ingénieurs regroupant sept établissements en France. Crée dans les années 1950 suite à la demande grandissante d'ingénieurs au lendemain de la seconde guerre mondiale. Chacun de ces établissements propose des formations diverses. L'INSA Centre Val de Loire est réparti sur deux campus, l'un à Bourges, l'autre à Blois. Parmi les trois premiers INSA de France avec Toulouse et Lyon, il propose une formation variée allant du paysagisme au génie mécanique en passant par l'informatique industrielle et la maitrise des risques industriels. Comme tout institut de formation post-bac, l'INSA dispose de diverses unités de recherche dont le laboratoire d'informatique fondamentale et appliquée de Tours (LIFAT).

6. Le Laboratoire d'Informatique Fondamentale et Appliquée de Tours

Les recherches menées au LIFAT visent à concevoir et développer des modèles, méthodes et algorithmes ainsi que des ressources et logiciels permettant d'obtenir de bons résultats en un temps de calcul raisonnable. Ces recherches s'appliquent a trois domaines principaux : la santé, le big data et les sciences humaines numériques. Le LIFAT est divisé en trois groupes de recherche effectifs : (i) bases de données et traitement du langage naturel, (ii) recherche opérationnelle, ordonnancement et transport et (iii) reconnaissance de formes et analyse d'images. Ce dernier groupe focalise ses recherches dur l'apprentissage machine, l'exploration de données et le

traitement d'images. Il se concentre notamment sur le développement de méthodes interactives intégrant l'utilisateur et ses connaissances préalables dans les processus de reconnaissance et d'exploration visuelle des données.

7.RTR-DIAMS

Le Réseau Thématique de Recherche Données, Intelligence Artificielle, Modélisation et Simulation (RTR-DIAMS), est un réseau de chercheurs en région Centre-Val de Loire qui a pour objectif de structurer l'ensemble des scientifiques de la région concernés par les thématiques du numérique et de la donnée afin de favoriser le montage de projets de recherche. Ce réseau est financé par la région Centre Val de Loire pour mener des recherches pour une durée de 4 ans.

Ce réseau est organisé autour de 4 thèmes principaux : (i) la modélisation et la simulation comprenant la modélisation de fluides complexes, la modélisation multi-échelle en physique statistique, la spatialisation du sol ou le séquençage d'ADN en parallèle ; (ii) le traitement de données articulé autour de données textuelle, comme l'apprentissage automatique, linguistique computationnelle et linguistique de corpus, l'acquisition et l'exploitation de base de connaissances à partir de corpus textuel ; (iii) l'apprentissage, l'optimisation et l'aide à la décision, cela concerne le machine learning et son optimisation ; (iv) l'éthique et les bonnes pratiques scientifique. Dans ce contexte scientifique, des chercheurs de l'UMR BioForA et du LIFAT se sont associés dans le cadre d'un groupe de travail ayant pour but de segmenter des images hyperspectrales.

B. Le peuplier et l'anatomie de son bois

Le peuplier, de par sa rapidité de croissance et la qualité de son bois, a vu son exploitation augmenter au niveau mondial. Cet arbre est originaire des régions tempérées à froides de l'hémisphère Nord. Son bois est utilisé pour l'industrie du déroulage (emballage, contreplaqué) et, dans une moindre mesure, dans l'industrie du papier et du bois énergie. De plus, le peuplier dispose d'un intérêt en recherche fondamentale. Celui-ci permet la création de clone en replantant simplement une tige afin d'obtenir un nouvel individu. Aussi, la possibilité de transformation génétique fait de ce dernier une espèce modèle en biologie. Ainsi, divers laboratoires utilisent cet arbre afin de comprendre divers processus comme la formation du bois.

Les arbres peuvent atteindre des hauteurs et des durées de vie considérables grâce aux propriétés de leur bois. Le bois possède trois fonctions dans l'arbre : (i) la

conduction de l'eau, (ii) le support et le maintien de l'arbre vis-à-vis de son environnement et (iii) le stockage de réserves temporaire (Plomion et al, 2001 ; Laurans et al, 2006 ; Déjardin et al, 2010). Chez le peuplier, différents types cellulaires sont affectés à chaque fonction. Les fibres permettent le maintien de l'arbre, les vaisseaux la conduction de la sève et les rayons le stockage de ressources pendant les périodes de repos. Le bois est alors un assemblage complexe de ces différents types cellulaires (Fig.1 ; Plomion et al, 2001 ; Laurans et al, 2006 ; Déjardin et al, 2010).

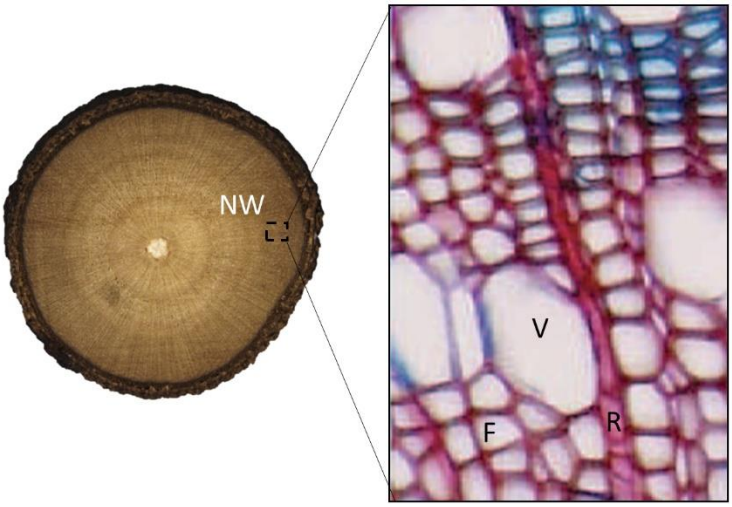

Figure 1. Coupe transversale d'une tige de peuplier (*Populus tremula* x *Populus alba*, clone INRA 717-1B4) après coloration au bleu alcian safranine.
NW : bois normal, V : vaisseau, F : fibre, R : rayon

Durant leur croissance, les arbres peuvent être sujet à des conditions environnementales récurrentes comme une forte pente ou un vent fort persistant. Ils vont alors répondre à ces conditions en modifiant la structure cellulaire de leur bois. Contrairement au bois normal qui dispose de la même structure sur toute sa circonférence, le peuplier va produire deux types de bois différents : le bois de tension et le bois opposé (Fig. 2 ; Déjardin et al, 2010). Le bois de tension, produit sur la face supérieure du tronc incliné, se caractérise par la présence d'une couche gélatineuse épaisse, dite couche G (Fig. 2 ; Onaka, 1949). Cette couche se caractérise par une composition chimique très particulière (Norberg et Meier, 1966 ; Joseleau et al, 2004 ; Pilate et al, 2004 ; Gorshkova et al, 2015 ; Guedes et al, 2017) en faisant un outil d'étude de choix dans la mise au point de méthodes de caractérisation à l'échelle cellulaire telles que l'imagerie par ATR-FTIR (Attenuated total reflectance – Fourier transformed infrared ; Cuello et al, 2020).

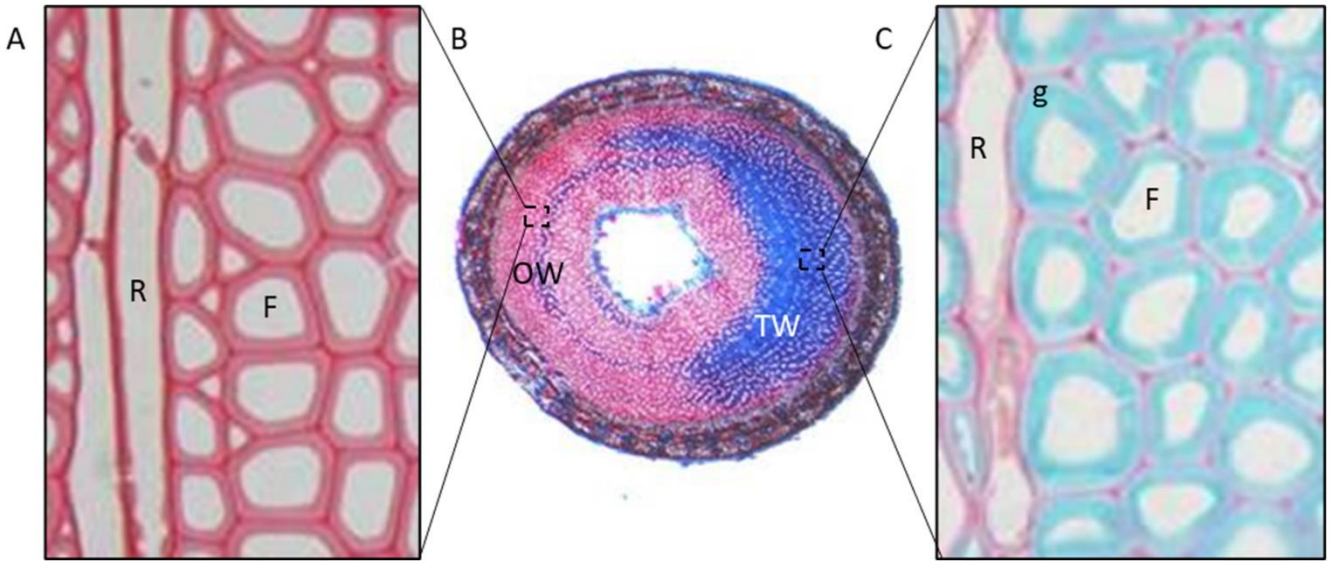

Figure 2. Coupe transversale d'une tige inclinée de peuplier (*Populus tremula* x *Populus alba*, clone INRA 717-1B4) après coloration au bleu astra safranine. (A) Bois opposé, (B) tronc, (C) Bois de tension.
OW : bois opposé, TW : bois de tension, F : fibre, R : rayon, g : couche gélatineuse

C. L'imagerie chimique en biologie végétale

L'imagerie hyperspectrale dispose d'une richesse de données très grande, ce qui explique son intérêt dans divers domaines : Imagerie via satellite au niveau de l'éco sciences, l'études des écosystème Côtiers, la Géoscience ou encore la Défense. Elle connait également une explosion dans le domaine agronomique. Ici, la notion de spectres continus, donc de bandes spectrales étroites et contiguës est essentielle et permet d'exploiter au mieux l'information. L'imagerie hyperspectrale permet, pour chaque pixel, de connaitre sa réponse spectrale afin de l'identifier (Fig. 3). Aussi, la position de leurs pics d'absorption dépend de leur composition chimique. Leur amplitude venant quant à elle donner des indications concernant la quantité des constituants présents. On peut également identifier les différents matériaux présents et en déterminer les concentrations et les caractéristiques physiques.

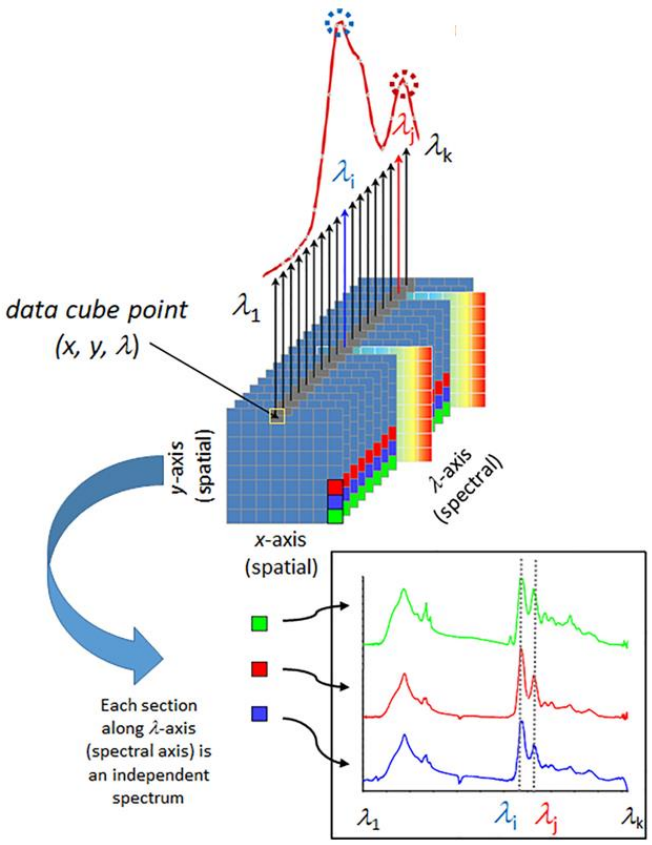

Figure 3. Exemple d'un data cube d'image hyperspectrale. Chaque pixel dispose de plusieurs valeurs d'absorbance suivant l'axe des spectres. (Beć et al, 2020)

Ces dernières années, de nombreuses techniques permettant de mesurer efficacement la biomasse végétale ont été développées. La plupart d'entre elles se basent sur l'infrarouge. L'ATR-FTIR est définie sur le principe optique de la réflectance. Ce phénomène optique bien connu se traduit par la proportion de lumière réfléchie par la surface d'un matériau. Elle se définie comme étant le rapport entre le flux lumineux réfléchi et le flux lumineux incident. Cette approche a récemment été couplé à un imageur et appliqué à des coupes transversales de bois de peuplier. Cette approche de microphénotypage non destructrice a permis, pour la première fois, de caractériser la composition chimique de l'arbre à l'échelle de la cellule. Cuello et collaborateurs (2020) ont alors développé une approche nouvelle de segmentation d'images hyperspectrales (Fig. 4), les outils classiques ne s'étant pas montré performant. Pour ce faire, ils ont réalisé une sélection manuelle de pixel appartenant à chacune des classes souhaitées en vue d'obtenir un spectre de référence représentatif de la classe (D). Les pixels sont alors comparés un à un à cette référence et assignés à une classe suivant leur coefficient de corrélation. Cette méthode s'est montrée très efficace pour les fibres classifiant

correctement plus de 90% d'entre elles. Elle présente cependant une faiblesse dans l'identification des vaisseaux et des rayons.

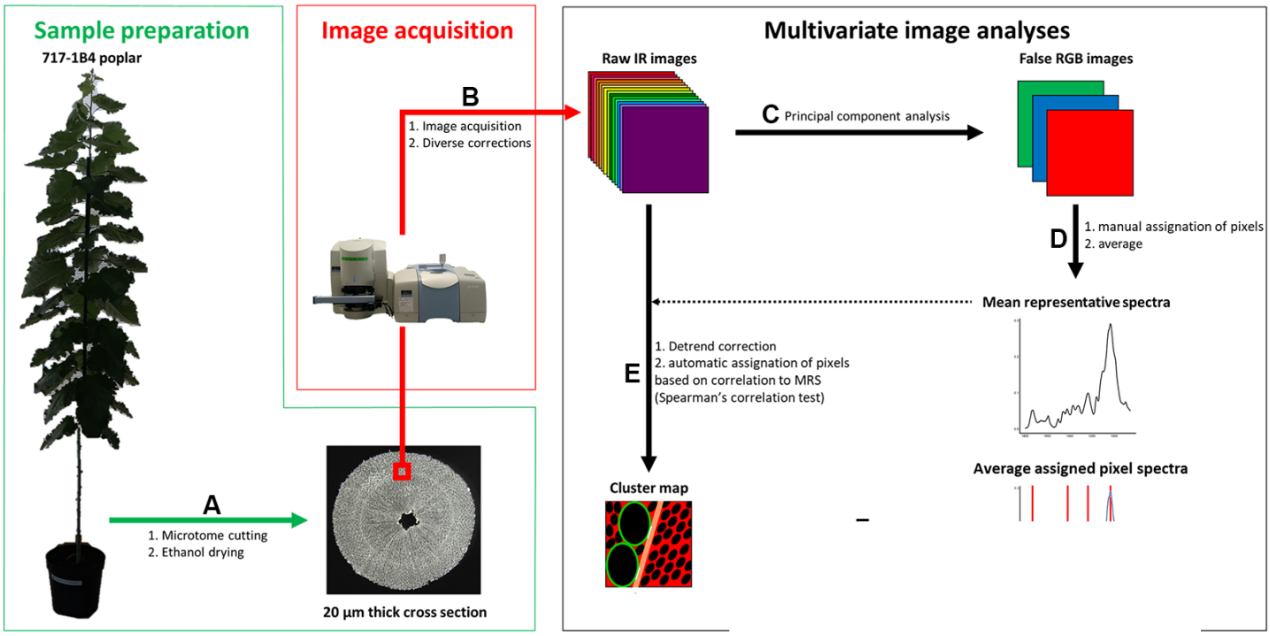

Figure 4. Préparation des échantillons : Peupliers ayant poussé trois mois dans une serre. (Cuello et al, 2020)

(A) Tranche de 20 µm d'épaisseur trempé dans l'éthanol. Des images 100x100µm² ont était acquise depuis ces coupes en utilisant un microscope infrarouge. Une réduction atmosphérique ainsi qu'une diminution du bruit ont été appliqué sur les images IR brut. (C) Analyse en composante principal a été utilisé sur les spectres bruts. Trois composantes ont été choisie pour créer une fausse image RGB. (D) 90 pixels ont été annotés manuellement pour chacun des rayon, fibre, vaisseau et couche G. (E) Une moyenne des spectres des pixels annotés a été effectué. Puis une corrélation entre cette moyenne et les spectres des pixels a été effectué en utilisant le test de corrélation de Spearman.

D. Objectifs du stage

Le bois est un tissu indispensable pour le développement et la pérennité des arbres. Comme nous l'avons vu, il est composé de plusieurs types cellulaires, chacun ayant une fonction bien particulière. Caractériser finement la mise en place des différentes structures cellulaires du bois est un enjeu majeur pour la compréhension de la réponse d'un arbre à son environnement. Dans ce cadre, des chercheurs de l'UMR BioForA tendent à caractériser les propriétés physico-chimiques des parois des différents types cellulaires du bois. Cette caractérisation passe par l'étude approfondie d'images ATR-FTIR. Ces images sont extrêmement riches. Contrairement aux habituels canaux RGB, elles contiennent l'ensemble des longueurs d'ondes du moyen infra-rouge.

L'objectif principal du stage est de poursuivre le travail déjà existant (Fig 4, Cuello et al, 2020) en testant divers outils de traitement d'image, afin de pouvoir développer un algorithme permettant de faire une assignation automatique des différentes classes composant les images hyperspcetrales.

Matériel et méthodes

A. Matériel d'étude

Pour la suite des opérations, j'avais à ma disposition douze images hyperspectrales, reparties selon les trois types de bois présentés en amont : bois droit (NW), bois opposé (OW) et bois de tension (TW). Les images sont composées de 1 626 variables représentant les nombres d'ondes entre 4 000 et 750 cm^{-1} avec un pas de 2 cm^{-1}. Les images utilisées sont tirées de l'étude de Cuello et collaborateurs (2020) et ont été nouvellement annoté pour aider à l'identification des classes (Table I).

Table 1. Nombre de pixel par classe sur les différentes images étudiées

Classe	NW1[a]	NW2[b]	NW3[c]	NW4[d]	OW1[e]	OW2[f]	OW3[g]	OW4[h]	TW1[i]	TW2[j]	TW3[k]	TW4[l]
Lum	2465	2011	1901	1919	2102	1640	2257	1355	1128	789	1066	855
FibS	985	1453	1723	1756	1221	2009	941	2373	894	1017	1583	985
Ves	264	242	229	201	495	115	413	135	330	166	206	232
Ray	382	390	243	220	278	332	485	233	234	179	282	290
FibG	NA	NA	NA	NA	NA	NA	NA	NA	1510	1945	959	1734

a-l : références des images utilisées dans Cuello et al, 2020 ; a : UT_12-D2, b : UT_12-C2, c : UT_12-D1, d : UT_12-C1, e : TT_11-A3, f : TT_11-A2, g : TT_11-A1, h : TT_12-B1, i : TT_11-A3, j : TT_11-A2, k : TT_11-A1, l : TT_12-B1, Lum : lumen, FibS : fibres, Ves : vaisseaux, Ray : rayon, FibG : couche G ; NA : non applicable

À cette fin, j'ai créé un outil permettant l'annotation d'images hyperspectrales à partir de csv contenant ces données. J'ai alors utilisé un programme déjà existant au LIFAT mais permettant de n'ouvrir que des images RVB (rouge, vert, bleu) classique. Par la suite, suivant les différentes idées d'améliorations, une mise à jour a été effectuée pour faciliter l'annotation d'images hyperspectrales : effectuer une ACP, analyse en composante principale qui consiste à transformer des variables corrélées, à savoir liées entre elles, en de nouvelles variables appelées composantes principales qui elles sont décorrélées. Cette ACP permet une meilleure visualisation des types de cellules par l'expert et donc une meilleure annotation.

Concernant les autres améliorations, nous avons une slide bar permettant le défilement entre les nombres d'ondes ainsi que certaines informations supplémentaires affichées en bas de page. Il y a également une amélioration des menus toujours dans l'optique d'aider à l'annotation. Cet outil peut également ouvrir des csv sortis d'un logiciel prédisant les classes afin d'aider à la correction des éventuelles erreurs. Cet outil a été développé en C++ avec l'aide d'opencv ainsi que de la librairie QtWidget. Il sera intégré dans un package R et mise à disposition de la communauté scientifique sur le cran (Comprehensive R Archive Network).

À l'aide de cet outil, un expert peut alors annoter les images hyperspectrales. Cette segmentation manuelle est un prérequis nécessaire à tout algorithme de classification. Ici, nous avons deux images une en RVB (Fig 5) permettant dans un premier temps de faire une confirmation visuelle avec les futures prédictions mais aussi une image ayant les classes numérotées de 1 à 4 ou 5 suivant le type de bois. Ces dernières sont bien évidemment à la même résolution à savoir 64x64 pixels.

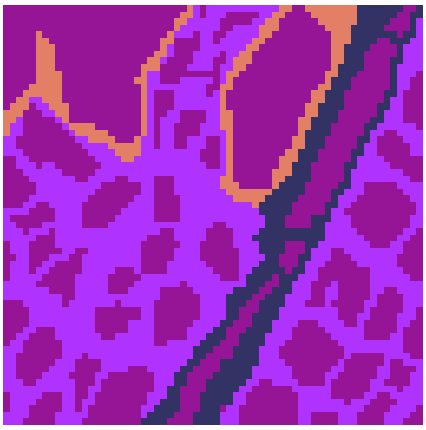

Figure 5. Exemple d'une image labélisé provenant d'un bois normal

B. Outils d'analyse d'images

Toutes les analyses d'images ont été réalisées sur python (v. 3.4) avec la librairie scikit learn (v.0.20). Le but d'utiliser une unique librairie est d'être sûr que les algorithmes utilisés sont de même qualité.

Afin de vérifier la possibilité de séparer linéairement les groupes, j'ai réalisé une LDA (linear discriminant analysis). Ceci nous permet de trouver s'il existe une combinaison linéaire entre les différentes longueurs d'ondes. Elle permet également la prédiction de classe. Un apprentissage a été effectué pour chaque type de bois indépendamment les uns des autres.

Afin d'utiliser des méthodes plus puissantes et plus sophistiquées, je me suis intéressé à deux méthodes d'apprentissage supervisée : le SVM (Support Vector

Machine) et le kNN (k-Nearest Neighbors). J'ai sélectionné ces deux algorithmes car il s'agit des plus reconnus dans le domaine du traitement d'image.

Pour commencer, le choix s'est porté sur un classique du machine learning de par sa robustesse et son efficacité à savoir le SVM (Cortes et Vapnik, 1995).

Le suppot vector machine, ou machine à vector de support, fait partie de la famille des apprentissages supervisés. Ils sont utilisés pour des problèmes de discrimination (déterminer la classe d'un échantillon), et de régression (prédire la valeur numérique d'une variable). Dans notre cas, nous utiliserons bien entendu une approche discriminante car le but est de pouvoir donner la classe à laquelle appartient un pixel issu d'une image hyperspectrale.

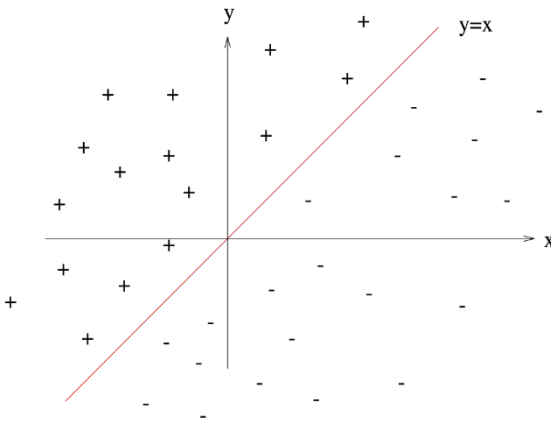

Figure 6. Représentation d'un cas simple de séparation linéaire

Le SVM est un classifieur linéaire. Dans les cas les plus simples, il va essayer de tracer une ligne (hyperplan) séparant les échantillons de classe de la manière la plus optimal possible (Fig 6). Ici, nous avons le cas le plus simple possible à savoir deux classes facilement séparables. Cependant nous pouvons aussi remarquer qu'il existe une infinité d'hyperplan possible.

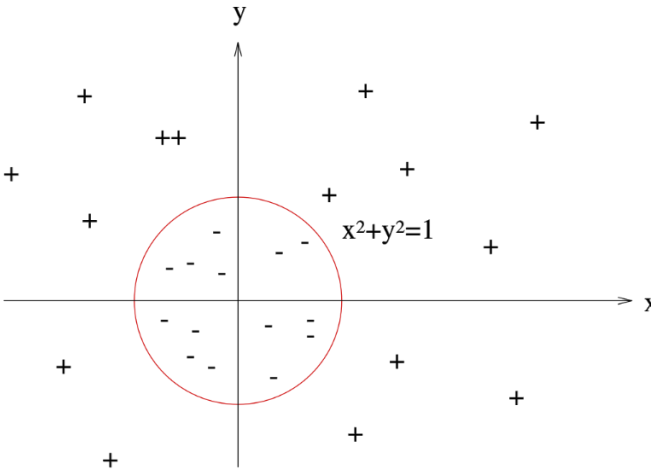

Figure 7. Représentation d'un cas simple de séparation non linéaire

À l'inverse, il est possible d'avoir également des cas au le problème n'est pas linéairement séparable (Fig 7). Ici, nous voyons visuellement que les deux classes ne peuvent en aucun cas être séparés par un seul hyperplan.

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal
laurent.chapillon@wanadoo.fr

J'ai également utilisé en comparaison un KNN (Cover et Hard, 1967), K plus proches voisins, qui est, encore une fois, un type d'apprentissage supervisé. Dans les grandes lignes, ce classifieur fonctionne en comparant la distance entre une nouvelle variable en entrée avec celle qu'il a déjà classée, ceci a pour effet de créer des clusters pour chaque classe.

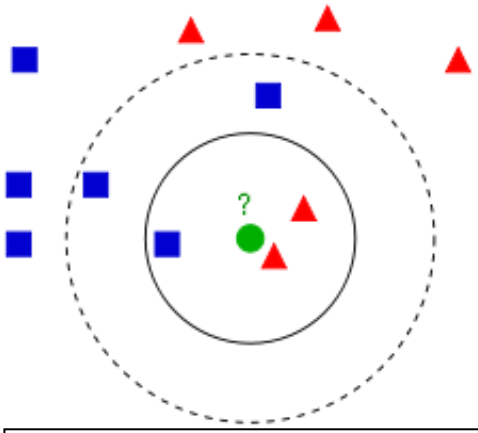

Figure 8. Représentation d'un cas de prédiction dépendamment du nombre de voisins pris en compte

On retrouve en figure 8 un exemple basique de classification à deux classes, les rectangles bleus étant la première et les triangles étant la seconde. Le cercle vert est la nouvelle entrée. Nous pouvons alors voir que le nombre de voisins pris en compte durant l'affectation de la nouvelle variable à une incidence sur le résultat. Par exemple, si seuls les trois plus proches sont choisis, alors la nouvelle variable appartient à la seconde classe. Si nous choisissons les cinq plus proches voisins, cette variable appartiendra à la première classe.

C. Ondelette de Gabor

J'ai également décidé d'utiliser une analyse de texture à savoir les ondelettes de Gabor disponible dans la librairie opencv (v. 4.4). Il s'agit d'un filtre linéaire qui va analyser l'image pour voir s'il existe des fréquences spécifiques dans l'image. Ces fréquences sont dans des directions spécifiques que nous aurons générées dans une banque de filtres. Cela va alors créer une matrice de convolution où nous pouvons faire varier l'angle d'orientation des ondelettes, ou encore leur grosseur. Cela a pour effet, sur l'image testée, de restituer les lignes de texture qui sont dans le sens opposée aux ondelettes.

Sur la Figure 9, par exemple, l'algorithme a été testé sur une voiture sur fond noir. Nous voyons alors qu'en fonction du sens ou encore de la taille des ondelettes utilisées, les contours de la voiture sont différents. Tous ces résultats peuvent ensuite être utilisés dans un classifieur pour prédire les classes.

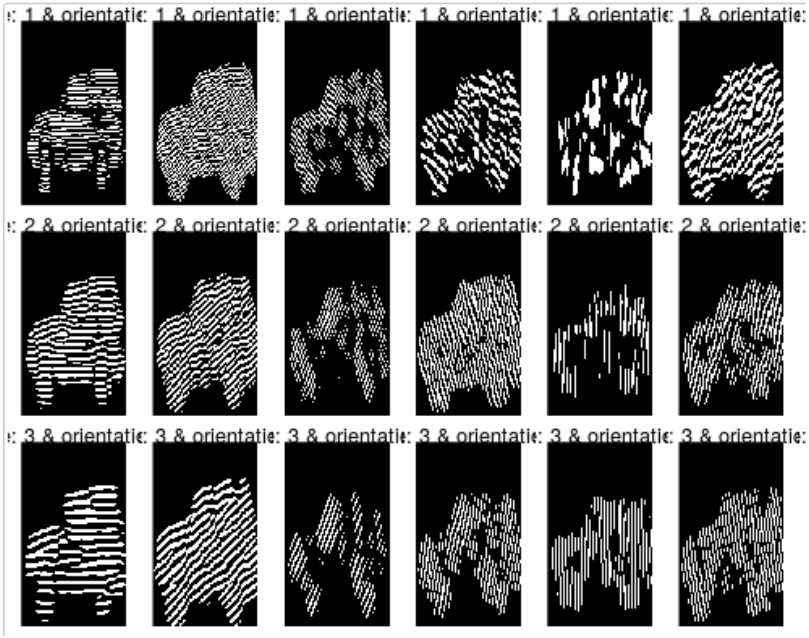

Figure 9. Exemple d'une banque d'ondelette de Gabor appliqué sur une voiture.

D. Architecture U-Net

Plus récemment encore, certaines architectures de réseaux de neurones convolutifs ont été développées. Parmi elles, on retrouve l'architecture U-Net (Fig. 10 ; Ronnerberger et al,2015). Elle est caractérisée par une structure en U composée d'un encodeur-décodeur. Le réseau est construit en deux partie, la première est un encodeur qui permet d'obtenir les variables des images, la seconde partie similaire à un décodeur permet une localisation précise. Ce type de structure ont une grande efficacité de calcul et ont la possibilité d'être entrainé avec un faible jeu de données. J'ai utilisé ce réseau car il s'agit d'un réseau de neurone convolutif, qui recherche la meilleure matrice de convolution possible. Pour cette analyse, j'ai utilisé la librairie keras (v. 2.3.1).

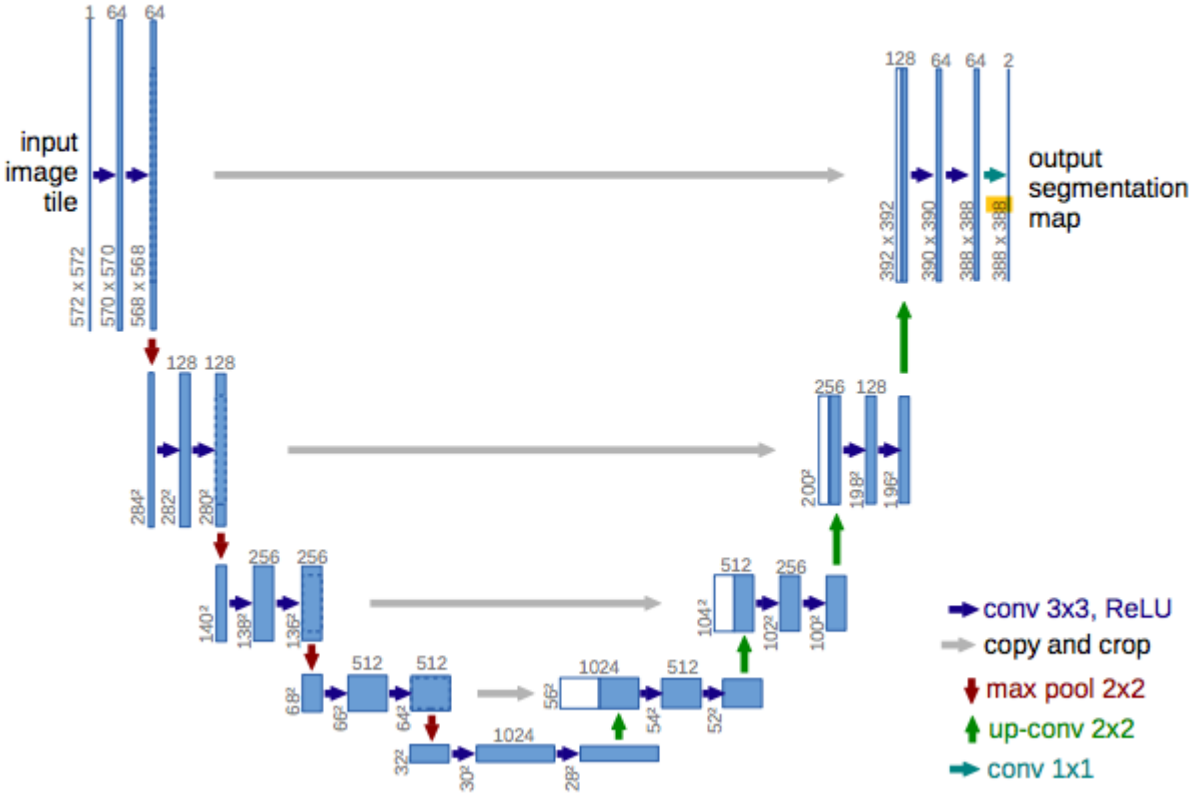

Figure 10. Architecture d'un U-Net (Ronneberger et al., 2015)

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal
laurent.chapillon@wanadoo.fr

Résultats et Discussions

A. Discrimination linéaire des classes

L'analyse par LDA m'a permis dans un premier temps d'étudier la possibilité de discriminer linéairement les différentes classes. Une représentation 3D des trois premiers axes de la LDA nous permet de distinguer les différentes classes, malgré un léger chevauchement (Fig. 11). Les données semblent donc linéairement séparables. Cette première étape est essentielle pour s'assurer de pouvoir utiliser des classifieurs linéaires.

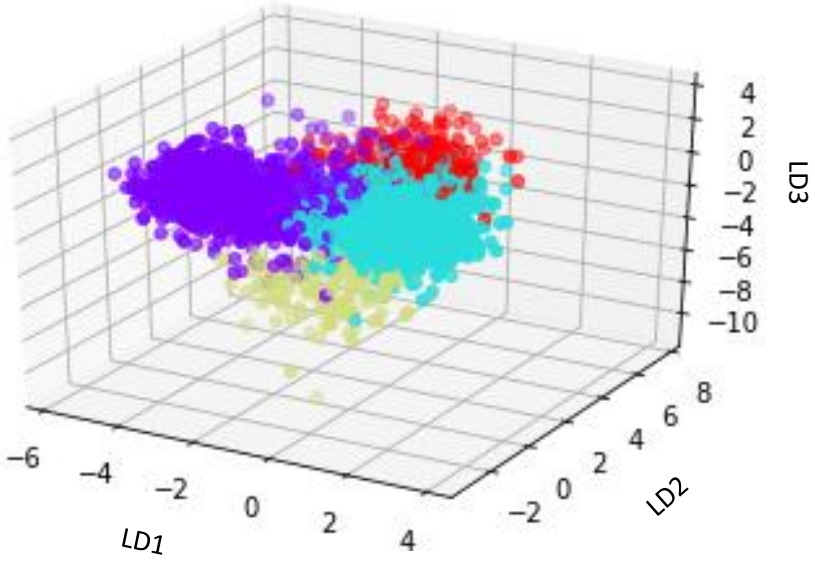

Figure 11. Représentation 3D des quatre premières composantes de la LDA sur une image de bois droit

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal
laurent.chapillon@wanadoo.fr

B. Résultat des classifieurs

Les premiers tests ont été effectués uniquement avec les données d'absorbance ainsi que les images labélisées. J'ai donc testé les différents classifieurs et ai comparé les résultats obtenus avec l'algorithme déjà existant (Fig. 12).

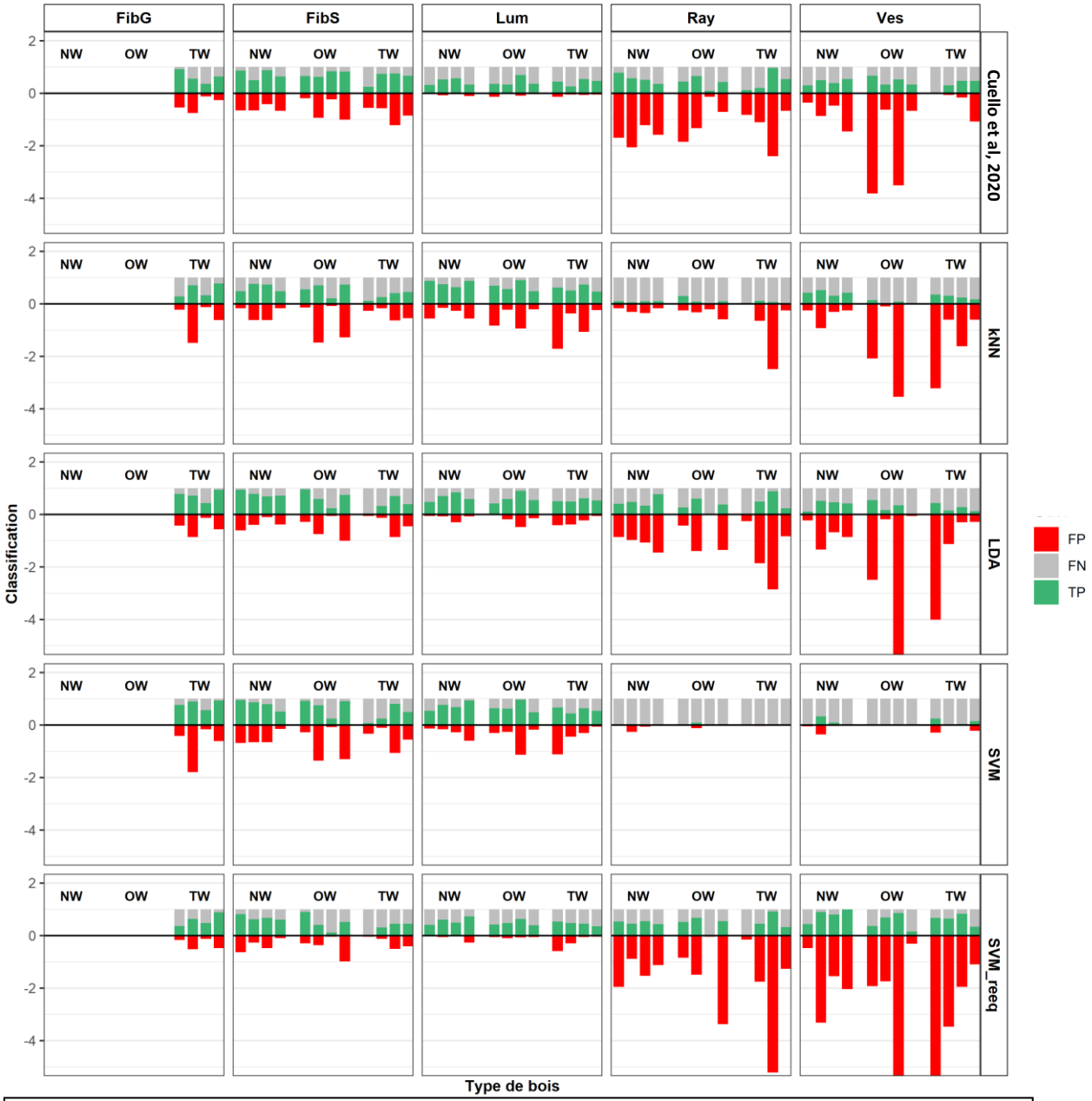

Figure 12. Proportion de faux négatif (FN), faux positif (FP) et vrai positif (TP) obtenu par chaque classifieur utilisant les 1626 nombres d'ondes en entrée.
FibG : couche G, FibS : fibres, Lum : Lumen, Ray : Rayon, Ves : vaisseaux, NW : bois normal, OW : bois opposé, TW : bois de tension, SVM_reeq : SVM avec rééquilibrage des classes. FP limité à 500%.

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal
laurent.chapillon@wanadoo.fr

On remarque dans un premier temps que, malgré ce qu'annoncent Cuello et ses collaborateurs (2020), l'efficacité de segmentation est variable suivant les images, même pour les fibres. En effet, on remarque le taux de vrai positif pour TW1, par exemple, est largement inférieur à 50% (Fig. 12). Cependant, on remarque que les classifieurs kNN, LDA et SVM ne sont pas plus efficaces. On remarque également qu'un grand nombre de faux positif est détecté quel que soit le classifieur utilisé (Fig. 12). Cela est principalement le cas pour les classes Ray et Ves. On distingue également que, pour ces deux mêmes classes, seule peu de vrais positifs sont identifiés (Fig. 12). On remarque que les pixels de lumen et fibres sont prépondérants dans le jeu de donnée (Table I). Explication possible au problème de détection des rayons et vaisseaux, j'ai alors réalisé un SVM après rééquilibrage des classes. On observe alors une amélioration dans la détection des rayons et vaisseaux avec, toutefois, une augmentation importante du nombre de faux positifs détectés dans ces mêmes classes.

En conclusion, les résultats obtenus par les classifieurs testés ne sont satisfaisant sur la seule base des données d'absorbance (Fig. 13). J'ai alors essayé d'ajouter des paramètres de texture à l'entrée de ceux-ci.

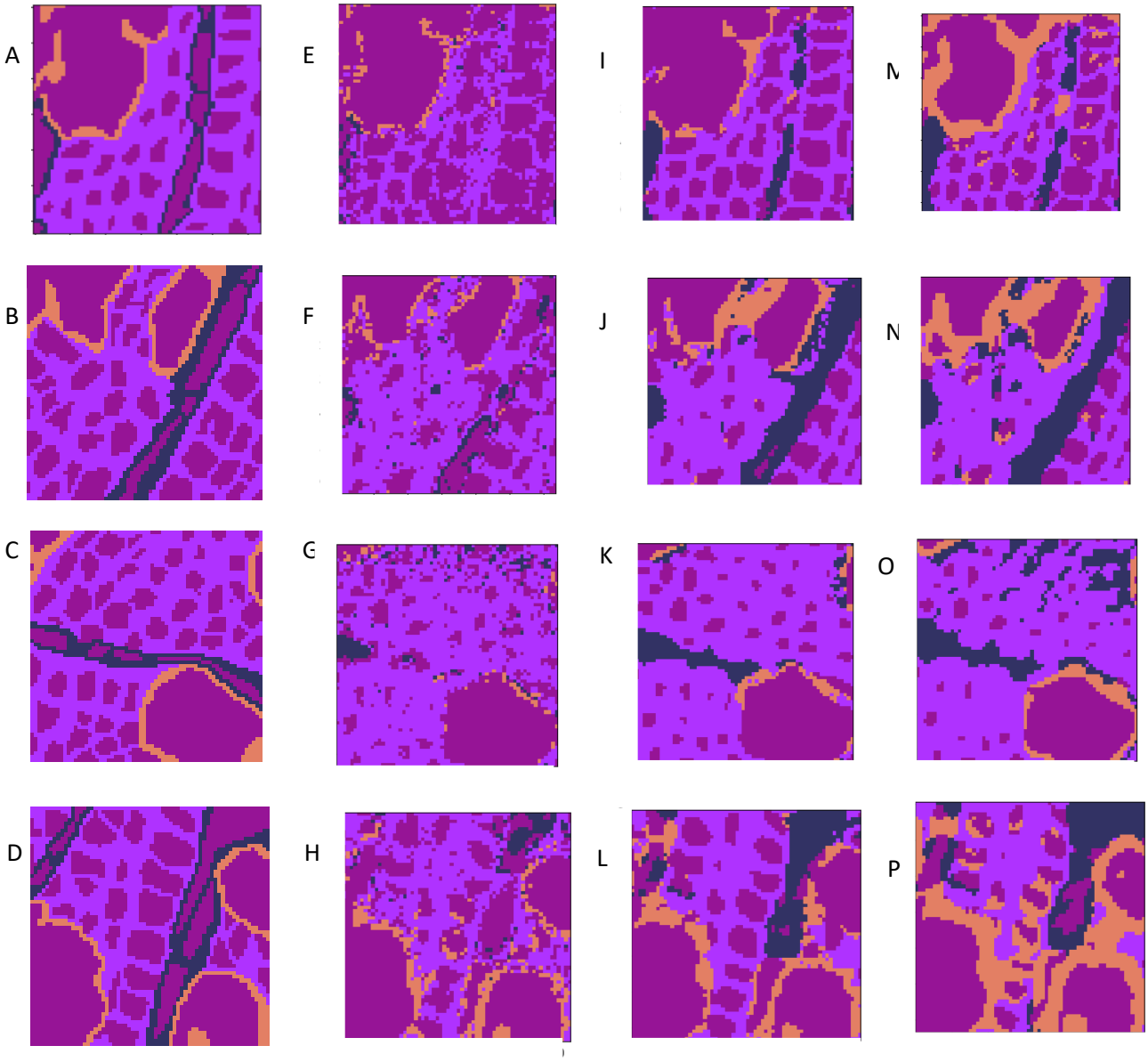

Figure13. Segmentation automatique par LDA (E-H), SVM (I-L) et K=kNN (M-P) des quatre images de bois normal (A-D)
Violet : lumen, rose : fibres, orange : vaisseaux, bleu : rayons

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal
laurent.chapillon@wanadoo.fr

C. Utilisation des ondelettes de Gabor

L'analyse de texture que j'ai réalisée est basée sur l'algorithme des ondelettes de Gabor. J'ai donc créé une banque de filtres. Le grand nombre de variables à ma disposition augmente considérablement le temps de calcul nécessaire. J'ai alors réalisé une analyse en vue d'identifier la redondance entre couche. J'ai calculé la distance entre chaque paire de variables afin d'identifier les variables à conserver. Ainsi, sur les 1626 variables d'entrée, j'en ai sélectionné 764. Par ailleurs, nous savons que l'information principale se trouve entre 1 800 et 850 cm^{-1}. Par conséquent, je n'ai retenu que 316 variables. Nous obtenons donc des images hyperspectrales de taille 64x64x316, puis chaque bande est passée une à une dans un filtre utilisant la banque d'ondelettes de Gabor.

L'utilisation des résultats obtenus par ondelettes de Gabor en entrée des outils de classification n'impacte pas ou peu la qualité de prédiction des fibres et lumen quel que soit le classifieur utilisé (Fig. 14). On remarque une fois encore un grand nombre de faux positif détecté quel que soit le classifieur utilisé (Fig. 14). Ici, sans rééquilibrage des classes, il est toujours difficile d'identifier les pixels de rayons et de vaisseaux. Le rééquilibrage des classes ou la recherche d'hyperparamètres pour la SVM semble avoir un impact important sur la capacité qu'a ce classifieur à identifier des pixels de rayons et de vaisseaux. Le nombre de faux positifs tend cependant toujours à augmenter avec ces modifications. En revanche, la combinaison du rééquilibrage de classe et de la recherche d'hyperparamètre semble limiter cette augmentation du nombre de faux positifs détecté tout en conservant le nombre de vrais positifs.

Lorsqu'on ajoute les 316 nombres d'ondes conservés aux résultats issus des ondelettes de Gabor, les résultats ont tendance à être détériorés. En effet, on remarque un grand nombre de faux positifs est détecté quel que soit le classifieur utilisé et peu à pas de vrais positifs, notamment pour les vaisseaux et les rayons (Fig. 15). Cette dégradation de prédiction peut s'expliquer par le déséquilibre entre variables issues des ondelettes de Gabor (> 6 000) et nombres d'ondes (316).

Après analyse des images segmentées, la méthode permettant la meilleure segmentation automatique semble être la SVM avec recherche d'hyperparamètres, les autres comportant énormément de bruit (Fig. 16).

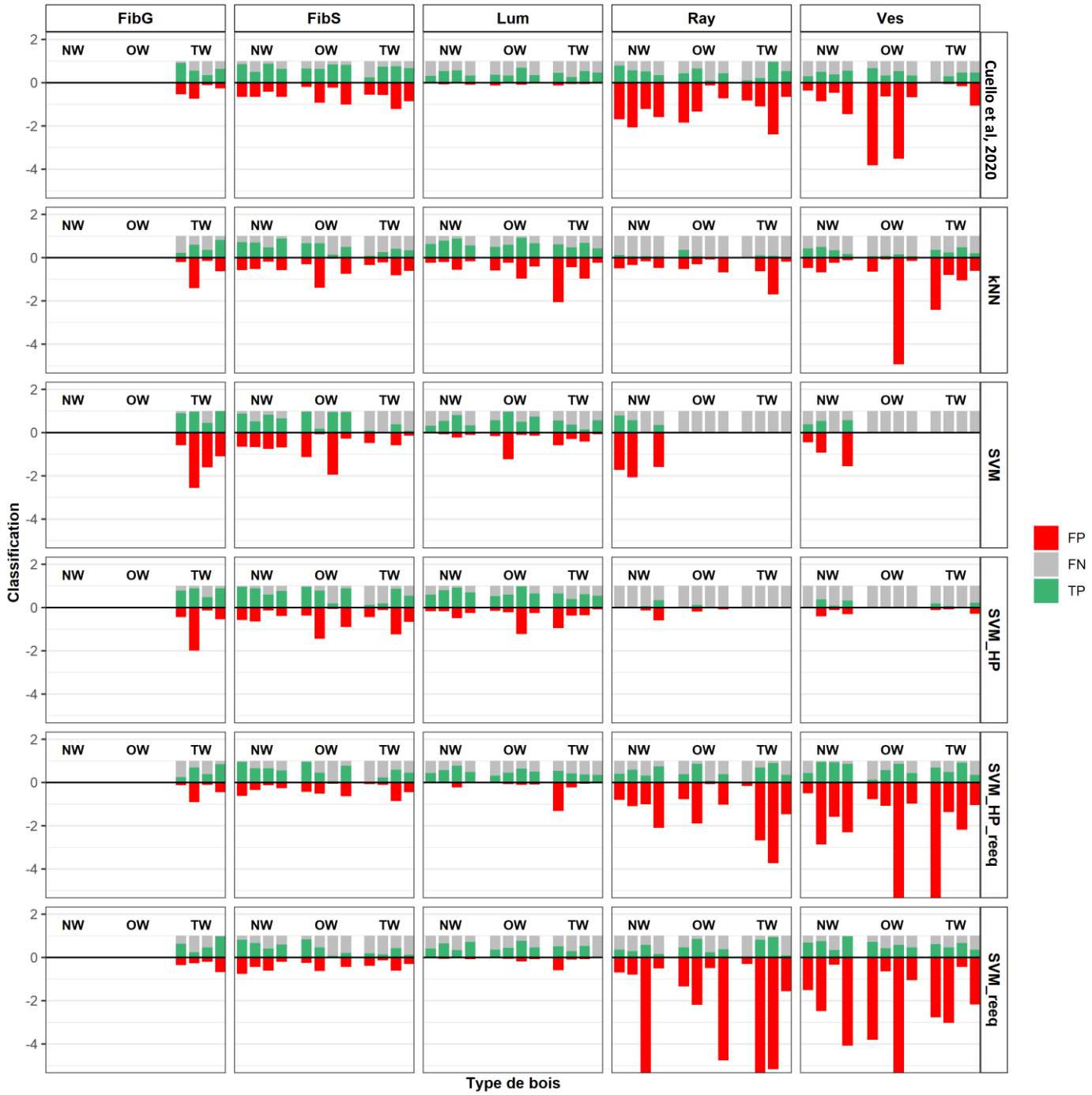

Figure14. Proportion de faux négatif (FN), faux positif (FP) et vrai positif (TP) obtenu par chaque classifieur utilisant les résultats des ondelettes de Gabor en entrée.

FibG : couche G, FibS : fibres, Lum : Lumen, Ray : Rayon, Ves : vaisseaux, NW : bois normal, OW : bois opposé, TW : bois de tension, SVM_reeq : SVM avec rééquilibrage des classes, SVM_HP : SVM avec recherche d'hyperparamètres, SVM_HP_reeq : SVM avec recherche d'hyperparamètres et rééquilibrage des classes. FP limité à 500%.

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal
laurent.chapillon@wanadoo.fr

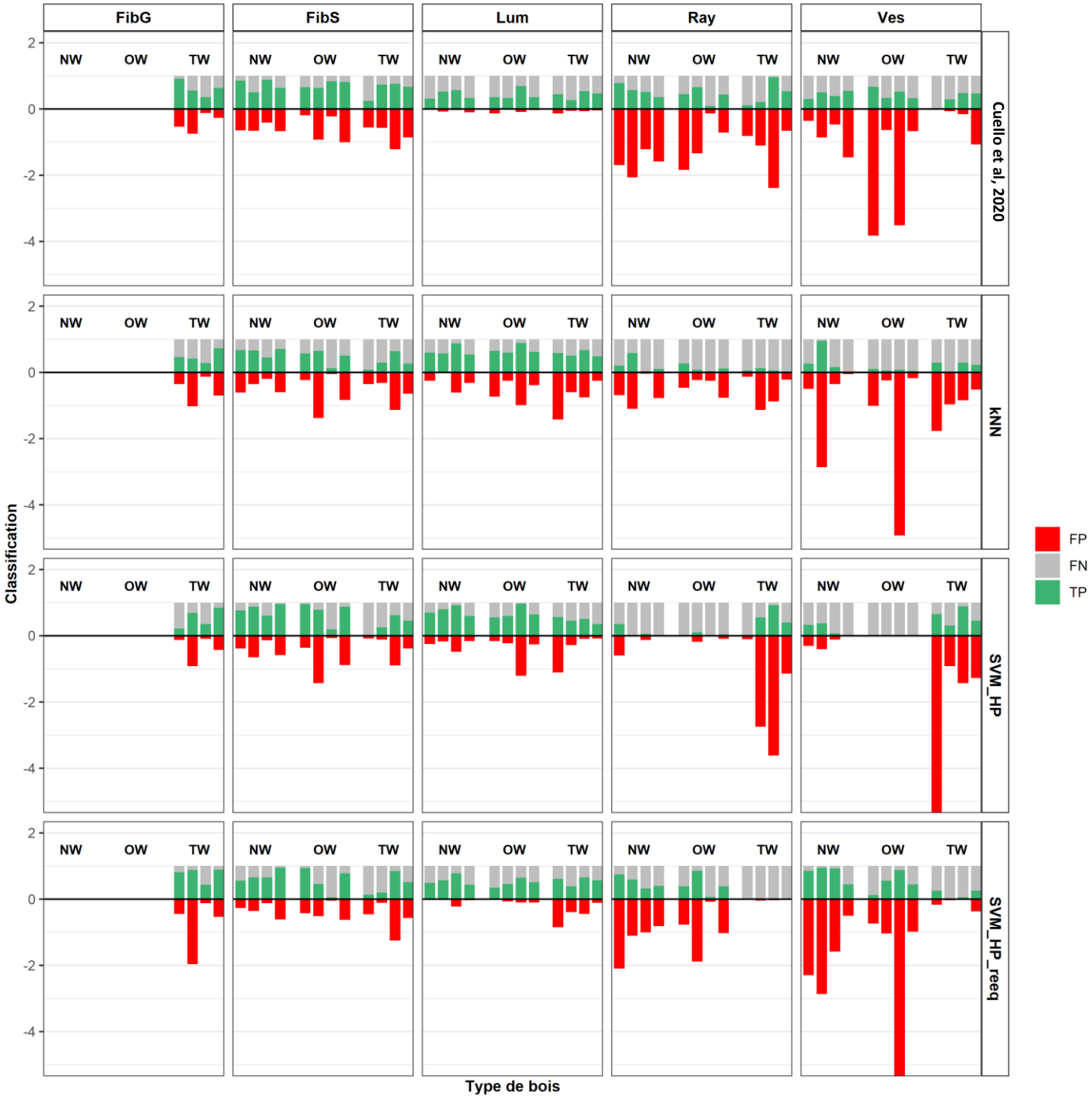

Figure15. Proportion de faux négatif (FN), faux positif (FP) et vrai positif (TP) obtenu par chaque classifieur utilisant les résultats des ondelettes de Gabor et les 316 nombres d'ondes en entrée.
FibG : couche G, FibS : fibres, Lum : Lumen, Ray : Rayon, Ves : vaisseaux, NW : bois normal, OW : bois opposé, TW : bois de tension, SVM_reeq : SVM avec rééquilibrage des classes. FP limité à 500%.

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal
laurent.chapillon@wanadoo.fr

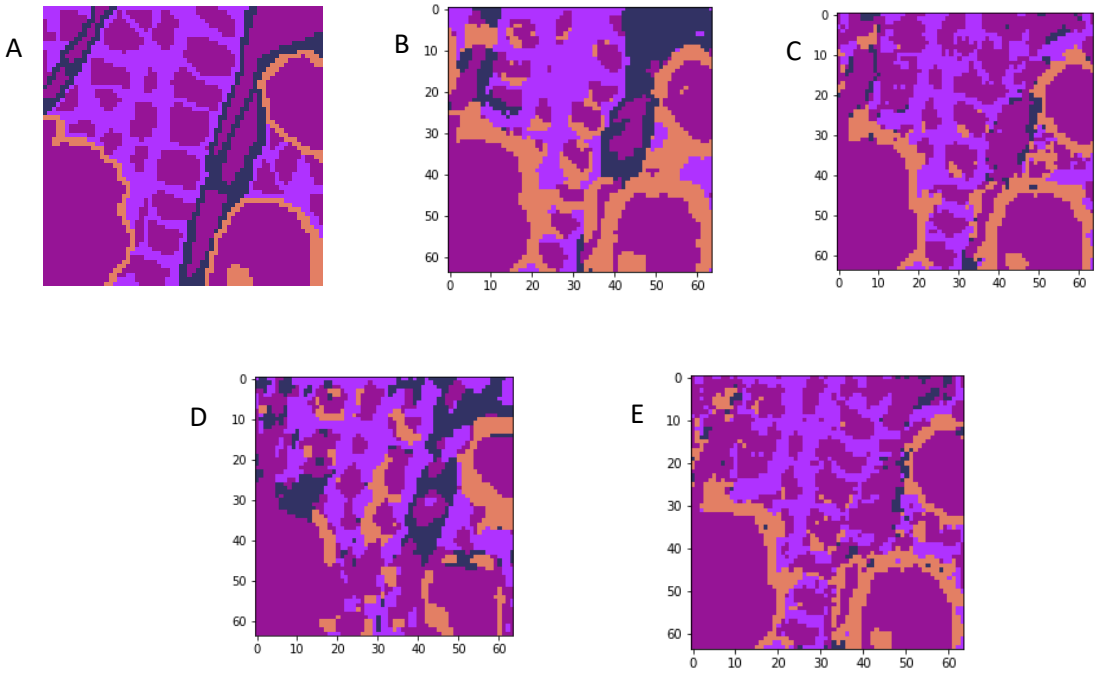

Figure 16. Segmentation automatique d'une image de bois normal (A) par SVM (B), SVM avec recherche d'hyperparamètres (C), kNN (D), tous trois utilisant les données de texture en entrée et SVM avec recherche d'hyperparamètres utilisant également les données d'absorbance (E).
Violet : lumen, rose : fibres, orange : vaisseaux, bleu : rayons

D. Résultat du U-Net

Afin d'améliorer les résultats de prédiction, j'ai utilisé une approche de deep learning en utilisant les 316 nombres d'ondes sélectionnés. Au vu de l'efficacité des outils de texture, je me suis orienté vers l'architecture U-Net car majoritairement utilisée pour la segmentation d'images dans le domaine biomédical.

Pour cette partie, je me suis focalisé sur les images de bois droit car elles semblent être les plus simples à prédire d'après mes résultats précédents. Qu'importe les images testées, ce réseau n'a été capable de distinguer que deux classes : les lumens d'une part et une classe regroupant les fibres, les rayons et les vaisseaux (Fig. 17). Ce problème peut être dû au déséquilibrage entre classes et au peu de données d'entrainement.

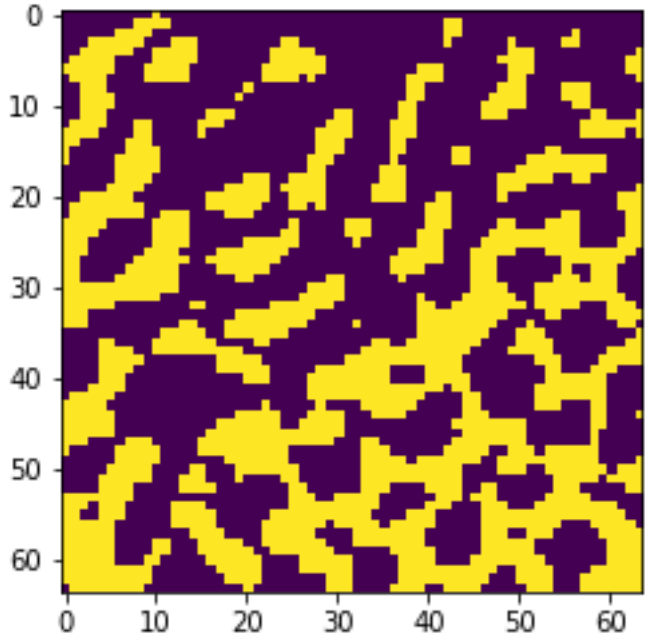

Figure 17. Prédiction fait par le U-Net pour la deuxième classe.

Pour obtenir de meilleurs résultats, nous pouvons augmenter la base de données et/ou effectué des prétraitements spectraux en amont du U-Net. Ces pistes n'ont pas pu être explorées dans le cadre du stage.

Conclusions et perspectives

L'objectif du stage était d'améliorer une méthode d'identification de types cellulaires provenant de plusieurs types de bois. Les différents tests menés ont montré que les approches d'analyse de texture donnent des résultats prometteurs. Cependant, il est encore difficile de dépasser les 75% de précision. Cette faiblesse peut s'expliquer par le manque d'images hyperspectrales en entrainement. Augmenter la base de données en la triplant pourrait avoir un impact positif sur la prédiction de l'ensemble des classes, notamment les vaisseaux et les rayons. Avec le jeu de donnée disponible, les meilleurs résultats sont obtenus soit par la méthode développée par Cuello et collaborateurs (2020) soit par SVM avec recherche d'hyperparamètres et rééquilibrage des classes. Par ailleurs, la prédiction par U-Net n'a pas pu être totalement finalisée. Malgré sa capacité à être entrainé sur de faibles jeux de données, il est probable que celui à ma disposition soit encore trop petit. Ainsi, augmenter la base de données me semble nécessaire avant de poursuivre l'exploration de cette méthode. Par ailleurs, j'ai également remarqué que l'efficacité de classification était fortement dépendante de l'ordre d'entrée des images d'apprentissage. En effet, le nombre d'images disponibles étant faible et le nombre de pixel par classe variant d'une image à l'autre, les classifieurs peinent à établir un modèle suffisamment robuste.

Étant donnée la variabilité importante au niveau du nombre de pixel contenant l'information des vaisseaux et des rayons, les méthodes de classification supervisée seules se sont rapidement révélées insuffisantes. Récemment, Corcel (2017) a développé une approche de classification non supervisée d'images multispectrales basée sur du K-means multi-échelle. Au vu des résultats prometteurs qu'ils exposent, il serait intéressant d'adapter cette approche aux images hyperspectrales.

Ce stage a tout de même permis de développer un outil d'annotation permettant à la communauté scientifique d'annoter simplement et efficacement de manière manuelle les images hyperspectrales. Cet apport non négligeable offre la possibilité d'agrandir facilement la base de données d'entrainement, le seul facteur limitant étant à présent l'acquisition des images.

Laurent Chapillon
Master 2 Traitement de l'Image et du Signal
laurent.chapillon@wanadoo.fr

Références

Beć K.B., Grabska J., Bonn G.K., Popp M., Huck C. W. (2020). Principles and Applications of Vibrational Spectroscopic Imaging in Plant Science: A Review. Front. Plant Sci. 11:1226. doi: 10.3389/fpls.2020.01226

Corcel M. Imagerie multispectrale en macrofluorescence en vue de la prédiction de l'origine tissulaire de particules de tiges de maïs [Thèse]. Nantes (France) : Université Bretagne Loire, 2017

Cortes C., Vapnik V. (1995). Support-Vector Networks. Machine Learning, 20, 273-297.

Covert T. M. et Hard P.E. Nearest Neighbor Pattern Classification. IEE transactions on information theory, Vol IT-I3, N°. 1.

Cuello, C., Marchand, P., Laurans, F., Grand-Perret, C., Laine-Prade, V., Pilate, G., et Déjardin A. (2020). ATR-FTIR microspectroscopy brings a novel insight into the study of cell wall chemistry at the cellular level. Front. Plant Sci. 11:105:105. doi: 10.3389/fpls.2020.00105

Déjardin, A., Laurans, F., Arnaud, D., Breton, C., Pilate, G., and Leplé, J.-C. (2010). Wood formation in Angiosperms. Comptes Rendus Biologies 333, 325–334. doi:10.1016/j.crvi.2010.01.010

Gorshkova, T., Mokshina, N., Chernova, T., Ibragimova, N., Salnikov, V., Mikshina, P., et al. (2015). Aspen tension wood fibers contain β-(1→4)-galactans and acidic arabinogalactans retained by cellulose microfibrils in gelatinous walls. Plant Physiology, pp.00690.2015. doi:10.1104/pp.15.00690

Guedes, F. T. P., Laurans, F., Quemener, B., Assor, C., Lainé-Prade, V., Boizot, N., et al. (2017). Non-cellulosic polysaccharide distribution during G-layer formation in poplar tension wood fibers: abundance of rhamnogalacturonan I and arabinogalactan proteins but no evidence of xyloglucan. Planta 246, 857–878. doi:10.1007/s00425-017-2737-1

Joseleau, J.-P., Imai, T., Kuroda, K., and Ruel, K. (2004). Detection in situ and characterization of lignin in the G -layer of tension wood fibres of Populus deltoides. Planta 219, 338–345. doi:10.1007/s00425-004-1226-5

Laurans, F., Déjardin, A., Leplé, J., and Pilate, G. (2006). Physiologie de la formation des parois de fibres de bois. Revue des composites et des matériaux avancés 16, 25–40. doi:10.3166/rcma.16.25-40

Norberg, P. H., and Meier, H. (1966). Physical and Chemical Properties of the Gelatinous Layer in Tension Wood Fibres of Aspen (Populus tremula L.). Holzforschung 20, 174–178. doi:10.1515/hfsg.1966.20.6.174

Onaka F. (1949). Studies on compression-and tension-wood. Mokuzai Kenkyu 1, 1–88

Pilate, G., Chabbert, B., Cathala, B., Yoshinaga, A., Leplé, J.-C., Laurans, F., et al. (2004). Lignification and tension wood. Comptes Rendus Biologies 327, 889–901. doi:10.1016/j.crvi.2004.07.006

Plomion, C., Leprovost, G., and Stokes, A. (2001). Wood formation in trees. Plant Physiol. 127, 1513–1523

Ronnerberge O., Fischer P., Brox T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Proceedings of the Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science, 9351 (2015), pp. 234-241, 10.1007/978-3-319-24574-4_28