Aide au diagnostic du cancer de la prostate par IRM multi-paramétrique : une approche par classification supervisée

Émilie Niaf

Numéro d'ordre : 271-2012 Année 2012

THÈSE

Délivrée par

L'UNIVERSITÉ CLAUDE BERNARD LYON 1
Spécialité : Traitement du signal et de l'image

DIPLÔME DE DOCTORAT

(arrêté du 7 août 2006)

ÉCOLE DOCTORALE : Interdisciplinaire Sciences-Santé (EDISS)

Emilie NIAF

Aide au diagnostic du cancer de la prostate par IRM multi-paramétrique : une approche par classification supervisée

Jury

Isabelle BLOCH	Professeur, Télécom ParisTech, LTCI, Paris	Rapporteur
Michèle ROMBAUT	Professeur, Université Joseph Fourier, Gipsa-Lab, Grenoble	Rapporteur
Olivier BASSET	Professeur, Université Claude Bernard, CREATIS, Lyon	Examinateur
Stéphane CANU	Professeur, INSA de Rouen, LITIS, Rouen	Examinateur
Carole LARTIZIEN	Chargée de recherche au CNRS, CREATIS, Lyon	Co-directrice de thèse
Olivier ROUVIÈRE	Professeur-Praticien Hospitalier, LabTAU, Lyon	Co-directeur de thèse

Résumé

Le cancer de la prostate est le cancer le plus fréquent et la deuxième cause de mortalité par cancer chez l'homme, en France. Actuellement, la méthode de référence pour le diagnostic du cancer de la prostate reste les biopsies écho-guidées. Le cancer de la prostate étant mal visible en échographie, les biopsies écho-guidées sont en réalité réalisées "au hasard" et distribuées de façon systématique dans la glande (biopsies dites *randomisées*). Ces biopsies *randomisées* n'étant qu'un échantillonnage, elles peuvent passer à côté de certains foyers tumoraux et apprécient mal le volume et l'agressivité des cancers détectés. L'imagerie par résonance magnétique multiparamétrique (IRM-mp) se positionne actuellement comme la technique la plus prometteuse pour la prise en charge de ce cancer. En permettant une visualisation précise et complète de la glande, elle pourrait, à terme, permettre un diagnostic plus précoce (biopsies ciblées plutôt que *randomisées*) et rendre enfin possible le traitement focal des cibles malignes, alternative à la prostatectomie radicale. Cependant, l'IRM-mp reste très difficile à interpréter, notamment lorsque les résultats des différentes séquences sont contradictoires et on observe une forte variabilité inter-lecteur. D'où l'intérêt porté au développement de systèmes experts pour "accompagner" les radiologues dans leur tâche de diagnostic.

Dans une première partie, nous présentons un système d'aide au diagnostic (*Computer-Aided Diagnosis*, CAD en anglais) reposant sur un ensemble de caractéristiques descriptives extraites de trois séquences IRM (imagerie T2-w, dynamique et de diffusion) et intégrant de l'information statistique, structurelle et fonctionnelle. Appuyant notre étude sur une base de données cliniques de 30 patients pour lesquels nous disposons d'une référence histologique, nous comparons plusieurs stratégies de classification combinant quatre méthodes de sélection de caractéristiques à quatre algorithmes de classification supervisée. Le schéma optimal, basé sur un séparateur à vaste marge (SVM) couplé à un test t, est ensuite évalué en pratique clinique auprès de douze radiologues, de 6 mois à 7 ans d'expérience, dont on teste les performances diagnostiques sans et avec l'outil CAD utilisé en "second avis". Les résultats obtenus montrent, de manière systématique, une amélioration des performances de discrimination et de la confiance dans l'interprétation, même pour les radiologues les plus expérimentés.

L'analyse des limites du système proposé et des schémas classiques de systèmes supervisés en général souligne l'influence majeure de la base d'apprentissage sur les performances. Celle-ci doit en effet être à la fois riche et fiable. L'établissement d'une telle base de données, où les lésions malignes sont annotées, sur les images IRM-mp, de manière précise et exhaustive, est une tâche difficile et fastidieuse. De nombreuses études ne peuvent pas s'appuyer sur une corrélation anatomo-radiologique et la vérité terrain utilisée est souvent le résultat d'une analyse en aveugle réalisée par un radiologue expert, entachée d'incertitude. Or, si les SVM sont des algorithmes de discrimination efficaces, ils ne peuvent cependant pas être appliqués directement dans le cas où les données contiennent des étiquettes de classe certaines (sain/pathologique) et des étiquettes incertaines correspondant à un score de suspicion de malignité. Dans la seconde partie de ce travail, nous proposons une nouvelle formulation des SVM permettant d'intégrer l'incertitude de l'expert sur la classe de certains exemples, communiquée sous la forme d'une probabilité d'appartenance. L'idée est d'apprendre une fonction discriminante qui, à la fois, maximise les performances de classification sur les étiquettes certaines et prédit au mieux les probabilités sur les étiquettes incertaines. L'évaluation de ce nouvel algorithme, noté P-SVM (pour *Probabilistic* SVM), est d'abord réalisée sur des exemples jouets puis sur la série de données cliniques. Nous montrons que le P-SVM permet de pondérer l'influence des cas d'interprétation difficile et d'obtenir de meilleures performances que celles réalisées en utilisant une vérité expert binaire.

Table des matières

Liste des abréviations

AC agent de contraste
Acc Précision
ACP analyse en composantes principales
ADC coefficient apparent de diffusion
AIF fonction d'entrée artérielle
ADL analyse discriminante linéaire
ANN réseaux de neurones artificiels
AUC aire sous la courbe ROC
AUGC aire sous la courbe de réhaussement en produit de contraste
CAD système d'aide au diagnosis
CADe système d'aide à la détection
CADx système d'aide à la décision
CLARA-P corrélations anatomo-radiologiques en IRM de prostate
CNB classifieur naïf de Bayes
CRF champs aléatoires conditionels
CT tomodensitométrie
DCE *dynamic contrast enhanced*
DCT tranformée en cosinus discret
DP densité de protons
DSC indice de Dice
DWI imagerie pondérée en diffusion
EEE espace extravasculaire/extracellulaire
FMRF champs de Markov flous
FN faux négatifs
FP faux positifs
FOV champ de vue
Gd-DOTA gadotérate méglumine
GE *graph Embedding*
GLCM matrice de co-ocurrence des niveaux de gris
Hz Hertz
HES Hématoxyline-Eosine-Safran
HIFU ultrasons focalisés de haute intensité
IM information mutuelle
IRM imagerie par résonance magnétique
IRM-mp IRM multi-paramétrique
Kep constante de flux entre l'EEE et le compartiment vasculaire
k-PPV k-plus proches voisins
Ktrans constante de transfert entre le compartiment vasculaire et l'EEE
LLE *locally Linear Embedding*
LOO *leave-one-out*
LOPO *leave-one-patient-out*

M malin
MCE Mean Cross-Entropy
mg milligramme
min minute
mm millimètre
MRF champs de Markov flou
MRS spectroscopie par résonance magnétique
mRMR *minimum-Redundancy, Maximum-Relevancy*
ms milliseconde
MSE Mean Squared Error
N normal
Nb nombre
NS non malin mais suspect
PIN néoplasie intraépithéliale de la prostate
PSA antigènes spécifiques à la prostate
P-SVM SVM probabiliste
QP problème quadratique
RBF fonction à base radiale
RDF forêt aléatoire de décision
RF radiofréquence
RMN résonance magnétique nucléaire
ROC *receiver operating characteristic*
ROI région d'intérêt
RVM *relevance vector machine*
s seconde
SE sensibilité
SFMA stroma fibro-musculaire
SI signal
SNR rapport signal sur bruit
SP spécificité
SVM séparateur à vaste marge
T Tesla
TE temps d'écho
TEMP tomographie d'émission monophotonique
TEP tomographie par émission de positron
TR temps de répétition
T2-map cartographie T2
T2-w T2-pondérée
T1-w T1-pondérée
US ultrason
Ve volume de l'EEE
VN vrais négatifs
VOI volume d'intérêt
VP vrais positifs
VPN valeur prédictive négative
VPP valeur prédictive positive
WI *wash-in*
WO *wash-out*
ZC zone centrale
ZP zone périphérique
ZT zone transitionnelle
2-D 2 dimensions
3-D 3 dimensions

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

Détecter et localiser les foyers tumoraux dans la prostate : un besoin clinique important

Le cancer de prostate est le cancer le plus fréquent et la deuxième cause de mortalité par cancer chez l'homme dans les pays développés [2, 8, 44, 52]. Il s'agit donc d'un réel problème de santé publique. Alors que l'imagerie s'est concentrée pendant longtemps sur le bilan d'extension locale du cancer (franchissement capsulaire, envahissement ganglionnaire), il apparaît maintenant qu'il est au moins aussi important de disposer d'informations précises sur la position exacte des foyers tumoraux dans la glande et sur leur agressivité. Malheureusement, aucune méthode d'imagerie ne donne actuellement cette information. L'échographie notamment est peu sensible et peu spécifique et les biopsies écho-guidées sont en réalité des biopsies réalisées au hasard dans la glande, l'échographie servant surtout à s'assurer de leur espacement régulier (on parle d'ailleurs de "biopsies randomisées"). Ce diagnostic "au hasard" du cancer pose de nombreux problèmes :

- il peut être source de délais diagnostiques, surtout si la tumeur est dans un territoire peu accessible aux biopsies (tumeurs antérieures) ;
- de très petits foyers de cancer peu agressifs peuvent être détectés par hasard, risquant de conduire à des traitements inutilement lourds et mutilants ;
- les patients ayant une seule biopsie randomisée positive sur moins d'un millimètre peuvent se voir proposer une simple surveillance, pour ne pas risquer de sur-traiter un foyer millimétrique ; cependant, certains de ces patients ont en fait un foyer tumoral plus volumineux et/ou agressif que ne le laisse penser le résultat des biopsies

et mériteraient un traitement immédiat. Par ailleurs, la surveillance se fait sur des paramètres (évolution du taux des antigènes spécifiques à la prostate (PSA), nouvelles séries de biopsies randomisées) qui ne sont pas optimaux pour attester d'une croissance tumorale ;
- il est impossible de réaliser un traitement focal du cancer (détruire seulement les foyers tumoraux en évitant les complications de l'ablation complète de la prostate) alors que l'on dispose de traitements (par ultrasons focalisés notamment) qui s'y prêteraient ;
- les biopsies ne sont pas des actes anodins. Invasives et agressives, les biopsies peuvent être notamment sources d'infections ou d'hémorragies.

Localiser précisément les foyers tumoraux permettrait 1) d'assurer un diagnostic plus précis et précoce, en guidant les biopsies vers les foyers suspects, 2) de diminuer le risque de sur-diagnostic en dirigeant les biopsies sur des cibles macroscopiques, 3) de réserver la surveillance à ceux ayant réellement de petits cancers peu agressifs et de surveiller ces cancers sur le seul paramètre carcinologique qui ait du sens : le temps de doublement tumoral et 4) de développer enfin des traitements focaux rationnels.

IRM de prostate : des progrès récents

Actuellement, la méthode qui permet le mieux de localiser le cancer dans la prostate est certainement l'IRM. Son utilisation pour le diagnostic et la localisation du cancer est aujourd'hui très discutée au sein de la communauté médicale [1]. Alors que la classique imagerie en T2-w était limitée, d'autres techniques ont récemment démontré leur potentiel dans la caractérisation des zones suspectes dans la prostate : IRM dynamique, IRM de diffusion, spectroscopie [42, 61, 98, 102, 139]. D'autres techniques, comme l'élastographie IRM, pourraient s'ajouter dans l'avenir. Toutes ces séquences, considérées individuellement, souffrent cependant d'une sensibilité et/ou d'une spécificité de détection insuffisantes. Des études récentes ont montré que ces techniques se complétaient les unes les autres et que l'analyse conjointe de plusieurs de ces modalités permettait au radiologue d'améliorer ses performances diagnostiques (mesurées via les courbes ROC) [22, 42, 61, 123, 140]. C'est pourquoi le concept d'IRM multi-paramétrique (IRM-mp) s'est récemment développé : il s'agit d'associer, dans un même examen IRM, des séquences T2-w, dynamique, de diffusion voire de spectroscopie. Il subsiste cependant le problème, majeur, de savoir comment croiser ces informations et notamment de savoir quelle attitude pratique adopter quand les résultats de ces différentes techniques sont contradictoires chez le même patient.

Problématique de la thèse

Le développement actuel de l'IRM multi-paramétrique impose au radiologue d'analyser et fusionner à l'œil un gros volume de données. C'est une tâche complexe et fastidieuse, en particulier pour des radiologues peu expérimentés. On observe d'ailleurs une forte inter-

et intra-variabilité dans l'analyse des images IRM [34, 38, 103, 115] puisqu'il n'existe pas de directives fiables pour l'interprétation des différentes séquences.

La fusion d'images IRM-mp est encore peu abordée dans la littérature. Certains services d'imagerie clinique proposent la visualisation concomitante des différentes séquences sur une interface conviviale permettant une interprétation plus aisée. Seules quelques études se sont intéressées aux méthodes de fusion et d'analyse automatique de plusieurs séquences (T2, dynamique, diffusion par exemple) à des fins de diagnostics assistés par ordinateur [5, 59, 63, 83]. La plupart des travaux récents ont proposé des modèles prédictifs combinant des informations extraites d'une [65, 67, 93, 124, 133] ou deux [18, 132] séquences seulement.

Nous proposons de construire et d'évaluer une chaîne de traitement des images IRM pour l'aide au diagnostic du cancer de la prostate. Ce travail a été réalisé à l'initiative du Pr. Olivier Rouvière, directeur de cette thèse, radiologue et chef du service de radiologie urinaire et vasculaire de l'hôpital Edouard Herriot (Lyon, France). Olivier Rouvière est un spécialiste de l'imagerie urologique et un promoteur de l'imagerie IRM multi-paramétrique pour le cancer de la prostate. Il est le porteur d'un projet INCa (Institut National du Cancer) sur ce sujet. Il a également initié le développement de la banque de corrélations anatomo-radiologique CLARA-P (pour *corrélations anatomo-radiologiques en IRM de prostate*), dont le projet a été déposé en 2008. La construction effective de la base CLARA-P a débuté en 2009 et a été enrichie tout au long de ce travail de thèse grâce à la collaboration de deux radiologues et deux anatomo-pathologistes.

Ayant à notre disposition une base de données cliniques, notre travail s'oriente vers des méthodes de type classification supervisée. Leur principe consiste à élaborer un modèle de prédiction à partir d'une base d'apprentissage de données annotées. Ce modèle empirique permet ensuite de classer une nouvelle image test en fonction de ses caractéristiques.

Le développement d'une méthode de classification de données suppose d'effectuer différents choix méthodologiques. Il faut d'abord choisir les caractéristiques (ou *features*) décrivant les données (par exemple, des paramètres statistiques, structurels ou fonctionnels extraits des images IRM-mp) à utiliser pour la classification. Différentes stratégies sont possibles pour réduire ce nombre de caractéristiques de manière à ne conserver que celles qui sont les plus discriminantes. Il faut enfin définir un classifieur. Différentes approches ont été proposées et testées dans la littérature. Elles ont toutefois été évaluées sur des populations différentes, ce qui rend délicate toute comparaison de leurs performances. En nous appuyant sur la base CLARA-P, nous proposons de comparer les performances de différentes stratégies, combinant et optimisant différentes approches de sélection de caractéristiques et de classification, afin d'identifier le schéma optimal dont les performances seront évaluées en pratique clinique.

Un tel système peut être utilisé de deux manières : 1) le système peut se comporter comme une aide à la décision (CADx) en proposant un "deuxième" avis au lecteur sur une

cible qui lui paraît suspecte ; 2) le système peut extraire de manière automatique à partir des images IRM de prostate une carte de suspicion de présence de lésions tumorales et ainsi proposer une aide à la détection (CADe). C'est l'orientation CADx, définie par le Pr. Rouvière comme besoin clinique prioritaire, qui a été privilégiée dans cette thèse.

L'originalité de notre approche est de combiner trois modalités d'imagerie IRM permettant d'extraire un grand nombre de caractéristiques des différents types d'images parmi lesquelles nous proposons d'isoler les plus pertinentes pour la discrimination des tissus sains des tissus malins. La force de ce travail de thèse est de s'ancrer sur une base de données cliniques (en constante évolution), qui recueille, pour chaque patient traité par prostatectomie radicale, la description de toutes les zones suspectes sur toutes les séquences IRM, le degré de suspicion affecté par le radiologue et le résultat de l'analyse histologique de la zone correspondante sur la pièce de prostatectomie. Enfin, l'apport de notre système de classification sur le diagnostic est mesuré en pratique clinique auprès de douze radiologues d'expérience variable.

Si notre étude peut s'appuyer sur une base de données d'apprentissage fiable, issue de la mise en correspondance des données histologiques et radiologiques, de nombreuses équipes n'ont pas cette opportunité. En effet, l'établissement d'une base de données d'apprentissage riche, où les lésions malignes sont annotées, sur les images IRM-mp, de manière précise et exhaustive, est une tâche difficile et fastidieuse, nécessitant la mobilisation de radiologues et d'anatomo-pathologistes. De ce fait, la vérité terrain utilisée est souvent le résultat d'une analyse en *aveugle* réalisée par un radiologue expert, entachée d'incertitude. Dans cette thèse, nous proposons une nouvelle méthode de classification généralisant le schéma optimal précédemment obtenu en permettant de traiter le cas de ces données pour lesquelles les étiquettes de classes (sain/pathologique) s'apparentent à des probabilités de malignité.

Plan de la thèse

La première partie de ce manuscrit présente le contexte médical et scientifique de la thèse. Nous décrivons les différents aspects du cancer de la prostate, le principe de l'imagerie par résonance magnétique et les séquences utilisées dans cette thèse. Enfin, nous présentons le principe et les étapes majeures de la construction des systèmes d'aide au diagnostic et dressons un état de l'art des schémas CAD pour le cancer de la prostate en imagerie IRM proposés dans la littérature.

La seconde partie est consacrée à notre première contribution : nous détaillons, comparons et évaluons différentes chaînes de traitement des images IRM-mp pour l'aide au diagnostic du cancer de la prostate. Nous exposons les choix méthodologiques que nous avons faits pour répondre à notre problématique clinique. Nous présentons ensuite la base de données cliniques sur laquelle s'appuie notre étude et décrivons chaque étape des traitements proposés : extraction et sélection des caractéristiques image et choix des classifieurs

utilisés. Enfin, nous présentons la stratégie d'évaluation retenue et les résultats obtenus. Cette première étude nous permet de proposer un système efficace de classification permettant de discriminer les zones cancéreuses des zones suspectes mais bénignes de l'image. Nous finalisons enfin l'évaluation des performances de ce système en pratique clinique, par l'intermédiaire d'une étude ROC réalisée auprès de douze radiologues.

La troisième partie présente notre deuxième contribution : une extension du séparateur à vaste marge (SVM) au cas des étiquettes probabilistes. Après avoir explicité les motivations pratiques de ce développement théorique, nous rappelons les fondements de la formulation classique du SVM. Nous introduisons ensuite la nouvelle formulation que nous proposons puis détaillons les étapes de sa résolution. Après avoir exposé les performances de cette méthode sur des exemples simulés afin de mieux en appréhender l'utilité, nous proposons une évaluation clinique sur la base de données d'images IRM multi-paramétrique (IRM-mp) et comparons les performances diagnostiques de ce nouveau système de classification avec celles du système CADx proposé dans la partie 2.

Finalement, nous discutons, dans un dernier chapitre, du bilan et des perspectives de ce travail de thèse.

I Contexte médical et scientifique de la thèse

Chapitre 2

Le cancer de la prostate

2.1 Introduction

Ce premier chapitre est dédié à la description de la prostate, son anatomie et sa fonction au sein de l'appareil reproducteur masculin, et du cancer de la prostate, de son diagnostic à sa prise en charge.

2.2 Epidémiologie

Le cancer de la prostate est de loin le cancer le plus fréquent chez l'homme après 50 ans, avec 71 500 nouveaux cas diagnostiqués en France en 2010. Il est le deuxième en termes de mortalité chez l'homme, avec 8 790 décès estimés en France en 2010 [2,8,52] et la quatrième cause de décès par cancer tous sexes confondus. Le taux de mortalité standardisé sur la population mondiale était de 11,2 pour 100 000 hommes et le taux d'incidence (standardisé monde) de 128,8/100 000.

C'est le cancer dont l'incidence a le plus augmenté ces 25 dernières années, loin devant les cancers du poumon et du côlon-rectum, sans doute en bonne partie du fait du dépistage individuel par dosage du PSA (lire section 2.5) mais aussi du vieillissement de la population.

Facteurs de risque du cancer de la prostate

L'âge est en effet le principal facteur de risque du cancer de la prostate : 69% des cancers de la prostate surviennent après 65 ans, avec un âge moyen de diagnostic en 2005

de 71 ans. D'autres facteurs augmenteraient la probabilité de survenue : les antécédents infectieux, les antécédents familiaux précoces (moins de 55 ans) dans les parents au premier ou deuxième degré et l'origine géographique des populations (l'Afrique subsaharienne et les Antilles ont des prévalences supérieures à la moyenne au contraire des pays d'Asie du Sud-Est pour lesquels l'incidence de ce cancer est faible). Des facteurs environnementaux sont fortement soupçonnés : de nombreuses études sur des facteurs alimentaires supposés promoteurs du cancer, les polluants carcinogènes, le tabac, le stress sont en cours.

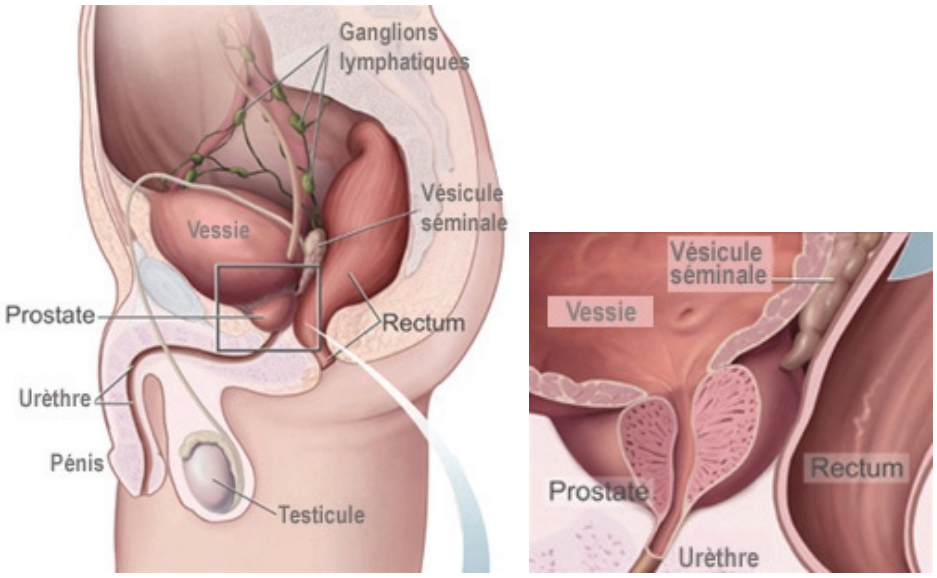
2.3 Anatomie et fonctions de la prostate

La prostate est une glande de l'appareil génital masculin. Elle est située dans le bassin, sous la vessie en avant du rectum et entoure le début de l'urètre, ce canal permettant d'éliminer l'urine de la vessie (voir figure 2.1). Une prostate saine a la forme d'une châtaigne d'environ 3 centimètres de hauteur et 4 centimètres de large, elle ne pèse pas plus de 20 grammes à l'âge adulte et est entourée d'une capsule. La prostate est formée de plusieurs lobes : un lobe prostatique antérieur, deux lobes latéraux et un lobe médian, aussi appelé lobe de Home. Elle se divise en 3 zones (voir figure 2.2) :

- **une zone périphérique (ZP)** : c'est la région de la prostate la plus proche du rectum. Elle constitue la plus grande zone de la prostate.
- **une zone centrale (ZC)** : c'est la partie de la prostate située à la base entourant les canaux éjaculateurs.
- **une zone transitionnelle (ou de transition, ZT)** : c'est la zone située au milieu de la prostate en avant des zones périphérique et centrale. Elle entoure l'urètre, canal qui traverse la prostate et représente environ 5% de la prostate jusqu'à l'âge de 40 ans. Avec le vieillissement, cette zone augmente en taille pour devenir la plus grosse partie de la prostate. C'est ce qu'on appelle un adénome de la prostate (également appelé hypertrophie bénigne de la prostate) qui survient chez presque tous les hommes de plus de 70 ans. L'augmentation de taille de la zone de transition a pour effet de pousser la zone périphérique vers le rectum.

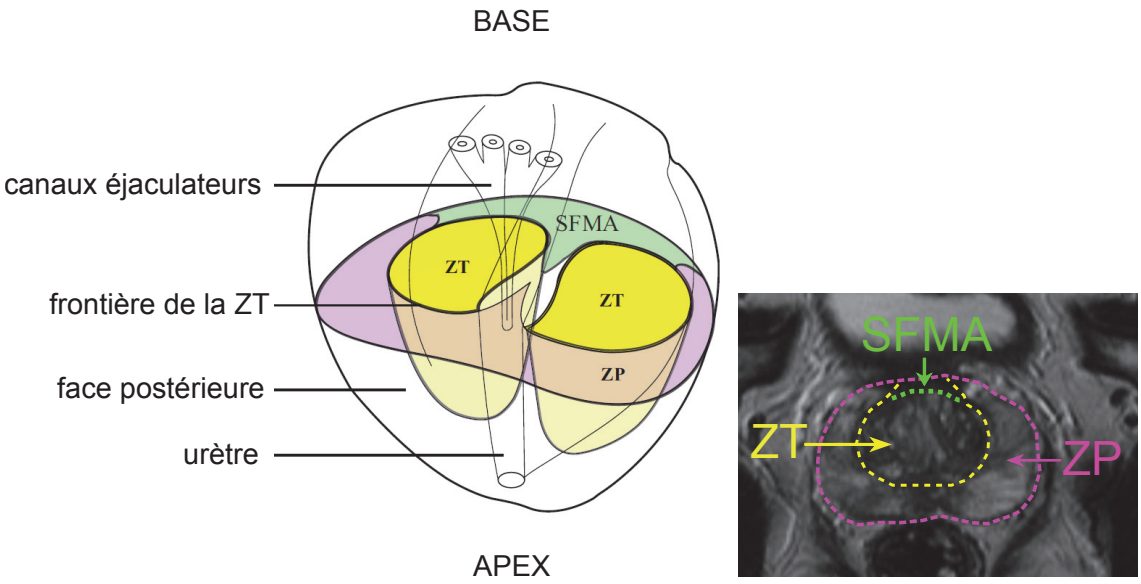
Tout autour de l'urètre, un ensemble de fibres musculaires (SFMA) regroupées sous la prostate forme le sphincter urinaire qui contrôle le passage de l'urine en se contractant ou se relâchant, permettant ainsi la continence. La zone centrale (ZC) est souvent regroupée avec la zone périphérique (ZP) dans le "compartiment externe" tandis que zone transitionnelle (ZT) et SFMA forment le "compartiment interne". Par abus de langage, on assimile la ZC à la ZP dans la pratique courante.

C'est à l'intérieur de la prostate que se fait la jonction entre l'urètre venant de la vessie, les canaux déférents et les vésicules séminales. La prostate sécrète 10-30% du liquide séminal, le reste est produit par les vésicules séminales. Le liquide séminal se mélange aux spermatozoïdes, qui viennent des testicules par les canaux déférents, dans l'urètre prostatique au moment de l'éjaculation.



(a) Pelvis masculin **(b)** Prostate et structures adjacentes

Figure 2.1 – *Représentation schématique du pelvis masculin, de la prostate et des diverses structures anatomiques adjacentes. Source : US government agency National Cancer Institute.*



(a) Anatomie zonale de la prostate **(b)** Coupe axiale en imagerie IRM T2-w

Figure 2.2 – *Anatomie zonale de la prostate. (a) Représentation schématique [69](b) acquisition T2-w d'une section axiale moyenne de la prostate permettant de localiser la zone périphérique, la zone transitionnelle et le stroma fibro-musculaire.*

2.4 Caractéristiques du cancer de la prostate

Les cancers prostatiques sont localisés dans la ZP dans 75% des cas, dans la ZT dans 25% des cas.

Stadification du cancer

Comme pour la plupart des cancers, on utilise la classification "TNM" (*Tumor, Node, Metastasis*) pour préciser le stade (ou extension) du cancer de la prostate (voir figure 2.3). La tumeur est ainsi caractérisée par 4 stades :

- **T1** : tumeur non palpable lors du toucher rectal ;
- **T2** : tumeur palpable, limitée à la prostate ;
- **T3** : tumeur étendue en dehors de la prostate ;
- **T4** : tumeur atteignant les organes voisins de la prostate (vessie, rectum, urètre).

Les cancers limités à la prostate, sans atteinte ganglionnaire ou métastatique sont des cancers localisés pouvant bénéficier d'un traitement à visée curative. Les cancers étendus en dehors de la prostate vont nécessiter des stratégies associant souvent plusieurs traitements pour retarder l'évolution du cancer.

Le score de Gleason

Le cancer est souvent hétérogène, composé de cellules d'agressivité différente. Lors de l'examen des tissus au microscope (obtenus par biopsies ou après prostatectomie), les cellules cancéreuses sont séparées en 5 classes qui correspondent au degré de différenciation et donc d'agressivité du cancer : les cellules cancéreuses classées "grade 1" ressemblent aux cellules normales de la prostate (cellules bien différenciées), celles classées "grade 5" sont très éloignées des cellules normales de la prostate (elles sont peu différenciées)(voir figure 2.4). Plusieurs grades peuvent se rencontrer au sein d'un même tissu. Le score de Gleason additionne les deux grades les plus représentés dans la région suspectée et varie donc de 2 (1+1) à 10 (5+5). Le grade majoritaire est le premier terme de l'addition : une tumeur de score de Gleason 7(3+4) a un contingent de grade 4 minoritaire, à l'inverse d'une tumeur de score de Gleason 7 (4+3). Le score de Gleason est un marqueur de l'agressivité tumorale. On considère qu'un cancer est agressif lorsqu'il supérieur à 7.

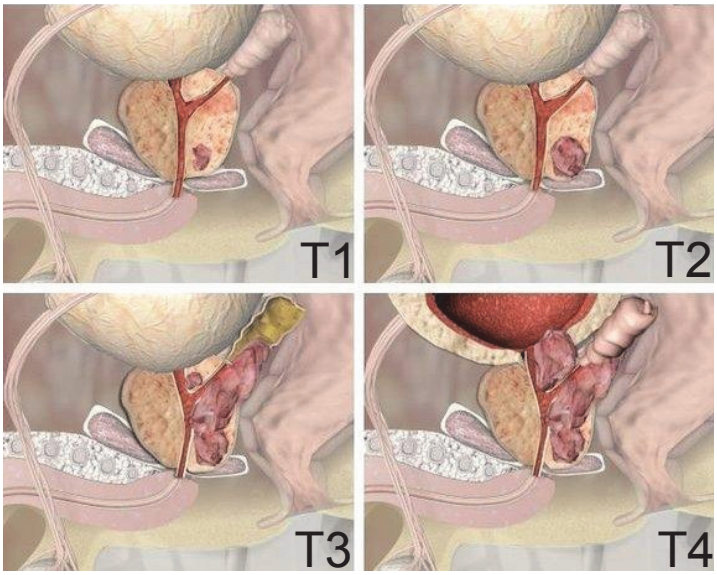


Figure 2.3 – *Différents stades du cancer de la prostate. T1 : seul un taux de PSA élevé peut trahir l'existence de la tumeur. T2 : un toucher rectal ou une échographie permet de la détecter. T3 : le tissu avoisinant ou les vésicules séminales sont touché(es). T4 : les organes proches sont touchés. Source : medipedia*

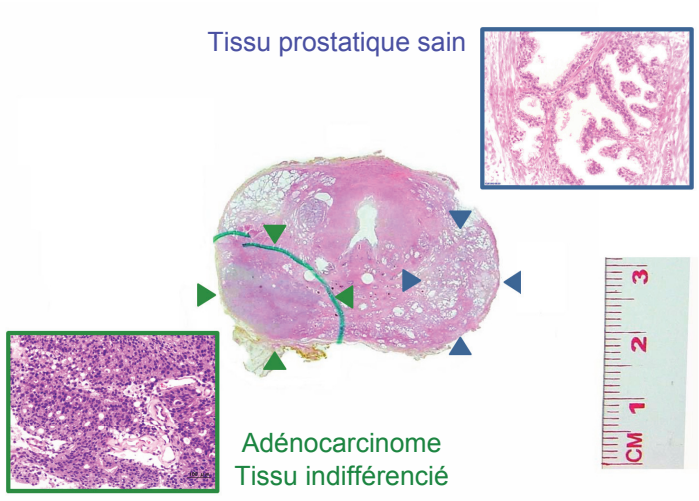


Figure 2.4 – *Histopathologie du cancer de la prostate. Coupe histologique médiane d'une prostate présentant un foyer malin (mis en évidence par les flèches vertes) de score de Gleason 9 (4+5). Contrairement au tissu prostatique sain (flèches bleues), le tissu malin est caractérisé par une absence de différenciation des cellules.*

2.5 Diagnostic du cancer de la prostate dans la pratique clinique

Le diagnostic du cancer de prostate est généralement envisagé devant une anomalie du toucher rectal (TR) et/ou une élévation du taux de PSA. Des biopsies *randomisées* sont alors réalisées.

Seul le cancer de la zone périphérique (ZP) peut être détecté par le **toucher rectal** devant une déformation des contours de la glande qui crée un bombement (voir figure 2.3). Le manque de sensibilité du TR et la sous-stadification qu'il implique le rendent peu fiable pour apprécier l'extension locale de la tumeur.

Le terme **PSA**, abréviation de *Prostatic Specific Antigen*, est une molécule qui n'est fabriquée que par la prostate. Son taux dans le sang varie en fonction de l'âge avec une normale inférieure à 4 ng/ml entre 60 et 70 ans. Le cancer est le plus souvent suspecté devant une élevation de ce taux de PSA. Cependant, le PSA est une molécule spécifique de la prostate mais pas du cancer, le taux de PSA peut ainsi être élevé dans d'autres situations (infection de la prostate - prostatite -, adénome, hypertrophie bénigne, etc). A noter que les deux grandes études prospectives européennes publiées récemment [4,110] montrent qu'un dépistage de masse du cancer de la prostate par le dosage du PSA conduit à un risque de sur-diagnostic important, et à l'absence d'effet bénéfique sur la survie spécifique ou globale.

La biopsie de la prostate (voir figure 2.5) est un examen qui consiste à prélever plusieurs fragments de la prostate. Elle est indiquée en cas de suspicion de cancer de la prostate, suite à un toucher rectal ou au résultat du PSA. Seule la biopsie (et l'examen au microscope des fragments prélevés) permet d'affirmer le diagnostic de cancer de la prostate. La prostate est le seul organe pour lequel on réalise classiquement des biopsies à l'aveugle, sans cible. Les recommandations officielles concernant le dépistage individuel du cancer prostatique reposent sur la réalisation systématique de 12 biopsies dès que le dosage sérique du taux de PSA dépasse 4ng/ml [117]. Celles-ci sont guidées par échographie et espacées régulièrement sur la surface de la glande. On parle de biopsies "pseudo-aléatoires" ou *randomisées* car elles ne réalisent évidemment qu'un échantillonnage de la glande. Ces biopsies ne représentant qu'un échantillonnage, elles peuvent rater certains foyers tumoraux et apprécient mal le volume et l'agressivité (score de Gleason) des cancers détectés. Le risque de sur et sous-diagnostic est en ainsi élevé. On estime que 17 à 21 % des hommes ayant eu une première série de biopsies négatives ont des re-biopsies positives [74,112].

En cas de biopsies positives, le taux de PSA, le pourcentage de biopsies positives et leur score de Gleason sont les éléments qui déterminent la gravité de l'état du patient [25,54] et donc le traitement à effectuer. Les biopsies *randomisées* ne permettant pas d'évaluer

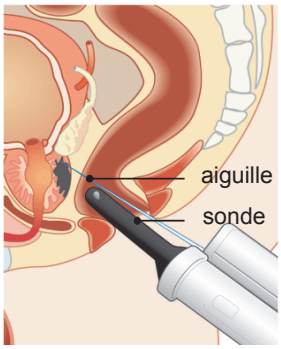


Figure 2.5 – *Diagnostic du cancer de la prostate. Réalisation de biopsies de prostate par voie endorectale et écho-guidée. Source : UroFrance*

précisément le volume et le nombre de foyers tumoraux, le traitement peut être inadapté.

Après une première série de biopsies négatives, la réalisation d'une nouvelle série de biopsies peut être proposée. Il n'y a pas de protocole validé de surveillance ou d'indication systématique de biopsies en cas de persistance simple d'une élévation du taux de PSA suspecte. L'attitude adoptée est donc très dépendante du praticien.
A noter que la multiplication et la répétition des biopsies au hasard dans la glande finissent souvent par ramener par "chance" un micro foyer tumoral de faible Gleason, permettant d'affirmer la présence d'un cancer qui n'aurait finalement peut être pas dû être diagnostiqué [126]...

2.6 Bilan d'extension

Lorsque la présence du cancer est avérée, il est également nécessaire de savoir si celui-ci est localisé à la glande prostatique ou s'il en a déjà franchi les limites, car les traitements proposés seront différents. Il est alors nécessaire de faire appel à des techniques d'imagerie. Dans le cas des cancers à risque élevé (Gleason ≥ 8, stade T3 ou PSA > 20 ng/ml), les recommandations officielles préconisent de réaliser une imagerie **IRM** en T2-pondéré pour le bilan d'extension locale (franchissement capsulaire) et un **scanner** ou **scintigraphie osseuse** pour détecter une éventuelle dissémination du cancer dans d'autres organes (ganglions lymphatiques, foie, os, etc). La réalisation de ce bilan d'extension permet alors de classer la tumeur du patient suivant la classification TNM.
Néanmoins, du fait du manque de fiabilité (faibles sensibilité et spécificité) de l'imagerie et de son coût, beaucoup d'équipes d'urologie choisissent de ne pas réaliser ces examens d'imagerie et évaluent le pronostic en se basant uniquement sur les tables de Partin [87] qui, à partir de paramètres cliniques (Gleason, PSA etc), fournissent une mesure de probabilité d'extension de la tumeur...

2.7 Traitement du cancer de la prostate

Les formes peu agressives de cancer, localisées dans la prostate seule, sont guéries dans 80% des cas. Lorsque la présence du cancer est révélée, différents types de traitements, curatifs ou palliatifs, peuvent être mis en œuvre. Nous les présentons ci-dessous.
A noter que le traitement de référence pour les cancers cliniquement localisés (pas de franchissement capsulaire détectés) est la prostatectomie radicale ou la radiothérapie, qui sont associés dans le cas des formes évoluées.

La prostatectomie. C'est un traitement local du cancer qui consiste à enlever la prostate dans sa totalité, ainsi que les vésicules séminales. C'est le traitement de référence du cancer de la prostate localisé. Malheureusement, ses effets secondaires inéluctables sur la continence urinaire et la fonction érectile malgré la préservation sphinctérienne et des bandelettes neuro-vasculaires, ont un impact significatif sur la qualité de vie des patients traités.

La radiothérapie externe. C'est un traitement local du cancer qui a pour but de détruire les cellules cancéreuses localisées au niveau de la prostate au moyen de rayonnements ionisants. Les photons de très haute énergie sont produits par des accélérateurs linéaires de particules qui constituent une source d'irradiation externe avec focalisation des faisceaux sur la prostate à irradier. La radiothérapie externe est un traitement de référence du cancer de la prostate localisé. C'est également un traitement de référence, en association avec l'hormonothérapie, du cancer de la prostate à haut risque et localement avancé. Elle s'accompagne également d'un risque d'impuissance et éventuellement d'incontinence urinaire, ainsi que d'un risque d'inflammation rectale (rectite radique).

La curiethérapie. C'est un traitement très localisé du cancer qui consiste à mettre en place des implants radioactifs à l'intérieur de la prostate. Ces implants émettent des rayonnements qui détruisent les cellules cancéreuses de la prostate. La curiethérapie est une modalité thérapeutique possible pour certains cancers de la prostate localisés à faible risque.

L'hormonothérapie. C'est le traitement de référence du cancer de la prostate avec atteinte ganglionnaire pelvienne et du cancer de la prostate métastatique. Elle vise à neutraliser l'hormone masculine (testostérone) dont dépend notamment la progression de la tumeur. Son effet n'est malheureusement que transitoire.

La chimiothérapie. Elle est indiquée pour le traitement des cancers métastatiques hormonorésistants dans le but de soulager la douleur ou maîtriser les symptômes de la maladie.

La surveillance active. L'objectif est de différer la mise en route du traitement curatif par une surveillance active initiale. Elle est indiquée pour les cancers jugés (là est tout le problème...) peu agressifs. L'évolution du cancer est surveillée par un examen clinique, un dosage du PSA total tous les 6 mois couplé à un toucher rectal ainsi qu'un bilan par biopsies à 1 an puis tous les 2 à 3 ans.

D'autres types de traitement curatifs mini-invasifs on récemment fait leur apparition : **l'ablathermie, la cryothérapie et le laser**.
A noter que les premiers essais cliniques de traitement par ablathermie, utilisant alors un prototype developpé au laboratoire LabTau (Lyon, France), ont été initiés en 1993 à l'hôpital Edouard Herriot (Lyon, France). C'est un traitement peu invasif qui traite le cancer de la prostate en concentrant des ultrasons focalisés de haute intensité (HIFU) qui vont détruire les cellules de la glande par la chaleur sans endommager les tissus environnants. Il est recommandé pour les patients porteurs d'un cancer localisé pour lesquels la chirurgie n'est pas recommandée en raison de l'âge (plus de 70 ans) et/ou de l'état général (par exemple obésité morbide). Il est également indiqué en traitement de rattrapage pour les patients présentant une récidive locale d'un cancer de prostate initialement traité par radiothérapie externe.

Ces nouvelles technologies alternatives pourraient permettre un traitement focal de la tumeur. Néanmoins, elles sont actuellement utilisées exclusivement en traitement global, pour détruire les tissus sur l'ensemble la glande prostatique. En effet, pour pouvoir envisager un traitement focal, il est nécessaire de connaître précisément la position des foyers tumoraux dans la glande, ce qui n'est actuellement pas réalisable. Pourtant, l'intérêt d'un traitement localisé est évident : traiter la lésion sans détruire l'organe, tout en préservant les fonctions urinaires et sexuelles, et en respectant les tissus environnants (rectum et vessie).

2.8 Imagerie du cancer de la prostate

L'utilisation de la tomographie par émission de positron (TEP), technique d'imagerie fonctionnelle utilisant la désintégration d'un traceur radioactif (le fluor 18) couplé à la choline pour mettre en valeur les zones de forte activité métabolique, est malheureusement d'utilité limitée pour la visualisation et la gradation des tumeurs prostatiques à cause du faible métabolisme des tumeurs, du faible contraste et d'une résolution spatiale limitée [31]. En tomodensitométrie (CT), le contraste dans les tissus mous est très faible [141] rendant, là encore, cette modalité très peu exploitée (voir figure 3.1c, page 30). Ces techniques trouvent en revanche leur place pour le bilan d'extension ganglionnaire et métastatique.

L'imagerie ultrasonore (US) avec sonde endorectale a longtemps été la modalité d'imagerie principale pour la prostate puisqu'elle est non-invasive, peu coûteuse, facile d'accès et peut être utilisée en temps réel. Néanmoins, les images sont difficiles à lire à cause d'un rapport signal/bruit faible et du bruit de speckle inhérent aux ultrasons [13] (voir figure 3.1b, page 30). L'échographie n'est ni sensible ni spécifique dans la détection du cancer de prostate (pas plus d'ailleurs que les techniques dérivées de l'écho-Doppler). Elle sert donc uniquement lors des biopsies, pour le guidage de la sonde endorectale et des aiguilles. De nouvelles techniques d'imagerie échographique prometteuses sont néanmoins

en cours d'évaluation. Il s'agit notamment de l'élastographie par compression ou onde de cisaillement, l'échographie de contraste ou encore l'histoscanning.

Quel usage de l'IRM ?

En dehors du bilan d'extension local du cancer, l'imagerie médicale peine à trouver une place pour le diagnostic du cancer. De gros efforts de recherche ont été entrepris ces vingt dernières années pour tenter de développer une méthode d'imagerie fiable du cancer de prostate permettant d'établir une cartographie tumorale. Un telle modalité d'imagerie serait l'élément clef pour pouvoir assurer une prise en charge optimale du cancer de la prostate. Elle pourrait :

- permettre un **second tri des patients avant de pratiquer les premières biopsies**. On estime actuellement que 40-50% des biopsies sont actuellement pratiquées inutilement,
- être une **alternative à la pratique régulière de biopsies de contrôle**, pour la surveillance des patients à fort risque, présentant par exemple un PSA élevé, ou une cible de Gleason faible détectée mais non traitée, ...).
- **guider une première série de biopsies** sur des zones présentant un signal suspect,
- **guider une re-biopsie** : des biopsies ciblées peuvent être réalisées dans certains cas afin d'éviter un excès de biopsies après une série de biopsies négatives et en cas de suspicion clinique forte,
- contribuer à l'**évaluation non-invasive de la stadification du cancer** de prostate en fournissant une information précise sur sa localisation et son extension.
- permettre un **traitement focal rationnel**,
- permettre le **suivi post-traitement**.

L'IRM prostatique, à condition d'être multiparamétrique, se positionne actuellement comme la modalité la plus prometteuse pour la visualisation de ce cancer. En effet, elle permet une visualisation précise de la glande et des structures des tissus. Son utilisation est ainsi très discutée au sein de la communauté médicale [1]. Reste néanmoins à parvenir à définir de bons critères d'évaluation des images...

2.9 Conclusion

Le cancer de la prostate est un problème de santé publique majeur. Depuis dix ans, les publications scientifiques sur le cancer de la prostate regroupant la recherche fondamentale, les études épidémiologiques, cliniques et les essais thérapeutiques ont été extrêmement nombreuses. Néanmoins, le cancer de la prostate détient aujourd'hui le triste record de premier cancer chez l'homme, en France et de deuxième cause de mortalité par cancer après le cancer du poumon.

L'évaluation pronostique à partir du couple TR/PSA et du résultat des biopsies est

malheureusement insuffisante pour décider du traitement approprié et le risque de sous- et sur-traitement est grand.

Bien que des technologies récentes (HIFU, cryothérapie, laser), qui pourraient permettre un traitement focal efficace du cancer, aient démontré leur intérêt thérapeutique, la prostatectomie totale reste actuellement le traitement de référence des tumeurs localisées. En effet, un traitement focal ne peut être envisagé que si une cartographie précise de la localisation des foyers malins dans la glande peut être dressée, ce que ne permettent pas les biopsies.

L'imagerie IRM-mp se positionne actuellement comme la technique la plus prometteuse pour l'amélioration du diagnostic du cancer de la prostate. Comme nous le verrons section 3.8, page 45, de nombreuses études s'attachent à démontrer son potentiel pour la visualisation de ce cancer. Les enjeux cliniques sont considérables puisqu'elle ouvrirait ainsi la voie au traitement focal, alternative à la prostatectomie radicale.

Chapitre 3

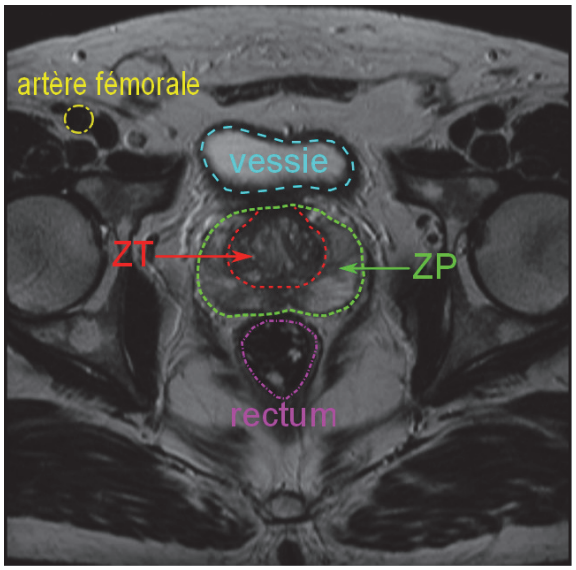
Imagerie par Résonance Magnétique de la prostate

3.1 Introduction

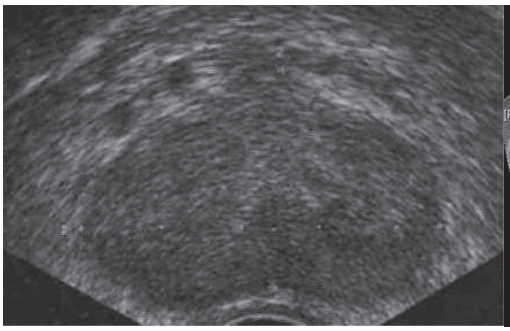
L'utilisation de l'imagerie par résonance magnétique pour l'inspection de la prostate (bilan d'extension locale) a débutée au début des années 1990 avec la séquence en T2-pondéré et a alors permis des avancées majeures par rapport aux autres modalités d'imagerie existantes, scanner CT et échographie notamment (voir la figure 3.1). En effet, contrairement à ces autres modalités, l'IRM permet une visualisation précise des contours prostatiques (ZP, ZT) et des structures pelviennes avoisinantes (rectum, vessie et urètre notamment) grâce à une haute résolution spatiale et en contraste, et offre la possibilité de réaliser une imagerie multi-plans avec un large champ de vue [21].

Développement de l'IRM prostatique multi-paramétrique (IRM-mp)

L'IRM prostatique, à condition d'être multi-paramétrique, est aujourd'hui considérée comme la modalité de choix pour l'étude du cancer de la prostate. L'exploration prostatique ne se limite plus à la seule séquence T2-w. Au début des années 2000, la séquence d'imagerie de perfusion a fait son apparition, suivie, quelques années plus tard, de la séquence de diffusion. De nos jours, les centres d'imagerie modernes réalisent un examen d'IRM multi-paramétrique pour le bilan d'extension local du cancer. Celui-ci combine différents types de séquences permettant d'accéder à une information à la fois morphologique (séquence T2-w) et fonctionnelle (séquences de diffusion ou de perfusion). Plusieurs



(a) IRM



(b) Echographie

(c) CT

Figure *3.1* – *Différentes modalités d'imagerie de la prostate*

équipes d'urologie en évaluent aujourd'hui l'apport pour le guidage des re-biopsies après une première série de biopsies négatives.

Après avoir brièvement décrit le principe de l'imagerie par résonance magnétique, nous détaillons dans les sections suivantes les spécificités des trois séquences IRM sus-citées, les plus fréquemment exploitées pour l'analyse de la prostate et qui sont en particulier utilisées dans notre étude.

3.2 Principe physique de l'Imagerie par Résonance Magnétique

Cette section présente les principes de base de l'IRM. L'imagerie IRM, à travers différentes séquences, permet de caractériser les propriétés physiologiques des tissus, en se basant sur la densité des protons et leur propriété de relaxation dans les tissus. En effet, elle consiste en l'observation de la résonance magnétique nucléaire (RMN) du noyau d'hydrogène ^{1}H, appelé proton, le plus facile à détecter par la RMN, grâce à sa grande concentration dans le corps humain (constitué à 70% d'eau) et à son grand rapport gyromagnétique γ. Cette observation nous permet ainsi de voir la répartition en eau dans l'organe que nous souhaitons étudier.

3.2.1 Résonance Magnétique

La RMN consiste à étudier l'aimantation d'un élément lorsqu'on le place dans un champ magnétique. On peut représenter le comportement d'un proton par son moment magnétique de spin μ. En l'absence de champ magnétique, les spins des protons sont orientés de manière totalement aléatoire, avec une résultante nulle ($\mathbf{M} = \sum \mu = 0$). Sous l'action d'un champ magnétique $\mathbf{B_0}$ fort et homogène, les spins s'alignent parallèlement à $\mathbf{B_0}$ ou anti-parallèlement (voir figure 3.2). L'orientation parallèle correspond à un niveau d'énergie faible, donc les protons sont plus nombreux dans la configuration parallèle que dans la configuration anti-parallèle et on a $\mathbf{M} = \sum \mu \neq 0$. En pratique, les spins ne s'alignent pas exactement selon le vecteur $\mathbf{B_0}$: les spins d'une partie des protons tournent autour de ce dernier en formant un angle α, nommé angle de précession. Chaque spin tourne autour de $\mathbf{B_0}$ avec une vitesse angulaire (vitesse de précession) ω_0 donnée par la relation de Larmor :

$$\omega_0 = \gamma B_0$$

où B_0 est l'amplitude du champ $\mathbf{B_0}$ et γ le ratio gyromagnétique, de l'ordre de 42.58 MHz/T dans le cas des protons.

3.2.2 Excitation et relaxation

On va maintenant perturber les moments magnétiques de spin décrits précédemment en appliquant un deuxième champ magnétique $\mathbf{B_1}$, perpendiculaire à $\mathbf{B_0}$ et de fréquence

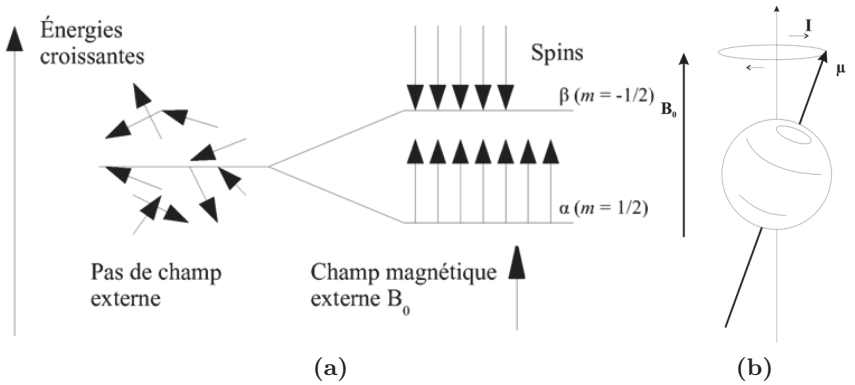


Figure *3.2* – *(a)Moments magnétiques sans champ et sous l'action d'un champ* $\mathbf{B_0}$. *(b) Mouvement de précession du spin autour de l'axe de* $\mathbf{B_0}$. *Source : Université de Laval.*

égale à la fréquence de Larmor ω_0. En IRM de prostate, les champs magnétiques appliqués sont de l'ordre de 1.5 - 3 Tesla (T). Ainsi, dans le cas de protons plongés dans un champ magnétique à 1.5 T (correspondant au champ utilisé dans cette étude), la fréquence de $\mathbf{B_1}$ sera d'environ 64 MHz. Cette fréquence étant située dans le domaine de fréquence des ondes radiophoniques, le champ $\mathbf{B_1}$ est une onde (électromagnétique) de radiofréquence (onde RF). On appelle période d'excitation la durée d'application de cette onde. Etant donné qu'on applique une onde radiofréquence (RF) de fréquence égale à la fréquence de Larmor ω_0, les spins entrent en résonance et il y a transfert d'énergie entre les spins en précession et l'onde RF, causant alors un basculement de l'énergie tissulaire. En d'autres termes, le vecteur macroscopique $\mathbf{M}$, tout en continuant à tourner (précesser) autour de $\mathbf{B_0}$ (Oz), à la fréquence angulaire ω_0, va se mettre également à précesser autour de $\mathbf{B_1}$ à la fréquence angulaire

$$\omega_1 = \gamma B_1.$$

En d'autres termes, les spins passent de l'orientation parallèle (énergie faible) à l'orientation anti-parallèle (énergie élevée). L'angle de bascule est donné par l'équation suivante :

$$\beta = \gamma \int B_1 dt.$$

Cet angle dépend donc directement de l'intensité de $\mathbf{B_1}$ et de sa durée d'application.

Lorsqu'on arrête d'appliquer l'onde RF, l'excitation s'arrête. Les spins vont alors retourner à l'équilibre, pendant ce qu'on appelle la phase de relaxation. Ce retour à l'équilibre correspond à l'émission d'une autre onde RF qu'on enregistre et qui constitue alors le signal RMN.

Cette relaxation est caractérisée par deux constantes de temps $T1$ et $T2$. Celles-ci correspondent respectivement aux relaxations des composantes longitudinale $\mathbf{M_z}$ et transversale $\mathbf{M_{xy}}$ du vecteur d'aimantation $\mathbf{M}$. Pour la composante longitudinale, on a :

$$M_z(t) = M_0(1 - e^{-\frac{t}{T1}}),$$

où M_0 représente l'intensité de l'aimantation tissulaire à l'équilibre (lorsque tous les spins sont alignés). La constante de temps $T1$ correspond au temps que met $\mathbf{M_z}$ pour revenir à 63% de sa valeur initiale et dépend directement du tissu. Cette relaxation $T1$ est aussi appelée *relaxation longitudinale*.

Pour la composante transversale, on a :

$$M_{xy}(t) = M_{xy}(0)e^{-\frac{t}{T2}},$$

où $M_{xy}(0)$ correspond à l'aimantation transversale au début de la relaxation. La constante de temps $T2$, correspond au temps que met $\mathbf{M_{xy}}$ pour atteindre 37% de sa valeur initiale. Cette relaxation $T2$ est aussi appelée *relaxation transversale*. Elle est décroissante et beaucoup plus rapide que la relaxation $T1$.

En réalité, du fait d'une hétérogénéité du champ magnétique $\mathbf{B_0}$, la relaxation transversale est donnée par une constante de temps $T2^*$ effective, différente de $T2$, définie par :

$$\frac{1}{T2^*} = \frac{1}{T2} + \gamma\Delta B,$$

où ΔB représente les inhomogénéités dans le champ $\mathbf{B_0}$. Plus ΔB est élevé, plus la relaxation $T2^*$ sera rapide. Ce principe est utilisé dans la technique d'écho de gradient, où on applique un déphasage rapide suivi d'une relaxation rapide des spins en utilisant un champ $\mathbf{B_0}$ inhomogène contrôlé.

3.2.3 Encodage de l'image

Nous avons vu comment former des signaux qui vont être captés par l'antenne réceptrice. Pour reconstruire l'image, il faut localiser l'origine du signal dans l'espace. Cela se fait en ajoutant un gradient directionnel dans les trois directions de l'espace sur le champ magnétique $\mathbf{B_0}$, grâce aux bobines de gradient de champ magnétique. Un gradient de sélection de coupe est tout d'abord appliqué, le plan de coupe sélectionné ayant alors une largeur dépendant de la bande passante et de l'intensité de ce gradient. On différencie ensuite les deux directions du plan de coupe avec un gradient de phase et un gradient de fréquence. La réponse à ces différents gradients permet de remplir un espace de Fourier appelé *espace des k*. On utilise ensuite des algorithmes de transformée de Fourier rapide afin de revenir dans le domaine image et reconstruire celle-ci. Différentes techniques d'acquisition permettent de créer cet espace des k plus ou moins rapidement, en fonction des besoins cliniques. On parle alors de stratégie de remplissage de l'espace des k, qui peut être linéaire, sphérique ou en spirale par exemple. De nos jours, la résolution spatiale en IRM prostatique est de l'ordre du millimètre.

3.2.4 Séquences de base en IRM

Pour former le signal RMN décrit précédemment, on utilise différents types de séquences IRM. Ces dernières sont des séries d'impulsions et d'applications de gradients

faites à des instants bien précis, que l'on utilise selon le type d'image que l'on souhaite obtenir.

Echo de spin

La séquence d'écho de spin (ou spin-écho) est considérée comme la séquence de base en IRM. Dans la technique d'écho de spin (ou spin-echo), on émet une onde RF longitudinale de 90° qui va causer le basculement des spins dans le plan transverse xy. Au bout d'un temps $\frac{T_E}{2}$, on va appliquer une onde RF transversale de rephasage de 180° qui va avoir comme effet de rephaser les protons déphasés du fait des hétérogénéités du champ $\mathbf{B_0}$. On lit le signal au temps T_E et on répète l'enchaînement au temps T_R. T_E est appelé temps d'écho et T_R est appelé temps de répétition. On notera que l'impulsion de 180° de rephasage permet d'obtenir un signal $T2$ vrai et non pas $T2^*$.

Echo de gradient

Dans ce type de séquences, on rephase les spins en utilisant un gradient. La séquence d'écho de gradient utilise un angle de bascule plus faible que celui de la séquence écho de spin, ce qui permet un retour à l'équilibre plus rapide et donc de raccourcir le temps T_R. Il en résulte donc un temps d'acquisition plus court que pour les séquences par écho de spin. Ici, le signal d'écho est généré en appliquant d'abord un gradient de codage de fréquence pour déphaser les spins, puis un gradient identique inversé pour les rephaser. Dans ce cas, contrairement aux images en écho de spin, les inhomogénéités du champ $\mathbf{B_0}$ ne sont plus corrigées, on obtient donc des images pondérées en $T2^*$.

3.2.5 Pondérations

Les séquences IRM peuvent être plus ou moins pondérées en $T1$, $T2$ et densité de protons. Nous proposons ci-après une présentation succincte du principe de la pondération des séquences.

Pondérations $T1$, $T2$

Soit T_E le temps d'écho, la durée qui sépare le milieu de l'onde RF d'excitation et le milieu du temps de lecture du signal émis par les protons lors de la relaxation. On rappelle que T_R est le temps de répétition, c'est-à-dire le temps au bout duquel on répète le processus d'excitation. La pondération des séquences en $T1$, $T2$ ou en densité de proton est réalisée en faisant varier les constantes de temps T_E et T_R.

Pondération T1

En effet, si le temps T_R est la durée entre deux ondes RF de 90°, il conditionne la relaxation longitudinale des tissus étudiés, valeur qui dépend du $T1$. Plus le T_R est long (supérieur à 1800 ms avec une IRM à 1.5 T), plus M_z tend vers M_0, rendant alors les tissus difficiles

à différencier car les protons ont fini leur relaxation et retrouvé un état d'équilibre au moment de la nouvelle bascule. Une diminution du temps T_R (inférieur à 600 ms avec une IRM à 1.5 T) permet de pondérer l'image en $T1$ (T1-w, pour *T1 weighted*) car on va mettre en évidence les différences entre les tissus. En pratique, sur les images finales, ceux qui ont un temps $T1$ court apparaîtront en blanc tandis que ceux caractérisés par un temps $T1$ long apparaîtront en noir.

Pondération T2

Dans le cas où on utilise un temps T_E court (inférieur à 50 ms avec une IRM à 1.5 T), on ne peut pas différencier les tissus en fonction de leurs $T2$ respectifs car les aimantations transversales M_{xy} des différents tissus n'auront pas eu le temps de se différencier. Si le temps T_E est long (supérieur à 60 ms avec un IRM à 1.5 T), la séquence permet alors de distinguer les tissus en fonction de leur propriété $T2$; elle est dite pondérée en $T2$ (T2-w, pour *T2 weighted*). En pratique, sur les images finales, les tissus ayant un temps $T2$ court apparaîtront en noir tandis que ceux ayant un $T2$ long apparaîtront en blanc.

La figure 3.4 présente des images de prostate pondérées en $T1$, $T2$.

On remarque qu'il est plus facile d'obtenir une image bien pondérée en T2 ("exclusivement") qu'une image bien pondérée en T1. En effet, pour bien dépondérer en T1 il suffit d'allonger le temps T_R (seule contrainte : augmenter la durée de la séquence...), par contre, pour bien pondérer une image en T1 ("exclusivement"), il faudrait théoriquement réduire T_E "à zéro". Mais il est difficile de beaucoup réduire le T_E car il doit (toujours) être égal à deux fois le temps $T_E/2$ (contraintes instrumentales : délai le plus court pour appliquer une impulsion RF de 180° après l'impulsion RF de 90° $\backsimeq$ 20ms).
Le contraste est meilleur sur une séquence en T2 mais le rapport signal sur bruit est plus faible car les mesures sont réalisées "tardivement" (T_E long) sur la courbe d'atténuation du signal en T2 (signal plus faible). Une séquence pondérée "exclusivement" en T2 sera souvent notée T2-map (pour T2 *mapping*).

A noter que si la séquence n'est pondérée ni en $T1$ ni en $T2$ (T_R long et T_E court), les tissus sont alors différenciés en densité de protons.

Pondération en diffusion

Cette technique exploite le phénomène de diffusion, c'est-à-dire les mouvements microscopiques (browniens) liés à l'agitation thermique des molécules d'eau (voir figure 3.3). Ces déplacements peuvent se faire dans toutes les directions de l'espace de façon équilibrée (dite "isotrope"), ou de façon déséquilibrée, dans une orientation particulière (dite "anisotrope"). La séquence comporte l'application successive de deux gradients intenses et symétriques autour d'une impulsion spin-écho de 180°, qui ont pour rôle d'appliquer aux protons un déphasage de précession qui dépend de leur position. Le premier gradient donne à la précession des protons une avance de phase proportionnelle à leur position sur l'axe

du gradient. Le second gradient est exactement opposé au premier et impose un retard de phase de même angle aux protons. Les protons n'ayant pas bougé entre les deux impulsions sont donc rephasés ("déphasage" nul) et émettent du signal. Les molécules mobiles, qui se sont déplacées entre chaque application de gradient, en revanche, ne sont pas ou sont incomplètement re-phasées, ce qui entraîne une perte de signal proportionnelle à leur amplitude de déplacement (dans la direction étudiée, i.e. l'axe des gradients de diffusion). Les gradients peuvent être appliqués dans n'importe quelle direction de l'espace. Ils sont habituellement appliqués selon trois axes deux à deux perpendiculaires et l'image isotrope est produite en moyennant les trois signaux recueillis. Tout comme le degré de pondération en T2 dans une séquence d'écho de spin est défini par le temps d'écho T_E, le degré de pondération en diffusion est établi par la force des gradients et leur temps d'application qui sont intégrés dans ce qui est appelé le facteur de gradient b (lire section 3.6).

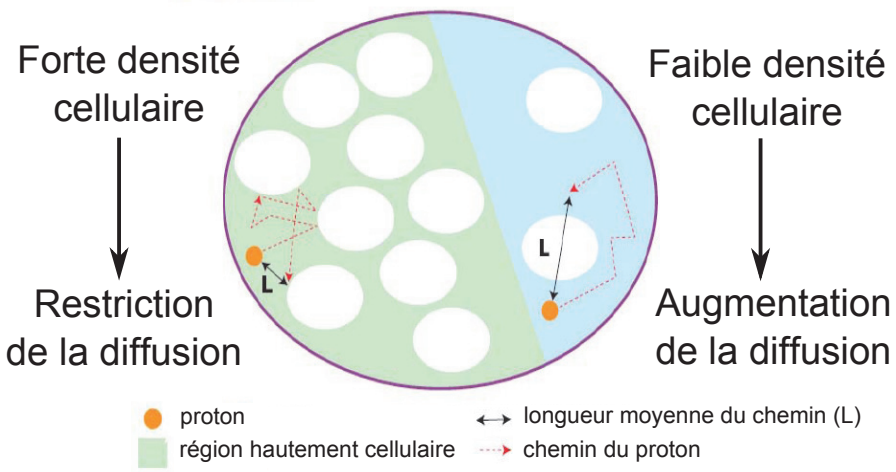


Figure *3.3* – *Principe de la séquence de diffusion : mesure des déplacements microscopiques des protons de l'eau (mouvements browniens). Cette mobilité de l'eau reflète indirectement la micro-structure des tissus. Source : Brunelle et coll. [12].*

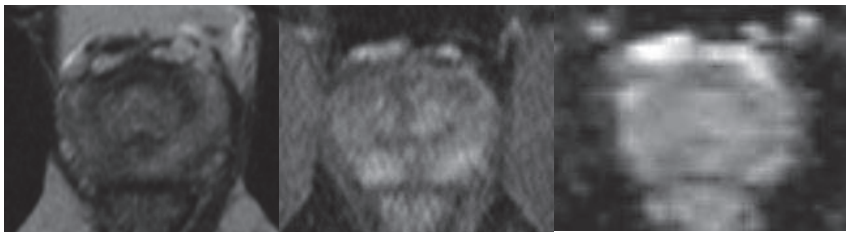
Technique de saturation sélective du signal de la graisse (*fatSat*)

Les protons de la graisse ont une fréquence de résonance différente des protons de l'eau. Cette différence est de 3,25 ppm soit environ 208 Hertz (Hz) à 1.5 T. Il est ainsi possible de supprimer la composante graisseuse des tissus. On réalise cela en incorporant dans la séquence une impulsion sélective centrée exactement sur le pic de résonance de la graisse (impulsion dont la fréquence est décalée de 208 Hz par rapport à l'impulsion habituelle) d'où l'annihilation de l'aimantation longitudinale de ce tissu (saturation). Ainsi lors de l'impulsion de 90° suivante, le signal de la graisse n'aura pas eu le temps de "repousser" par rapport aux autres tissus : on réalise de cette façon une suppression de son signal. Cette technique est habituellement appelée *FatSat*. Elle est très intéressante pour mettre

en évidence des lésions à proximité de structures graisseuses en particulier après injection de gadolinium (cf. la section 3.5 dédiée à l'imagerie dynamique avec injection de produit de contraste). De plus, cette technique permet de supprimer le signal de la graisse sans altérer la visualisation des tissus ayant un $T1$ équivalent.

3.3 L'IRM prostatique multi-paramétrique en pratique clinique : aspects techniques

Il est aujourd'hui admis que, pour apporter le plus d'informations, un examen de type IRM multi-paramétrique (IRM-mp) de la prostate doit se composer au moins des séquences en T2-pondérée, de diffusion et de perfusion avec injection d'agent de contraste (acquise en T1-pondérée) (voir figure 3.4). Ces trois types d'acquisition, qui seront d'ailleurs utilisés dans le cadre de cette thèse, sont présentés dans les sections 3.4, 3.5 et 3.6 ci-après. Dans cette section, nous présentons des aspects techniques spécifiques à l'acquisition IRM-mp de prostate.



(a) T2-w **(b)** DCE (T1-w) **(c)** ADC

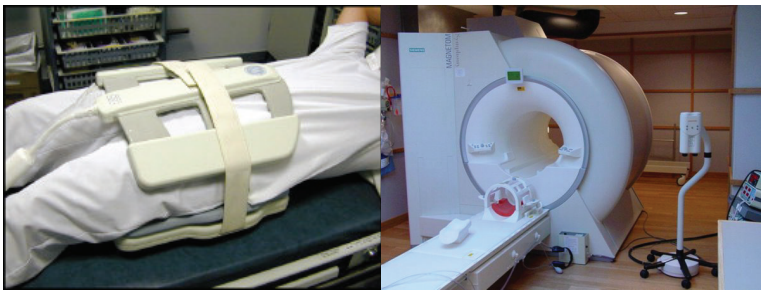
Figure *3.4* – *IRM multi-paramétrique de prostate réalisée sur un patient âgé de 53 ans. Coupes axiales dans le plan médian.*
(a) Image en T2-pondérée (T2-w) turbo-spin-écho.
(b) Image en T1-pondérée (T1-w) écho de gradient FatSat, extraite de la séquence dynamic contrast enhanced (DCE) aquise 45 s. après injection de l'agent de contraste (AC).
(c) Cartographie du coefficient apparent de diffusion construite à partir de la séquence de diffusion (DWI).

3.3.1 Acquisition

L'imagerie de référence est la séquence acquise en T2-pondérée (T2-w). Elle est souvent réalisée dans plusieurs plans orthogonaux (au moins deux, parfois trois) :

- coupes axiales, perpendiculaires au bord postérieur de la prostate (figure 3.6a) ;
- coupes sagittales (figure 3.6b) ;
- coupes coronales obliques, dans le plan des vésicules séminales (figure 3.6c).

Le plan de référence est le **plan axial** oblique perpendiculaire à la paroi rectale. Il doit être privilégié pour les séquences de diffusion et de perfusion. Il est important que les coupes axiales des différentes séquences aient la même épaisseur et le même positionnement pour



(a) Antenne pelvienne **(b)** Imageur IRM

Figure 3.5 – *(a) Imageur IRM Siemens Magnetom Symphony MR 2004A (Siemens Medical Systems, Erlangen, Allemagne) et (b) antenne pelvienne en réseau phasé, utilisés pour les acquisitions IRM-mp présentées dans cette thèse.*

permettre une comparaison précise du signal d'une zone donnée sur chaque séquence. On réalise préférentiellement des coupes de 3 à 4 mm d'épaisseur incluant l'ensemble de la prostate et des vésicules séminales.
Le protocole d'acquisition utilisé dans notre étude est détaillé au chapitre 6.4.

3.3.2 Antennes

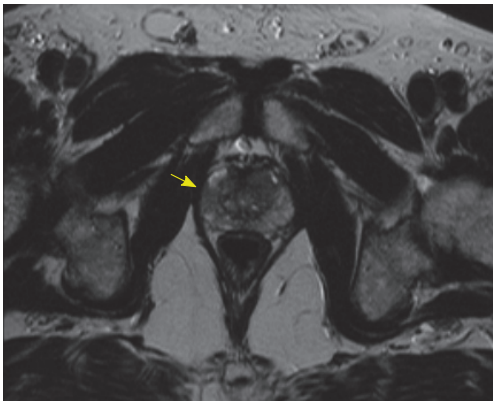
En imagerie de la prostate, l'acquisition est obtenue soit avec une antenne pelvienne haute résolution en réseau phasé (dite "de surface" par abus de langage...), soit avec une antenne endorectale. L'usage de l'antenne endorectale seule n'est cependant plus recommandé (signal trop faible en avant de la prostate). Elle peut être utilisée en complément de l'antenne pelvienne.

Antenne endorectale

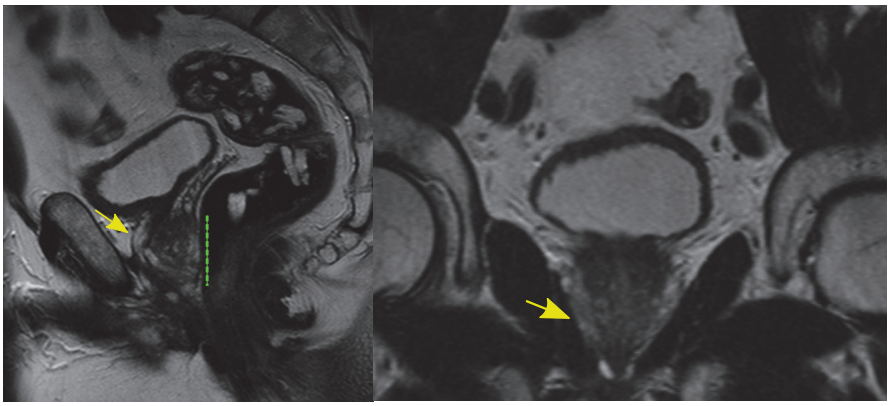
Elle est située à l'intérieur d'un ballon gonflable placé dans le rectum. Son avantage principal est son excellente résolution spatiale, avec une très bonne visibilité des territoires postérieurs de la prostate. Sa limite est la chute du signal dans les territoires antérieurs. Ses inconvénients sont l'inconfort du patient, les artefacts de mouvements dus aux contractions rectales involontaires, les artefacts de brillance avec un signal inhomogène, la compression de la glande et enfin le surcoût de l'examen.

Antenne pelvienne en réseau phasé

Ces antennes sont formées de plusieurs éléments recevant le signal sur une petite surface, ce qui augmente le rapport signal sur bruit. Elles furent initialement développées pour des applications cardiaques et étendues ensuite à l'imagerie prostatique. Elles sont placées sur le pelvis du patient. Les avantages de ces antennes sont la qualité et l'homogénéité du signal acquis, le confort du patient et le moindre coût. Elles autorisent un champ de vue réduit de 16 cm avec une résolution spatiale quasi-identique à celle d'une antenne



(a) Coupe axiale (référence)



(b) Coupe sagittale **(c)** Coupe coronale

Figure *3.6* – *Plans d'acquisition des images IRM de prostate. Acquisition en T2-pondérée (T2-w) turbo-spin-écho, réalisées sur un patient de 63 ans présentant 3 foyers malins. La position de la prostate est indiquée par une flèche jaune. (a) Acquisition selon le plan de référence axial ; (b) acquisition selon le plan sagital (paroi rectale marquée par les pointillés verts) ; (c) acquisition selon le plan coronal.*

endorectale (voir figure 3.5a).

La nécessité de l'utilisation d'une antenne endorectale reste débattue. S'il est admis par la majorité des auteurs que l'antenne endorectale permet une meilleure détection des envahissements extra-capsulaires de petite taille, il n'est pas certain que son apport soit déterminant dans la détection des foyers tumoraux eux-mêmes. D'excellents résultats ont été obtenus dans ce domaine sans antenne endorectale, du moins quand une approche multi-paramétrique était suivie [100].

Notre étude utilise uniquement une antenne pelvienne haute résolution en réseau phasé.

3.3.3 Artefacts et mouvements

Acquisition après biopsies

Si l'IRM est réalisée après biopsies, il est recommandé d'attendre au moins 8 semaines pour éviter les artefacts sanguins post-biopsie qui gênent l'interprétation et peuvent soit masquer des tumeurs soit être causes de faux positifs. Malheureusement, ce délai de 8 semaines, s'il atténue les artefacts, ne les fait pas tous disparaître. Une acquisition en T1-pondérée sans injection d'agent de contraste (généralement intégrée à la séquence dynamique) doit donc être systématiquement réalisée pour repérer ces foyers hémorragiques.

Mouvements du rectum

Outre les possibles mouvements du patient (difficilement contrôlables) qui peuvent entraîner un décalage entre les images de différentes séquences, les contractions du rectum peuvent également introduire un biais dans la comparaison des différentes séquences. L'injection de 1 milligramme (mg) de glucagon avant le début de l'examen réduit les artefacts liés au péristaltisme intestinal mais ne les supprime pas tous.

3.4 Imagerie en pondération T2 (T2-w)

3.4.1 Principe

La séquence en pondération T2 (voir figure 3.7) est souvent désignée sous le nom de séquence "morphologique" : elle permet en effet de visualiser les différentes zones prostatiques puisqu'on observe un contraste notable de signal entre la ZP hyper-intense et la ZT en hypo-signal T2, due au fait que la ZP contient un espace liquidien plus important que la ZT. Comme nous l'avons déjà mentionné, la séquence en T2-pondéré reste la séquence de référence pour la visualisation des tissus prostatiques.

3.4.2 Acquisition

Le principe de l'acquisition des images pondérées en T2 (T2-w) a été donné section 3.2.5. La séquence T2-w est généralement réalisée en écho de spin rapide aussi appelée *turbo spin écho* (TSE). C'est une variété de séquence d'écho de spin où x échos successifs sont utilisés pour remplir x lignes de pas de codage de phase de l'espace de Fourier (cf. section 3.2.3), réduisant ainsi la durée d'acquisition d'un facteur x (facteur turbo). La pondération T2 de l'image est accrue. C'est ce type de séquences qui est utilisé pour l'acquisition des images T2-w de notre étude.

3.4.3 Inspection visuelle

La présence d'une lésion cancéreuse est suspectée sur des plages d'hypo-signal focalisées, asymétriques (voir 3.7).

Figure *3.7* – *Coupes axiales de la prostate en IRM T2-w turbo-spin-echo (T_E=109ms, T_R=7750ms) acquises chez un patient de 53 ans.*
L'image en haut à gauche représente la base. L'image en bas à droite représente l'apex (cf. section 2.3 page 18 dédiée à l'anatomie de la prostate).
Ce patient présente deux foyers malins en ZP postérieure droite et gauche.

3.5 Imagerie de perfusion (DCE)

3.5.1 Principe

L'IRM de perfusion (DCE pour *dynamic contrast enhanced*) tente d'exploiter les caractéristiques néo-angiogéniques des tumeurs. Différentes études ont en effet montré qu'au-delà de 200 μm, les tumeurs induisent une néoangiogenèse indispensable à leur croissance, responsable d'une augmentation de la microvascularisation. Celle-ci diffère de celle du tissu normal par une complexité et une augmentation de volume du flux sanguin échangé, une perméabilité accrue par fragilisation de la barrière endothéliale et enfin un développement de l'espace interstitiel, dit aussi extravasculaire/extracellulaire (EEE) [94].

La séquence de perfusion est basée sur une injection intraveineuse en bolus d'un agent de contraste (AC) paramagnétique de faible poids moléculaire (le chélate de Gadolinium) diffusant facilement dans l'espace interstitiel et permettant d'analyser la perméabilité de la barrière endothéliale. On effectue une série d'acquisitions correspondant à un échantillonnage temporel fixé afin d'obtenir une acquisition dynamique de la cinétique de diffusion de l'AC. L'évolution du signal mesuré dans la glande peut être représentée par des courbes de réhaussement desquelles on peut extraire des informations semi-quantitatives ou quantitatives (que nous étudierons section 7.2.3, page 104).

L'interstitium étant théoriquement plus important dans les tissus cancéreux que dans les tissus sains, la concentration en produit de contraste y sera plus grande. Le flux d'entrée du produit de contraste est appelé "réhaussement" ou "wash-in" et son retour vers le plasma sanguin après la diffusion interstitielle est le "lavage" ou "wash-out".

Un exemple de courbe d'intensité du signal en fonction du temps est donné dans la figure 3.8.

Les zones prostatiques tumorales ont théoriquement des phases de wash-in et wash-out plus rapides que les zones saines [129], mais il existe un chevauchement important, incluant à la fois des faux positifs (tissu inflammatoire, composante fibromusculaire de l'hypertrophie bénigne, nodules stromaux etc), et des faux négatifs (cancers avec un faible réhaussement). Ce chevauchement dans l'échelle du signal est lié à des facteurs physiologiques (débit sanguin, perméabilité capillaire/tissulaire/tumorale, perfusion tissulaire, etc) et physiques (caractéristiques de l'injection _ lieu, débit _, dose et concentration du produit de contraste, choix du produit de contraste _ poids moléculaire, dispersibilité etc _, paramètre de la machine IRM et de la séquence choisie _ T_R, T_E, gain, homogénéité des gradients _, T1 et T2 natifs du tissu etc) modifiant le contraste après injection.

3.5.2 Acquisition

La séquence de perfusion est pondérée en T1 (T1-w), en écho de gradient, 2-D ou 3-D, avec de nombreuses variantes destinées à optimiser la résolution temporelle et/ou le rapport signal/bruit, mais qui peuvent rendre la séquence incompatible avec la pratique clinique (temps d'acquisition trop long, résolution spatiale trop faible) ou l'analyse

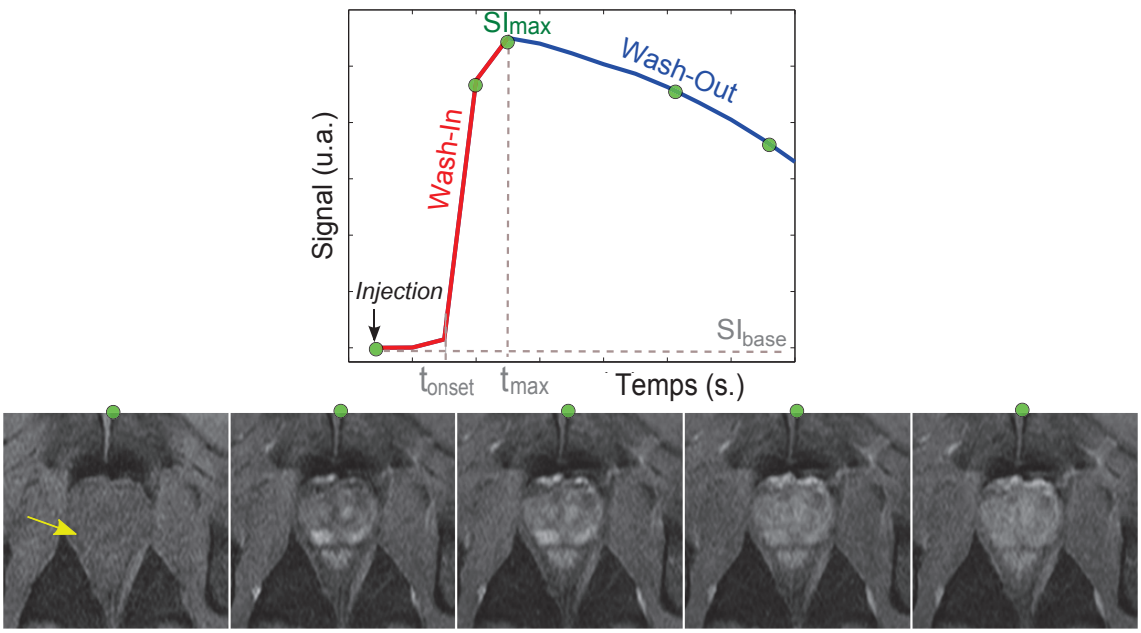


Figure *3.8* – *Courbe du signal moyen d'une région d'intérêt maligne (pointée par la flèche jaune) obtenue sur une séquence de perfusion IRM acquise en pondération T1 (environ 3 minutes). On observe initialement un signal "de base" juste après l'injection du produit de contraste mais avant son arrivée massive dans les capillaires de la région d'intérêt et son accumulation dans l'EEE, jusqu'à un pic d'intensité. Cette première phase s'appelle le "wash-in". Suit ensuite une décroissance du signal correspondant au retour du produit de contraste de l'EEE vers le compartiment vasculaire. Cette décroissance est plus ou moins lente en fonction de la perméabilité de la barrière endothéliale, du flux sanguin, et de la quantité de produit accumulée. On appelle cette phase le "wash-out". On peut aussi observer une phase de plateau quand la concentration en gadolinium est équilibrée entre le compartiment vasculaire et le compartiment extracellulaire.*

quantitative. L'extraction de paramètres quantitatifs (décrite section 7.2.3) nécessite en effet d'acquérir les images à haute résolution temporelle (5-15 secondes environ), ce qui n'est pas toujours compatible avec l'acquisition de plusieurs coupes permettant de couvrir l'ensemble de la glande avec de surcroît une résolution spatiale suffisante pour le diagnostic. Cette séquence est donc soumise à un compromis entre résolution spatiale, résolution temporelle et couverture anatomique.

3.5.3 Analyse des images

Les images après injection les plus précoces (celles où le produit de contraste apparaît) sont souvent les plus informatives car elles présentent les zones où le flux vasculaire ou la perfusion sont les plus importants. Un réhaussement focalisé, asymétrique, plus intense ou précoce que celui du reste de la glande oriente vers une lésion tumorale (cf. la cible pointée sur la figure 3.8). Afin de limiter les artefacts et d'accentuer le contraste, on peut soustraire le signal de la graisse à l'acquisition, ou soustraire chaque série par rapport à l'image acquise pré-injection (on obtient un signal de réhaussement "relatif", cf. section

7.2.1).

3.6 Imagerie de diffusion (DWI)

3.6.1 Principe

Dans une image pondérée en diffusion (DWI, pour *diffusion-weigthed imaging*), dont le principe a été expliqué section 3.2.5, les structures ayant une diffusion rapide sont sombres parce qu'elles sont sujettes à une plus grande atténuation de leur signal tandis que les structures caractérisées par une diffusion lente sont claires. Des cartes de coefficient apparent de diffusion (ADC) peuvent être construites à partir d'une série d'images pondérées en diffusion réalisées avec des valeurs différentes de gradients. Les images ADC fournissent une information quantitative par le calcul d'un coefficient de diffusion apparent inversement proportionnel à la densité cellulaire du tissu. Dans ces images, la diffusion seule est responsable du contraste ; il n'y a plus de contribution de T1 ni de T2 au contraste de l'image. Dans ces cartes de diffusion, le contraste est tel que les structures ayant une diffusion rapide sont hyperintenses alors que celles ayant une diffusion lente sont hypointenses. Néanmoins, elles souffrent d'une résolution spatiale limitée, y compris à 3 T. De plus, elles sont fortement sensibles aux artefacts (inhomogénéité de champ, air intrarectal, susceptibilité magnétique...) (cf. figure 3.4c). Ceux-ci peuvent être atténués grâce à la saturation du signal de la graisse et aux techniques d'imagerie parallèle.

Le facteur "b"

Nous avons vu, section 3.2.5, que la pondération en diffusion est déterminée par un "facteur de diffusion" nommé b (en s.mm^{-2}), lié à l'intensité des gradients de diffusion. Pour augmenter la pondération en diffusion, b doit être élevé. Une acquisition avec un facteur b élevé (supérieur à 500 s.mm^{-2}) sera sensible aux mouvements lents et courts des molécules d'eau, tandis qu'un facteur b faible (de l'ordre de 50 s.mm^{-2}) permettra de mieux détecter des mouvements plus rapides et plus amples.

On obtient pour chaque plan de coupe une image brute par valeur de b. Une cartographie du coefficient ADC calculé pour chaque pixel peut alors être réalisée si on dispose d'au moins deux séries d'images acquises avec des valeurs de b différentes. Pour l'étude de la prostate, on utilise généralement deux valeurs de b : une valeur à 0 s.mm^{-2} (notée b_0) et une entre 600 et 1000 s.mm^{-2}.

Il existe différentes techniques de calcul de l'ADC. La plus répandue fait l'hypothèse que l'atténuation du signal suit une loi de variation monoexponentielle en fonction de b. Le plus souvent, on utilise deux valeurs de gradient, b et b_0, et on calcule l'ADC comme suit :

$$\mathrm{ADC} = -\frac{1}{b - b_0} \ln(\frac{S}{S_0})$$

où S et S_0 sont les intensités mesurées après application des gradients b et b_0. On peut également utiliser plus de deux valeurs de b, l'ADC est alors calculé par régression mono-exponentielle.

Une autre technique, plus fine, consiste à tenir compte de la "fraction de perfusion" décrite par Le Bihan [11], exprimée pour les faibles valeurs de b (< 100 s.mm^{-2}) où deux composantes de l'ADC sont calculées par régression biexponentielle, la première correspond à la microcirculation, la seconde à la diffusion "vraie".

En pratique, un consensus récent [85] propose de s'affranchir de la première composante et de conserver un modèle monoexponentiel simple. C'est ce modèle qui est utilisé dans notre étude (cf. section 6.2 pour une description des paramètres utilisés).

3.6.2 Analyse des images

Ces séquences permettent de mesurer le mouvement des molécules d'eau. Ce mouvement aléatoire varie selon la taille cellulaire et la densité des structures extracellulaires. L'eau extracellulaire diffuse plus facilement que l'eau intracellulaire et son coefficient de diffusion est proche de celui de l'eau libre. Une tumeur, qui modifie le contenu cellulaire et le compartiment extracellulaire, va donc engendrer des perturbations du mouvement de l'eau entraînant une diminution du coefficient apparent de diffusion (ADC). Il est généralement observé que l'ADC est corrélé au score de Gleason. Plus le grade histologique est élevé, plus l'ADC est faible (plage en hypo-signal) [122].

3.7 Spectroscopie

L'imagerie par résonance magnétique spectroscopique est un outil permettant d'étudier la composition des métabolites d'un tissu. En spectroscopie, il existe une chute du pic de citrate et une élévation du pic de choline dans les foyers tumoraux prostatiques comparativement aux tissus sains. La spectroscopie peut être utilisée pour prédire la présence ou l'absence de cancer dans une région d'intérêt définie, mais n'améliore pas significativement la détection des foyers tumoraux en association à l'imagerie T2. Elle reste une séquence d'IRM fonctionnelle facultative, et, de par sa faible valeur ajoutée et son caractère très chronophage, a peu d'intérêt en pratique courante. Nous ne l'exploiterons pas dans notre étude.

3.8 Résultats actuels de l'IRM multi-paramétrique dans la détection tumorale

Nous proposons dans cette section un bref état de l'art des travaux d'évaluation des différentes séquences IRM par des observateurs humains, menés dans un contexte clinique. Dans la suite, les performances sont décrites par les critères classiques de sensibilité et

spécificité, définis dans la section 4.5, page 60.

D'après la littérature, l'imagerie T2-w seule n'a qu'un intérêt limité dans la détection tumorale avec une sensibilité de 25-60% et une spécificité de 57-98% [42, 51, 53, 95]. Elle ne permet pas non plus d'évaluation fiable du volume tumoral avec des sur et des sous-estimations par rapport au volume histologique [53]. Les causes de ces difficultés sont multiples : artefacts sanguins post-biopsie, anomalies bénignes apparaissant en hyposignal comme les cancers (zones de prostatite, de fibrose, d'atrophie glandulaire, etc), mais aussi un mauvais contraste entre certains cancers (notamment bien différenciés) et le tissu prostatique normal.

C'est ce constat de semi-échec qui a conduit à développer de nouvelles approches qui se sont intégrées dans ce que l'on appelle désormais l'IRM multi-paramétrique (IRM-mp).

Les séquences dynamiques ont nettement amélioré les performances de l'IRM en permettant notamment un gain significatif de sensibilité. C'est cette amélioration qui a véritablement lancé l'IRM en tant qu'outil de détection et de localisation tumorale. L'étude de Girouin *et coll.* [42], réalisée en 2007, compare les performances diagnostiques de trois lecteurs basée sur l'IRM T2-w seule, ou l'IRM DCE seule. La sensibilité de détection globale passait, pour les trois lecteurs, de 18-24% en T2 à 46-60% en dynamique (p-valeur <0.0001). En revanche, l'imagerie dynamique était légèrement (mais significativement) moins spécifique (94-98% versus 91-97%, $p<0.0001$). Parallèlement, Cornud *et coll.* [22] ont trouvé pour l'association T2/dynamique une sensibilité et spécificité respectivement de 77%, 91% pour les tumeurs de plus de 0.2 cc et de 90%, 88% pour celles de plus de 0.5 cc. L'IRM de diffusion, combinée à l'imagerie T2 et dynamique, a permis de poursuivre l'amélioration de la sensibilité et de la spécificité de l'IRM. Tanimoto *et coll.* [123] ont étudié une série de 83 patients avec un taux de PSA élevé (>4 ng/ml), ayant eu une IRM avant biopsie. L'aire sous la courbe ROC pour la détection des patients avec cancer était de 0.711 pour le T2 seul, de 0.90 pour l'association T2 et diffusion et de 0.96 pour l'association T2, diffusion, dynamique. Le groupe de Yoshizako *et coll.* [140] a également montré une amélioration progressive des performances de détection tumorale de l'IRM à mesure que le nombre de séquences était augmenté. L'aire sous la courbe ROC était ainsi de 0,62, 0.73, 0.84, et 0.92 pour des protocoles associant respectivement imagerie T2 seule, T2 + dynamique, T2 + diffusion et T2 + Dynamique + Diffusion. La valeur ajoutée de la spectroscopie, quant à elle, est plus discutée. La première (et seule) étude multicentrique comparant imagerie T2 seule et combinée à la spectroscopie a donné des résultats très décevants, avec des aires sous la courbe comparables pour les deux protocoles (0.60 pour l'imagerie T2 seule, 0.58 pour la combinaison T2 + spectroscopie) [136].

Cette analyse bibliographique montre qu'aucune séquence prise individuellement ne permet une détection fiable du cancer de prostate, mais souligne que la combinaison des données de plusieurs séquences permet une amélioration significative des performances diagnostiques de l'examen. On notera néanmoins que des études tentent de modérer les

résultats et conclusions parfois très positifs annoncés par certaines équipes en soulignant notamment que si ces résultats peuvent être obtenus dans des centres d'imagerie spécialisés, ils seront toutefois difficilement reproductibles dans d'autres centres [49, 55].

3.8.1 Limites d'interprétation de l'IRM prostatique multi-paramétrique

Dans la pratique clinique, le radiologue synchronise l'imagerie T2, l'imagerie de perfusion et l'imagerie de diffusion pour les analyser conjointement. Il cherche alors à apparier les zones suspectes repérées dans les différentes images, présentant un hyposignal T2, une restriction (hyposignal) de la diffusion et un réhaussement (hypersignal) précoce en perfusion. Néanmoins, cette interprétation à l'oeil d'un nombre d'images conséquent reste difficile et subjective et on observe une forte variabilité inter- et intra-observateurs [34, 38, 103, 115].

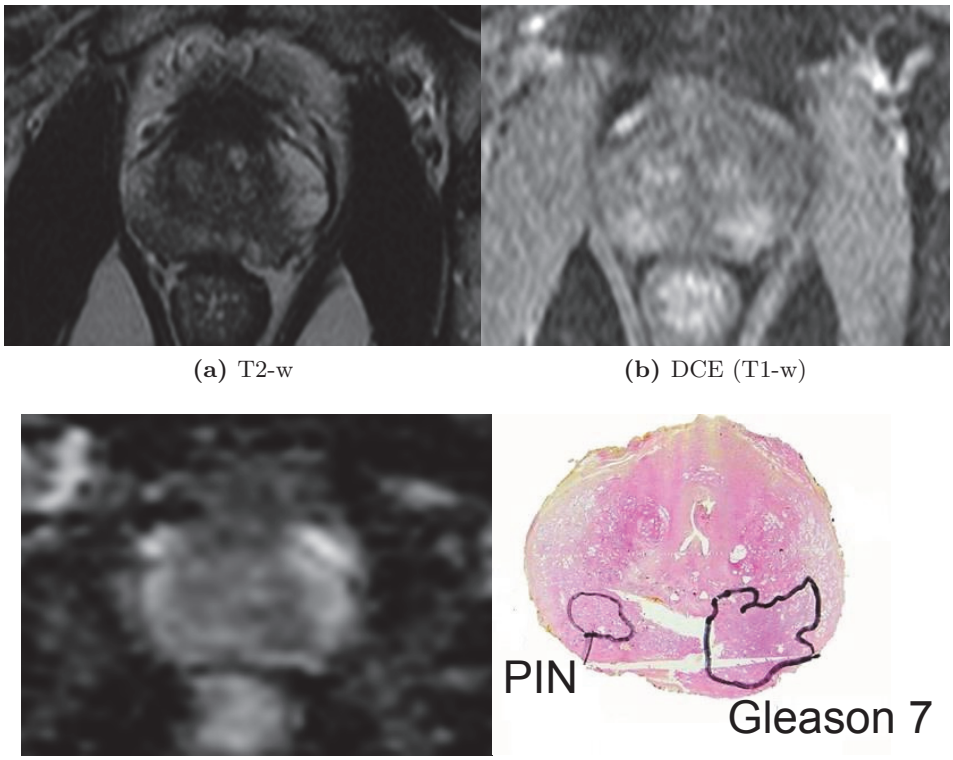
De plus, augmenter le nombre de séquences pose le problème de la gestion des données contradictoires : comment interpréter l'examen quand une séquence est positive, l'autre négative, et la troisième douteuse ? Les données d'une séquence doivent-elles l'emporter sur les autres en cas de discordance ?

La figure 3.9 est un exemple de ce cas d'incertitude dans lequel la séquence T2-w présente un hypo-signal notable, la séquence dynamique montre un réhaussement précoce mais l'ADC ne présente pas de plage sombre significative. Cette cible suspecte se révèle être un foyer bénin de néoplasie intraépithéliale de la prostate (PIN).

Il arrive également que certaines cibles présentent toutes les caractéristiques du cancer sus-citées mais se révèlent être des faux positifs. Par exemple, la figure 3.10 illustre le cas d'une plage en hypo-signal significatif sur le T2-w et la cartographie ADC associée à un réhaussement précoce sur la séquence dynamique correspondant finalement à un tissu prostatique sain.

A ce jour, il n'existe pas encore d'abaque précis permettant de distinguer avec certitude une tumeur sur l'IRM.

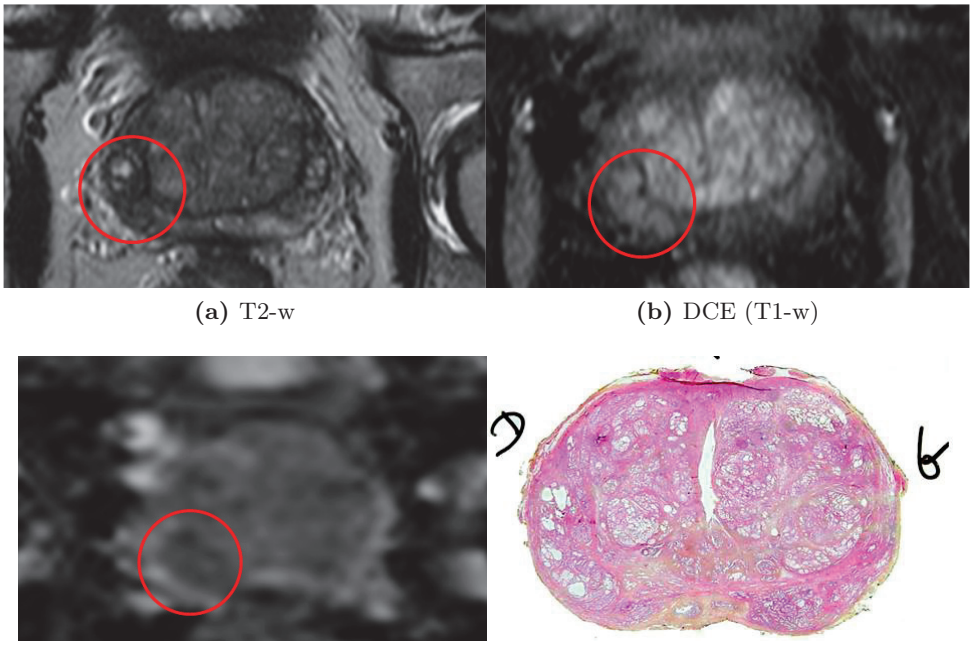
Comme nous le verrons dans le chapitre 4 suivant, les travaux de recherche actuels s'orientent donc vers l'établissement de systèmes d'aide au diagnostic pour assister le radiologue dans l'analyse des images IRM-mp.



(a) T2-w **(b)** DCE (T1-w)

(c) ADC **(d)** histologie

Figure *3.9* – *Cas clinique. IRM multi-paramétrique de prostate : coupes axiales dans le plan médian correspondant aux séquences (a) T2-w, (b) T1-w après injection de Gd-DOTA et (c) ADC ainsi que la coupe histologique correspondante, obtenus sur un patient âgé de 60 ans. Deux zones suspectes en ZP postérieure : à gauche de l'image, un signal en hypo-signal T2-w avec une restriction de la diffusion et un réhaussement localisé en dynamique ; à droite un signal suspect en T2-w et en dynamique, indifférent sur la carte ADC. Ces deux zones sont respectivement un adénocarcinome d'aspect typique de Gleason 7(3+4) et un faux positifs correspondant à un foyer de néoplasie intraépithéliale de la prostate (PIN).*



(a) T2-w **(b)** DCE (T1-w)

(c) ADC **(d)** histologie

Figure *3.10* – *Cas clinique. IRM multi-paramétrique de prostate : coupes axiales dans le plan médian correspondant aux séquences (a) T2-w, (b) T1-w après injection de Gd-DOTA et (c) ADC ainsi que la coupe histologique correspondante, obtenus sur un patient âgé de 69 ans. La zone suspecte entourée en rouge est un faux positif correspondant à un adénome (bénin) de la ZP.*

3.9 Conclusion

La détection, la localisation et l'appréciation de l'agressivité des foyers tumoraux dans la prostate sont devenues un enjeu clinique majeur, non seulement pour permettre un diagnostic plus précoce du cancer, mais aussi pour assurer une prise en charge optimale en diminuant les risques de sur- et de sous-traitement. L'IRM, à condition d'associer au moins l'imagerie T2, dynamique et de diffusion se positionne comme la modalité de choix pour le diagnostic du cancer. Il reste néanmoins le problème, majeur, de la difficulté d'interprétation diagnostique lié à la nature parfois contradictoire de l'information extraite des différentes séquences.

En outre, beaucoup d'urologues persistent à penser que l'IRM-mp n'est pas encore prête pour une utilisation de routine et que les résultats prometteurs présentés dans la littérature ne le sont que pour les centres de référence spécialisés, mais non reproductibles dans d'autres centres [49].

L'hypothèse que nous avons formulée avec les médecins est que les systèmes experts d'aide à la décision peuvent peut-être apporter une aide aux radiologues pour leur formation, mais aussi dans leur pratique clinique quotidienne.

Chapitre 4

Les systèmes d'aide au diagnostic pour l'imagerie du cancer

4.1 Introduction

Le diagnostic assisté par ordinateur (CAD) est devenu l'un des sujets de recherche les plus prolifiques en imagerie médicale. Les systèmes CAD sont des algorithmes permettant d'assister le praticien dans le diagnostic, la détection des zones pathologiques ou le classement des images médicales.

Un système CAD peut être assimilé à un observateur numérique analysant l'image et donnant le même type de décision qu'un clinicien. Cette décision numérique n'a pas vocation à remplacer l'avis du radiologue, mais à le compléter. Ainsi, ce second avis peut permettre au clinicien de revenir sur son appréciation initiale, en la confirmant ou en l'infirmant, ainsi que d'attirer son attention sur d'autres localisations potentiellement pathologiques. Ce domaine de recherche est interdisciplinaire et combine des éléments d'intelligence artificielle, de traitement d'images, et de connaissances biologiques et médicales.

Nous nous concentrons dans ce chapitre sur les systèmes CAD dédiés à l'imagerie du cancer. Le développement des systèmes CAD pour l'imagerie du cancer a débuté dans les années 1980 avec, notamment, l'aide à la détection des micro-calcifications en mammographie et des nodules pulmonaires [17, 41]. Les études [28, 35, 37, 96, 111, 120] présentent un état de l'art (non exhaustif) d'une grande partie des méthodes CAD existantes. Si l'apport des CAD n'est pas démontré pour toutes les pathologies, la majorité de ces études soulignent que ces systèmes permettent une meilleure prise en charge des

patients, en augmentant les performances en termes de diagnostic, de détection et de stadification et en diminuant les variations inter- et intra-cliniciens. Certaines méthodes ont d'ailleurs déjà donné lieu à des systèmes commerciaux. On citera en exemple Image Checker de R2 technologies (http ://www.r2tech.com), iCAD (http ://www.icadmed.com), Kodak mammography CAD engine (http ://www.kodak.com) et syngo mammoCAD de Siemens qui assurent l'aide au diagnostic en mammographie. Le système syngo CT de Siemens comporte un ensemble de modules dédiés au cancer du poumon ou du colon (http ://www.medical.siemens.com) pour l'imagerie TDM (tomodensitométrie), tandis que Philips utilise les programmes xLNA (www.medical.philips.com) pour l'imagerie TDM.

Les quelques méthodes de type CAD récemment proposées en imagerie IRM de la prostate n'ont pas encore fait l'objet d'un transfert industriel. Pourtant, les principaux constructeurs (Siemens, Philips . . .) sont associés à ces travaux par le biais de collaborations avec des équipes académiques. Comme souligné dans le chapitre 2, l'enjeu est majeur : améliorer la discrimination des foyers cancéreux sur les images pourrait permettre une prise en charge focalisée, par HIFU par exemple, du cancer et ainsi proposer une alternative à la prostatectomie radicale. Nous revenons plus en détails sur cet état de l'art dans la section 4.7.

Dans ce chapitre, nous présentons la méthodologie générale classiquement suivie pour l'élaboration des systèmes CAD. Nous proposons un rapide tour d'horizon des principaux algorithmes de classification sur lesquels reposent ces systèmes. Les règles permettant l'évaluation et la comparaison de ces systèmes sont données. Enfin, nous présentons un état de l'art des schémas CAD récemment développés en imagerie IRM de la prostate.

4.2 Systèmes d'aide à la décision (CADx) versus systèmes d'aide à la détection (CADe)

Deux types de systèmes CAD, aux objectifs différents, sont à distinguer dans la suite de ce chapitre : les systèmes d'aide à la détection CADe et les systèmes d'aide à la décision CADx. Les systèmes d'aide à la détection (CADe) se proposent, à partir d'une image ou de plusieurs images recalées du patient, de calculer une carte de probabilité de présence du cancer, combinant la localisation et la classification des anomalies sur une base voxelique, c'est-à-dire pour chaque voxel de l'image. Cette aide à la détection des tumeurs peut notamment être envisagée pour permettre un gain de temps en routine clinique, ou orienter le radiologue vers des régions qui ne l'auraient pas alerté et diminuer ainsi le nombre de faux négatifs (FN).

Les systèmes d'aide à la décision (CADx), quant à eux, retournent un score de suspicion de malignité pour une région d'intérêt (ROI) segmentée par le radiologue, sur laquelle s'interroge l'expert. Un système de type CADx ne réalise donc pas la tâche de détection. L'objectif est d'aider le radiologue à poser un diagnostic correct en proposant un score de malignité objectif et reproductible sur des cibles suspectées (proposer un "deuxième avis").

Bien que ces deux approches CAD n'abordent pas le même problème, elles reposent sur la même méthodologie générale et les mêmes algorithmes de discrimination, présentés ci-après.

4.3 Les systèmes CAD : un schéma méthodologique standardisé

Les systèmes CAD reposent généralement sur un schéma méthodologique standardisé. La figure 4.1 en synthétise les étapes majeures :

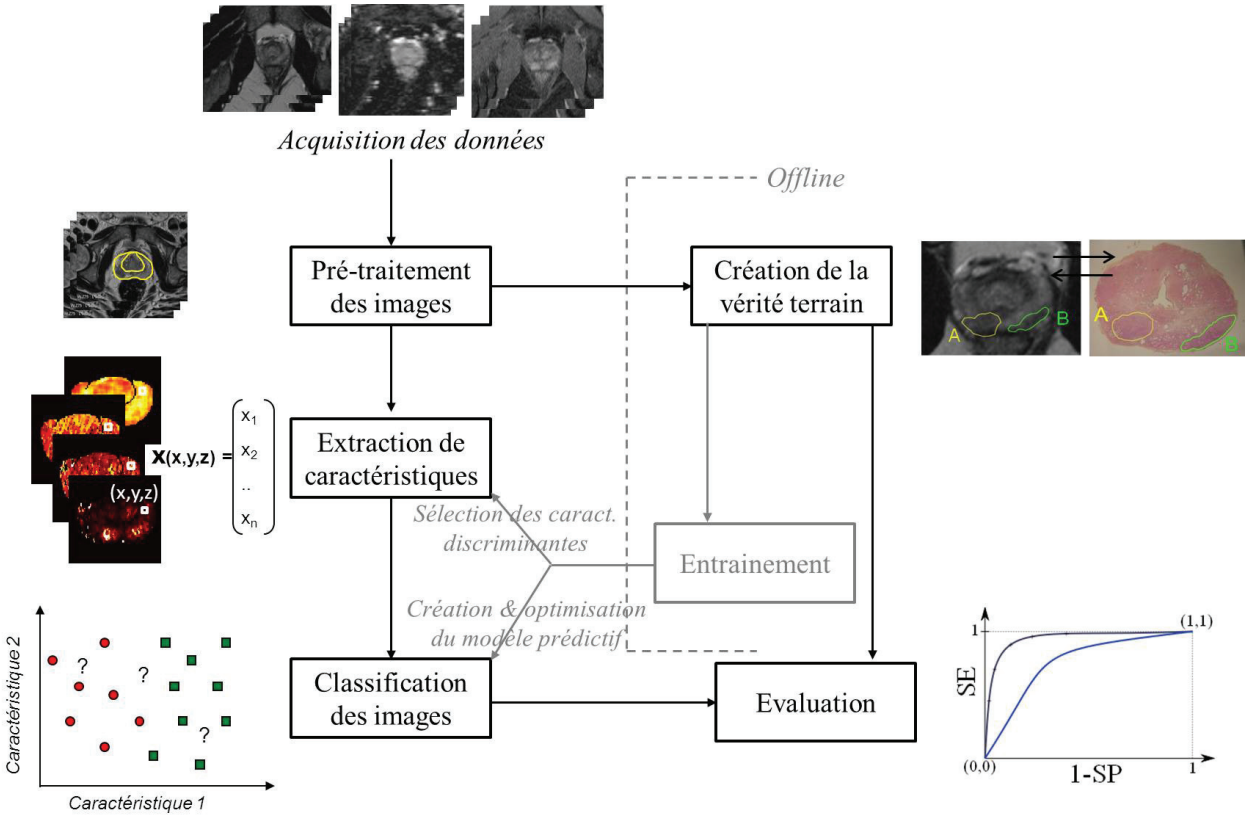


Figure 4.1 – *Les étapes classiques dans la construction et le fonctionnement d'un système CAD*

Pré-traitement des images L'objectif est de préparer et améliorer l'image initiale. Il peut s'agir de segmenter l'organe étudié. Lors d'une acquisition multi-modale ou multi-séquence, il est parfois nécessaire d'effectuer une étape de recalage des images pour permettre une correspondance voxel à voxel des données. Certaines images brutes nécessitent parfois d'être pré-traitées. Dans le cadre des images IRM par exemple, une correction de l'inhomogénéité de champ doit parfois être appliquée, la carte ADC peut être construite, etc.

Extraction de caractéristiques Un système de classification ne traite que rarement

l'image brute seule. Des caractéristiques descriptives de l'image ou du cas clinique (patient) sont généralement exploitées. Ces caractéristiques descriptives représentent l'interprétation numérique d'une image. Elles doivent notamment exprimer ce que l'œil humain analyse avant de poser son diagnostic sur la présence ou non d'une tumeur. Elles peuvent être de plusieurs types : statistiques, fréquentielles, géométriques, cliniques, sémantiques etc.

Classification des images L'objectif est d'attribuer, à une image ou à une région d'intérêt (ROI) par exemple, une classe ou une probabilité d'appartenance à une classe (sain/pathologique, par exemple). La section 4.4 est consacrée à la comparaison de deux grandes familles de classification : la classification supervisée et la classification non-supervisée.

Création de la vérité terrain Afin de pouvoir évaluer le système de classification et d'entraîner le modèle prédictif si on considère une méthode par apprentissage (on construit alors la règle de décision et on optimise les paramètres), il est nécessaire de pouvoir comparer la sortie proposée par le classifieur avec la vérité terrain. Cette dernière peut être une vérité expert, dépendant uniquement de l'analyse réalisée par un radiologue expérimenté qui localise les cancers sur les images ou une vérité histologique (analyse des résultats de biopsies et de coupes histologiques à la suite d'une chirurgie) sûre mais qui nécessite une mise en correspondance avec les images radiologiques acquises.

Évaluation Elle consiste à définir les critères qualitatifs et/ou quantitatifs de mesure des performances diagnostiques du système CAD et le protocole d'évaluation. La section 4.5 présente quelques notions et outils utilisés pour évaluer et comparer les performances de différents schémas CAD.

Entrainement Cette étape d'entraînement (ou apprentissage) est nécessaire pour les classifieurs de type supervisés qui apprennent les caractéristiques des tissus sur une base de données d'entraînement associant à chaque image analysée la vérité terrain correspondante. Cette étape n'entre pas en jeu dans le cadre des méthodes dites non-supervisées. La base d'apprentissage peut également servir à sélectionner, parmi toutes les caractéristiques descriptives extraites, celles qui sont réellement discriminantes dans la tâche de classification.

Sélection des caractéristiques L'objectif de cette étape est d'éliminer les caractéristiques non-informatives pour ne garder que les caractéristiques réellement significatives dans la tâche de discrimination des tissus sains et malins. Dans la littérature, on peut trouver deux types de méthodes de sélection de caractéristiques :

de type *filtre* ou *a priori* : la sélection des caractéristiques est indépendante du classifieur et se fait en amont (ex. : les méthodes de sélection basées sur l'information mutuelle).

de type *enveloppante* ou *wrapper* : la sélection des caractéristiques est intégrée dans la phase d'apprentissage du classifieur (ex. : les méthodes de sélection basées sur

des algorithmes génétiques).

Comme nous le verrons section 4.7 (page 69), la sélection des caractéristiques est une problématique peu traitée dans la littérature des systèmes CAD de prostate.

4.4 Classification supervisée ou non supervisée ?

Un algorithme de classification vise à "regrouper" des objets en sous-ensembles partageant des caractéristiques communes, c'est-à-dire appartenant à la même "classe". Il existe deux principales techniques de classification automatique : la classification supervisée et la classification non supervisée. On parle de classification supervisée lorsqu'on utilise un ensemble de points exemples étiquetés selon leur classe d'appartenance pour apprendre à classifier de nouveaux cas. L'objet des systèmes CAD étant essentiellement de discriminer les cas sains des cas pathologiques, nous nous intéressons principalement aux systèmes de classification en deux classes (binaires). Néanmoins, de nombreux systèmes de classification multi-classes ont été proposés dans la littérature. Dans la suite, nous présentons les principaux algorithmes de classification supervisée et non-supervisée proposés dans la littérature. Il ne s'agit pas de faire une présentation exhaustive de toutes les méthodes mais seulement de préciser les techniques les plus étudiées, que nous seront amené à utiliser et comparer dans le cadre de notre travail en fonction de leurs propriétés particulières.

4.4.1 Méthodes non supervisées

Un algorithme de classification non-supervisée (parfois aussi appelé 'clustering') vise à partitionner un groupe hétérogène d'objets en sous-groupes (ou classes) bien différenciés d'objets "similaires". Ce découpage s'effectue sans informations *a priori* sur la classe (ex. : sain/pathologique) des objets ou d'un échantillon d'objets (voir figure 4.2).

Notons que le nombre de classes peut être un des paramètres d'entrée de l'algorithme (notamment pour les k-moyennes), ou être déterminé de manière automatique.

Le partitionnement repose généralement sur une analyse statistique de la répartition des données et sur la définition d'une fonction de proximité entre individus adaptée (toutes les observations d'une même classe devront être proches au sens de cette fonction). Le partitionnement idéal est obtenu lorsque 1) la dispersion de chaque classe est minimisée (ex. covariance intra-classe minimisée), pour que chaque classe soit la plus homogène possible, et lorsque 2) la distance inter-classe est maximisée, pour que les classes soient les plus distinctes possibles.

Différents types d'algorithmes de classification non-supervisée sont décrits dans la littérature. On peut notamment citer, parmi les algorithmes les plus utilisés, le regroupement ascendant hiérarchique, la classification divisive et la classification avec un nombre fixe de classes telle que les k-moyennes (généralisées par les nuées dynamiques) ou les cartes auto-adaptatives :

Regroupement ascendant hiérarchique (ou 'par agrégation') : cet algorithme procède par fusions successives de sous-groupes (appelés *clusters*) déjà existants. A chaque étape, les deux clusters qui vont fusionner sont ceux dont la "distance" est la plus faible. A la première itération, toutes les observations sont des clusters ne contenant qu'une seule observation (une observation = une classe). La première étape consiste donc à réunir dans un cluster à deux observations les deux observations les plus proches. Puis l'algorithme continue, fusionnant à chaque étape les deux clusters les plus proches au sens de la distance choisie. Le processus s'arrête quand les deux clusters restants fusionnent dans l'unique cluster contenant toutes les observations [135].

Classification hiérarchique descendante (ou divisive) : cet algorithme procède de façon inverse du précédent. Il considère l'ensemble des données comme un gros cluster unique, et le scinde en deux clusters "descendants". La scission s'opère de façon à ce que la distance entre les deux descendants soit la plus grande possible, afin de créer deux clusters bien séparés. Cette procédure est ensuite appliquée à chacun des descendants (procédure récursive) jusqu'à ce qu'il ne reste plus que des clusters ne contenant qu'une seule observation (singletons).

k-moyennes (k-means) [66] : cet algorithme itératif associe à chaque point la classe dont le barycentre est le plus proche, puis remet à jour le barycentre des classes à l'itération suivante. Les observations sont donc divisées en k partitions dans lesquelles chaque observation appartient à la partition avec la moyenne la plus proche. Le nombre de classes doit être fixé *a priori* par l'utilisateur.

Cartes auto-adaptatives [56] : cet algorithme, type très particulier de réseaux de neurones, utilise des techniques dérivées des graphes pour partitionner les données. Les cartes de Kohonen reposent sur une projection optimale de l'espace des observations dans un espace de dimension inférieure tout en conservant leur positionnement relatif dans l'espace des données ("respect de la topologie").

Champs de Markov aléatoires : dans la modélisation markovienne, l'image est considérée comme une réalisation d'un champ aléatoire. Cet algorithme est fondé sur une modélisation statistique conjointe des régions et des niveaux de gris et repose sur la minimisation d'une fonction de vraisemblance (ou énergie). Cette fonction prend simultanément en compte la vraisemblance de l'appartenance du pixel à une région considérant son niveau de gris, et les régions auxquelles appartiennent les pixels voisins. Elle réalise un compromis entre la fidélité à l'image initiale et la régularité des régions segmentées. Cette méthode se distingue des méthodes classiques par la prise en compte des interactions locales entre chaque pixel et ses pixels voisins pour définir les différentes régions de l'image. Notons aussi que des techniques de classification supervisée utilisant des champs de Markov aléatoires ont aussi été développées.

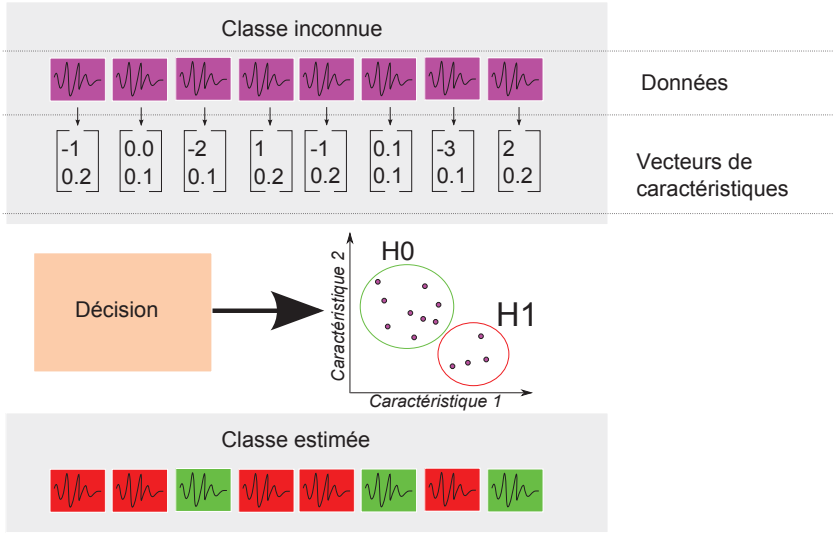


Figure 4.2 – *Fonctionnement d'un classifieur non supervisé : les données brutes sont envoyées au classifieur qui va les regrouper en classes en fonction de leur répartition dans l'espace des caractéristiques.*

4.4.2 Méthodes supervisées

Contrairement à la classification non-supervisée, la classification supervisée vise à classer des observations à partir de connaissances acquises *a priori* sur un ensemble de données d'apprentissage.

On entraine le classifieur en lui fournissant des exemples d'apprentissage étiquetés c'est-à-dire pour lesquels la classe d'appartenance (ex. : statut sain/pathologique) est connue.

A partir de cette base de données d'entraînement, le classifieur va générer un *modèle* prédictif permettant de classer de futurs exemples non encore connus, comme présenté dans la figure 4.3. Par essence, l'inconvénient principal de la classification supervisée repose sur la nécessité d'avoir une base d'observations suffisamment large et représentative pour l'apprentissage du classifieur.

De nombreux classifieurs supervisés ont été développés. Nous avons choisi de présenter ici certains des algorithmes les plus cités dans la littérature relative aux systèmes CAD : le séparateur à vaste marge (SVM), l'analyse discriminante linéaire (ADL), le classifieur naïf de Bayes (CNB), les k-plus proches voisins (k-PPV), les réseaux de neurones, et les arbres (étendus aux forêts RDF) de décision. Notons qu'une description plus détaillée des 4 premières méthodes, exploitées dans ce travail de thèse, est donnée dans les parties II et III :

Le séparateur à vaste marge (SVM) [128] : à partir d'un ensemble de données d'apprentissage, l'algorithme du SVM consiste à trouver l'hyperplan séparateur optimal qui maximise la distance (la "marge") entre l'hyperplan et les données des deux classes (lire chapitre 11, page 159).

L'analyse discriminante linéaire (ADL) [36, 50] : l'ADL consiste à déterminer des

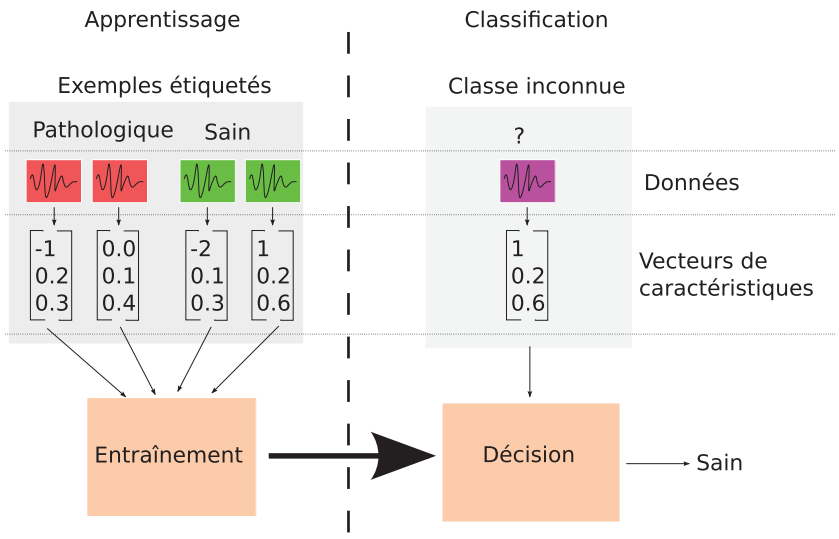


Figure *4.3* – *Fonctionnement d'un classifieur supervisé : Les données d'apprentissage servent à entraîner le classifieur pour générer un modèle. Ce modèle permettra de rattacher des observations aux classes apprises.*

facteurs discriminants, combinaisons linéaires des variables descriptives d'origine, qui prennent des valeurs les plus proches possibles pour des éléments de la même classe, et les plus éloignées possibles entre éléments de classes différentes. Mathématiquement, l'ADL se formule comme la recherche de la meilleure projection des données de manière à avoir une variance intra-classe minimale et une variance inter-classe maximale (lire chapitre 7.4.4, page 117).

Le classifieur naïf de Bayes (CNB) : c'est un classifieur probabiliste simple, basé sur l'application du théorème de Bayes. La décision Bayésienne [9] établit que la stratégie qui consiste à affecter une nouvelle observation à la classe ayant la plus grande probabilité a posteriori est optimale. La probabilité a posteriori correspond à la probabilité pour qu'une observation appartienne à une certaine classe connaissant son vecteur de caractéristiques. Ce classifieur est dit naïf dans le sens où il repose sur une hypothèse forte d'indépendance entre les caractéristiques des données. Bien qu'elle soit basée sur une hypothèse fausse en général (les attributs sont rarement indépendants), elle donne cependant de bons résultats dans les problèmes réels.

Les k-plus proches voisins (k-PPV) [26] : cette méthode assez simple repose sur la comparaison de toute nouvelle observation aux observations d'apprentissage. La méthode consiste à sélectionner les k plus proches voisins de la base d'apprentissage, au sens de la métrique considérée et à associer à l'observation la classe majoritaire de ses k voisins. Les principaux inconvénients de cette méthode sont le nombre d'opérations nécessaires pour classer une entité dans le cas d'une grande base de référence ainsi que sa sensibilité au bruit présent dans les données d'apprentissage.

Les réseaux de neurones [48] : Le réseau de neurones est un classifieur non linéaire

visant à imiter le fonctionnement des neurones du cerveau humain. Un réseau de neurones est constitué de neurones artificiels interconnectés et fonctionnant en parallèle qui miment les réactions des neurones biologiques. Un neurone permet de définir une fonction discriminante (généralement non-linéaire) dans l'espace d'entrée à partir d'une combinaison linéaire (coefficients = poids, et ajout d'une constante de biais) de son vecteur de caractéristiques. Chaque élément du vecteur de caractéristiques entré dans un neurone ira stimuler un ou plusieurs autres neurones (propagation de l'information). L'ajustement des poids (et du biais) se fait par comparaison entre la réponse du réseau (ou sortie) et la cible, jusqu'à ce que la sortie corresponde au mieux à la cible. La règle d'arrêt de l'apprentissage repose sur la minimisation du risque empirique (sans postulat sur les densités de probabilité donc). Les réseaux de neurones sont très performants, mais souffrent de difficultés de mise en place à cause de leur grand nombre de paramètres et d'un temps d'apprentissage très importants. De plus, ils présentent certains risques de sur-apprentissage, modélisant non seulement le concept mais également le bruit de fond qui l'accompagne. Enfin, l'interprétation synthétique des estimateurs qu'ils délivrent est rendue malaisée par leur structure interne complexe.

Les arbres/forêts de décision : les forêts aléatoires de décision (RDF pour *random decision forest*) sont constituées de plusieurs arbres aléatoires de décision qui peuvent voter chacun pour une classe. La classe ayant reçu le plus de vote constitue alors la réponse de la forêt. Partant d'une base d'apprentissage de N exemples, chaque arbre est entraîné à partir d'un tirage aléatoire avec remise de N exemples issus de la base d'apprentissage (c'est ce que l'on appelle le bootstrap). On construit un arbre en tirant aléatoirement un sous-ensemble de k caractéristiques parmi toutes les caractéristiques disponibles ainsi que k seuils associés. Ces couples (caractéristique, seuil) séparent les exemples en deux catégories : ceux dont la caractéristique est inférieure au seuil et les autres. On évalue alors les k couples afin de déterminer celui qui sépare le mieux les exemples : le meilleur couple constitue alors un nœud de l'arbre. On applique récursivement ces étapes de tirage de couples et choix aux fils gauches et droits du nœud. La récursion s'arrête lorsqu'elle dépasse une profondeur maximale fixée ou que l'ensemble à séparer pour un nœud est un singleton. On construit alors une feuille qui contient la classe qui a la probabilité la plus forte. Les forêts aléatoires ont pour avantage d'être simples à mettre en place et permettent de sélectionner des variables discriminantes automatiquement pour des problèmes contenant un grand nombre de caractéristiques. Un inconvénient de ces méthodes est qu'elles dépendent largement de la graine utilisée et du nombre d'arbres, fixé a priori (ou par *grid search*). De plus l'interprétabilité et les capacités d'analyse offertes par les classifieurs de type arbres de décisions sont perdues, du fait de l'utilisation de principes de "randomisation" au cours de leur induction.

4.5 Évaluation des performances des systèmes CAD

Comme nous pourrons le constater dans la section 4.7, il est très difficile de comparer les performances des systèmes CAD. Les études ne précisent en effet pas toujours toutes les règles d'évaluation utilisées, nécessaires à une interprétation correcte des performances. La première consiste à définir le comptage des vrais positifs (VP) et faux positifs (FP) décrits au paragraphe 4.5.1.

Plus largement, il faut également fixer une stratégie d'évaluation des performances en sensibilité et spécificité (lire paragraphe 4.5.1). Un certain nombre d'outils basés sur la méthodologie ROC, initialement prédéfinis pour l'évaluation de la détectabilité dans des études d'observateurs humains, sont généralement utilisés et adaptés pour la problématique de la classification.

Enfin, une règle propre aux systèmes CAD supervisés consiste à définir le partitionnement des images en base d'apprentissage et base de test. Les images d'apprentissage servent à estimer les paramètres optimaux du classifieur et les images de test permettent de calculer la proportion finale des erreurs de classification. La méthode de partitionnement peut en effet influencer les performances de manière pessimiste ou optimiste. Les paragraphes suivants présentent un bref état de l'art de l'ensemble de ces règles afin d'introduire la méthodologie choisie pour l'évaluation du système CAD développé dans cette thèse.

4.5.1 Définitions préliminaires

Notion de vrais/faux positifs (VP/FP), vrais/faux négatifs (VN/FN)

La nature des performances quantitatives d'un système de détection repose sur quatre notions essentielles que sont les vrais positifs (VP), faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN). Un VP quantifie la détection d'une vraie tumeur tandis qu'un FP correspond à la détection d'une tumeur qui n'en est pas réellement une. De même, les FN qualifient les vraies tumeurs non détectées par le système et les VN correspondent aux régions non-tumorales bien classifiées par le système.

Remarque sur le comptage des vrais positifs et faux positifs

Différents modes de comptage des VP et FP sont proposés dans la littérature. L'évaluation du nombre de VP/FP etc pour les schémas de type CADx est moins sujette à discussion puisque la sortie du CAD est généralement un score de malignité (ou parfois directement une valeur binaire) sur des tumeurs (ROI) pré-détectées : la comparaison à la vérité terrain peut se faire deux à deux. Il ne s'agit pas d'une détection initiale de tumeurs (CADe).

En revanche, la sortie d'un système CADe est généralement une carte paramétrique (binaire ou pas) pour laquelle un voxel ou groupe de voxels reçoit un score caractéristique de sa classe d'appartenance (tumeur/non tumeur). Il s'agit alors de déterminer si un voxel

ou un groupe de voxels étiqueté "tumeur" est bien un VP.
Certaines études [86] proposent de définir une distance d'acceptation entre la position trouvée et celle de la vérité terrain, en dessous de laquelle une tumeur suspecte est considérée comme correctement détectée. Cette distance est souvent calculée entre les centres des deux positions. D'autres études [5, 63] proposent plutôt de définir une mesure de recouvrement des surfaces de la région détectée positive et de la vérité terrain, l'indice de Dice (DSC) par exemple :

soit A et B deux surfaces à comparer, l'indice de Dice est défini par

$$\mathrm{DSC}(A,B) = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid} \tag{4.1}$$

où $\mid . \mid$ désigne l'aire de la surface (nombre de voxels pour les images numériques).

La méthode de comptage des tumeurs suspectes étant réellement des tumeurs (VP) ou des faux positifs (FP) peut être appliquée par image (= patient) ou normalisée par coupe (2 dimensions (2-D)). Les performances peuvent fortement différer selon un calcul par image ou par coupe. Lorsque le comptage est réalisé par coupe, le nombre réel de candidats FP visibles sur l'image n'est pas reflété. En effet, une coupe contenant un nombre N de fausses détections ($N \geq 1$) n'incrémente le compteur total des FP que d'une unité au lieu de N.

De même, l'unité d'évaluation du nombre de VP et de FP peut être une ROI cible (groupe de voxels) ou un voxel. L'équilibre (ou plutôt le déséquilibre) entre le nombre de cas sains et pathologiques dans l'image peut également influer sur les performances mesurées. Ces divergences de choix rendent la comparaison entre différentes études difficile.

Notion de sensibilité (SE) / spécificité (SP)

Une matrice de confusion, présentée dans le Tableau 4.1, propose une définition simple des mesures de la qualité d'un système CAD qui sont comprises entre 0 et 1 (ou 0 et 100%). Plusieurs valeurs peuvent être directement extraites de cette matrice de confusion

Table *4.1* – *Exemple de matrice confusion permettant de définir les notions de VP, FP, VN et FN :*

		Classe réelle		
		Pathologique	Sain	
Classe estimée	Pathologique	VP	FP	↔ VPP
	Sain	FN	VN	↔ VPN
		↕ SE	↕ SP	

de manière à quantifier les différentes capacités du CAD. C'est le cas notamment des mesures de sensibilité, spécificité, précision (Acc, pour *Accuracy*), valeur prédictive positive

(VPP) et valeur prédictive négative (VPN) :

$$\text{Sensibilité} = \frac{VP}{VP+FN}, \tag{4.2a}$$

$$\text{Spécificité} = \frac{VN}{VN+FP}, \tag{4.2b}$$

$$\text{Précision} = \frac{VP+VN}{VP+FP+VN+FN}, \tag{4.2c}$$

$$\text{VPP} = \frac{VP}{VP+FP}, \tag{4.2d}$$

$$\text{VPN} = \frac{VN}{FN+VN} \tag{4.2e}$$

La sensibilité (SE) correspond à la proportion de cas correctement évalués pathologiques par rapport au nombre total de cas réellement pathologiques (elle est également appelée fraction de vrais positifs). Elle permet d'évaluer la capacité du CAD à détecter les vraies tumeurs.

La spécificité (SP) correspond à la proportion de cas correctement évalués comme sains par rapport au nombre total de cas réellement sains (elle est également appelée fraction de faux positifs). Elle permet donc d'évaluer la capacité à ne pas détecter de tumeurs là où il n'y en a pas.

Le couple (sensibilité, spécificité) permet de caractériser les performances d'un test diagnostique binaire. Ces deux grandeurs sont complémentaires mais ne permettent pas à elles seules de comparer les performances de différents observateurs. En effet, un lecteur va souvent donner des notes, qui vont indiquer son niveau de confiance sur la présence de la pathologie [27] (à ne pas confondre avec des notations sur la gravité des lésions, comme les techniques de gradation de Gleason [43]).

Les techniques de comparaison de systèmes de décision comme l'analyse ROC (Receiver-Operating Curve), que nous détaillons dans la section 4.5.2 suivante permettent de prendre en compte ces incertitudes. Les courbes ROC servent à évaluer la capacité d'un ou plusieurs "observateurs" à discriminer des signaux entre deux classes "normale" et "anormale". Les informations de sensibilité et de spécificité se limitent à comparer les performances pour un niveau de confiance donné.

4.5.2 Méthodologie ROC - Receiver Operating Characteristic

Principe

La sensibilité et la spécificité d'un observateur (qu'il soit humain ou numérique), définis précédemment, mesurent son efficacité de discrimination entre deux classes à partir d'une réponse binaire (ici : sain/pathologique).

Les courbes de type ROC (Receiver Operating Characteristic) [72, 121] exploitent ces mesures en représentant graphiquement la spécificité et la sensibilité de l'observateur pour

différents niveaux de certitude. En effet, dans la majorité des études réalisées, l'observateur ne renvoie pas directement une réponse binaire mais plutôt un score de suspicion de malignité traduisant la difficulté du diagnostic. L'observateur peut être humain ou numérique (un algorithme de classification), et le score discret ou continu.

Ainsi, différentes échelles de notation (discrètes) adressées aux radiologues ont été proposés dans la littérature. Dans le cadre du diagnostic du cancer de la prostate, on notera par exemple les systèmes de scores proposés par Dickinson *et coll.* [27] (barème entre 0 et 4 avec : 0-définitivement bénin, 1- potentiellement normal, 2-équivoque, 3-potentiellement pathologique, 4-définitivement pathologique) ou plus récemment par Barentsz *et coll.* [6].

De même, la majorité des algorithmes de classification présentés dans la section 4.4 renvoient en sortie un score continu traduisant par exemple une distance entre classes ou une probabilité d'appartenance à une classe.

Par convention, plus le score (noté λ dans la suite) est élevé, plus l'observateur considère qu'il est en présence d'un cas pathologique. A l'inverse, une note basse indiquera un cas présumé sain ou normal.

Les courbes ROC, en prenant en compte ces valeurs de score (sans passer par l'utilisation d'une valeur de seuil arbitraire pour *binariser* la réponse) fournissent une mesure objective des performances d'un observateur dans une tâche de discrimination entre deux classes, prenant en compte son degré de certitude.

Cette méthode d'évaluation est très souvent utilisée comme critère pour permettre les comparaisons inter-observateurs ou fixer les différents paramètres des algorithmes.

Mise en oeuvre

L'évaluation d'un observateur par la méthode ROC nécessite la création d'un jeu de données étiquetées selon deux classes H0/H1, où H0 correspond aux cas négatifs (sains) et H1 au cas positifs (pathologiques).

Le tracé de la courbe ROC se fait en reportant la sensibilité et la différence "1 - spécificité" de l'observateur pour différentes valeurs de seuil de décision (ou niveau de certitude), au-delà duquel les observations sont considérées comme pathologiques (les exemples positifs au-dessus du seuil sont des vrais positifs, les exemples négatifs au-dessus du seuil sont des faux positifs etc).

Ce seuil permet de modifier de manière dynamique la répartition des observations dans la matrice de confusion (voir tableau 4.1). Cela permet d'enrichir la comparaison des observateurs par rapport à un couple (sensibilité/spécificité) seul. Par construction, la courbe va commencer au point $(0, 0)$ (tous les points sont diagnostiqués négatifs pour le seuil λ_{max}) et se terminer au point de coordonnée $(1, 1)$ (tous les points sont diagnostiqués positifs pour le seuil λ_{min}).

Différents exemples de courbes ROC sont présentés sur la figure 4.4.

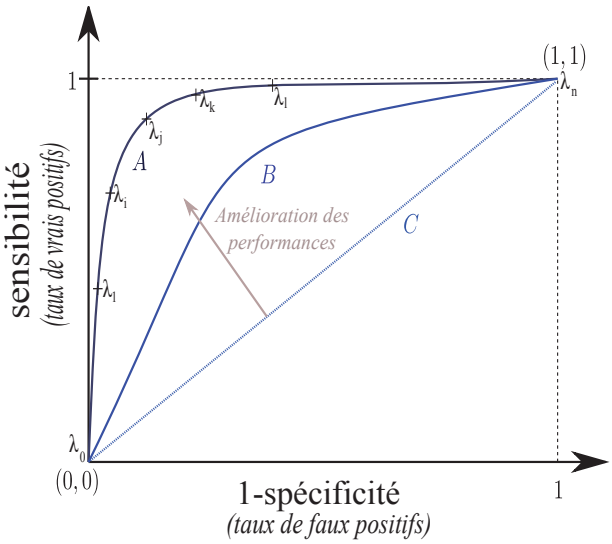


Figure 4.4 – *Exemples de courbes ROC.*
L'échelon constitue la courbe ROC idéale (toutes les détections sont justes, pas de faux positifs, AUC = 1).
La courbe A représente le résultat d'une évaluation ROC avec un barème à 6 niveaux (λ_i, $i = 1 \dots 6$). Pour chaque note du barème, le point correspondant est affiché en reportant la spécificité et la sensibilité (représentés par les croix).
La courbe B représente une autre évaluation, avec des performances inférieures : pour chaque niveau de "1-spécificité", la sensibilité de la courbe B est inférieure à celle de la courbe A.
La première bissectrice C constitue la droite de "chance" : elle représente le cas d'un observateur qui donne ses réponses de manière aléatoire sur une base de données équilibrée. L'aire sous cette courbe (AUC) vaut 0.5.
Les courbes passant en dessous de la courbe de chance utilisent mal les données (AUC < 0.5). En effet il leur suffit d'inverser leur pronostic ($f_{new}(i) = 1 - f(i)$) afin de remonter au-dessus de la courbe de chance.

4.5.3 Comparaison de courbes ROC

Un ensemble d'indicateurs permet de quantifier et de comparer les performances de détection à partir des courbes ROC.

Le plus simple consiste à choisir un niveau de spécificité et à comparer les sensibilités des différents observateurs. L'avantage de ce système est qu'il permet de comparer les performances dans des conditions proches de la réalité, où l'on cherche à rester dans un taux de spécificité α donné. Cependant, les résultats vont dépendre du paramètre α.

Une autre méthode consiste à trouver le point optimum de la courbe considérée, celui le plus proche du point de performance idéale (0,1) correspondant à des sensibilité et spécificité de 100%.

Une mesure plus globale est l'aire sous la courbe ROC (notée AUC). C'est un indicateur synthétique de performance. Elle représente la probabilité d'identifier correctement le cas pathologique quand un cas pathologique et un cas sain sont présentés simultanément à

l'observateur.

Étant donné que la courbe est nécessairement comprise dans un carré unitaire, la valeur de l'aire sera comprise entre 0 (l'observateur donne systématiquement les mauvaises réponses) et 1 (l'observateur donne toujours la bonne réponse).

Modélisation et analyse statistique

Pour comparer deux courbes ROC et estimer s'il existe une différence significative (au sens statistique) entre elles, différentes méthodes d'estimation de l'aire sous la courbe ROC (AUC) et des intervalles de confiance associés ont été proposées.

Parmi les méthodes d'analyse non-paramétrique [46, 64] (i.e. ne reposant pas sur un modèle statistique sous-jacent), la plus simple consiste à estimer l'AUC par la méthode des trapèzes. Une approche classique consiste à assimiler l'AUC à la statistique de Wilcoxon-Mann-Whitney, de la manière suivante :

$$AUC = \frac{1}{n_{\mathrm{H1}}.n_{\mathrm{H0}}} \times \sum_{\substack{i\in\{\mathrm{H1}\}\\ j\in\{\mathrm{H0}\}}} \psi(\lambda_i,\lambda_j),\ \text{où}: \psi(\lambda_i,\lambda_j) = \begin{cases} 1 \text{ si } \lambda_i > \lambda_j \\ 0,5 \text{ si } \lambda_i = \lambda_j \\ 0 \text{ si } \lambda_i < \lambda_j. \end{cases} \tag{4.3}$$

et d'en exploiter les propriétés statistiques pour déterminer les intervalles de confiance.

Des méthodes paramétriques [72, 73] ont également été développées et sont largement exploitées pour le tracé et l'analyse des courbes ROC. L'idée est d'ajuster les données mesurées sur une courbe théorique. Les méthodes d'analyse statistique paramétrique font

	Méthode non paramétrique	**Méhode paramétrique**
Avantages	Utilisation de tous les points expérimentaux. Pas d'extrapolation paramétrique des données expérimentales. La courbe passe par tous les points observés. Détermination de l'aire sous la courbe simple. Absence de biais pour la détermination de la sensibilité, de la spécificité et de l'aire sous la courbe.	Tracé "lissé". Comparaison de courbes possibles à toutes les sensibilités et spécificités.
Inconvénients	Tracé en marches d'escaliers. Sujets *ex aequo* source de sous-estimation des performances diagnostiques du test. Construction longue avec les échantillons de taille importante. Comparaison des courbes uniquement aux spécificités et sensibilités observées.	Extrapolation paramétrique des données expérimentales. Biais dans l'estimation de l'aire sous la courbe possible. Détermination de l'aire sous la courbe plus complexe. La courbe ne passe pas nécessairement par les points expérimentaux.

Table *4.2* – *Avantages et inconvénients des courbes ROC non paramétriques et paramétriques d'après Zweig* et coll. *[145]*

généralement l'hypothèse que les distributions de probabilités des scores des cas H0 et H1 suivent des lois gaussiennes (voir figure 4.5). Le modèle de décision suppose que l'ensemble des scores (λ) évaluées sur des cas sains H0 (respectivement sur les cas pathologiques H1) suit une distribution de probabilité $P(\lambda_0, \sigma_0)$ (resp. $P(\lambda_1, \sigma_1)$) de valeur moyenne λ_0 (resp. λ_1) et d'écart-type σ_0 (resp. σ_1). Selon cette hypothèse gaussienne, on peut montrer que la distance d, appelée indice de détectabilité, correspond à l'aire sous la courbe ROC :

$$d = \frac{\lambda_1 - \lambda_0}{\sqrt{\sigma_1 + \sigma_0}}. \tag{4.4}$$

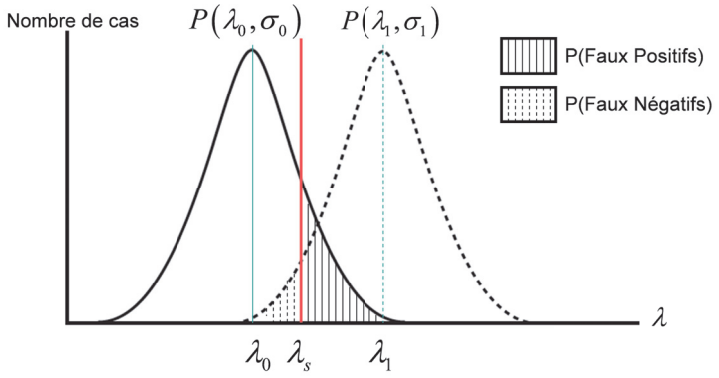


Figure 4.5 – *Modèle de la distribution de probabilité de la variable de décision pour les populations H0 ($P(\lambda_0, \mu_0)$) et H1 ($P(\lambda_1, \mu_1)$) dans les études ROC. La valeur λ_s représente le seuil à partir duquel une observation sera étiquetée H0 ou H1.*

On peut alors effectuer des tests statistiques sur les variables estimées pour évaluer la significativité de la différence entre deux courbes ROC.
D'autres distributions, telles que la distribution logistique, sont aussi employées pour la modélisation des courbes (nous en discuterons notamment au chapitre 9).

Plusieurs logiciels ont d'ores et déjà été développés pour estimer les paramètres des courbes ROC à partir d'approches paramétriques et non-paramétriques (AccuROC, Analyse-It, CMDT, DBM, GraphROC, LabROC, MedCalc, mROC, ROCKIT, and SPSS) ; la plupart d'entre eux sont comparés dans la publication de Carsten *et coll.* [119].
On notera qu'un domaine d'étude actuel concerne le cas des données corrélées, en particulier le cas des données *clusterisés* (plusieurs cibles proviennent d'un même patient). Cette approche a fait l'objet de travaux théoriques [80,106] qui n'ont pas permis à l'heure actuelle de déboucher sur une méthode consensuelle [81]. Nous reviendrons sur cette problématique dans la partie II de cette thèse.

La méthodologie ROC est limitée par le fait que l'observateur ne donne pas d'information de localisation de la pathologie dans l'image. On lui présente une image et il doit la noter sans indiquer la localisation de la zone suspecte. Si nous voulons comparer des observateurs qui réalisent une tâche de détection et de localisation de la zone suspecte, il faut non seulement savoir si des lésions sont présentes, mais aussi avoir leur nombre et leur

localisation. Pour éviter cette limitation, plusieurs extensions à la méthodologie ROC ont été proposées, notamment la méthodologie F-ROC (*Free Response Operating Characteristic*). Nous ne développons pas cette partie de l'analyse psychophysique que nous n'utilisons pas dans le cadre de cette thèse.

4.6 Sélection des bases d'images d'apprentissage et de test

4.6.1 Problématique de la définition de la vérité terrain

L'élaboration d'un schéma CAD supervisé nécessite généralement un grand nombre d'images de référence annotées pour lesquelles le statut carcinogène est connu, c'est ce que nous avons appelé la base d'apprentissage. Cette base d'apprentissage est non seulement utile pour la détermination des modèles de prédiction construits à partir de méthodes de classification supervisée, mais également pour le calcul des performances quantitatives détaillées précédemment (lire section 4.5).

L'annotation des données cliniques par le clinicien expert est un travail long et fastidieux. On distingue deux types de vérité diagnostic :

- le diagnostic réalisé par le radiologue lors de l'analyse des images acquises et qui est donc très dépendant de son niveau d'expertise.
- le statut carcinogène réel des tissus, obtenu après une analyse histologique totale (après chirurgie) ou partielle (après biopsies) de l'organe considéré.

Quelques bases de cas cliniques sont disponibles en imagerie IRM de la prostate, mais celles-ci ne sont pas encore distribuées à la communauté scientifique. Comme nous le verrons dans la section 4.7 dédié à l'état de l'art, les auteurs utilisent généralement des bases d'images limitées correspondant à quelques cas fournis par les hôpitaux partenaires de leur étude. Ce manque de bases d'images cliniques communes est un réel problème pour la recherche en CAD de la prostate puisqu'il rend la comparaison inter-étude difficile et freine l'amélioration des systèmes actuels.

Une alternative à l'utilisation d'images cliniques repose sur la simulation d'examens cliniques. Les images simulées sont obtenues à partir d'un programme numérique ou simulateur, reproduisant virtuellement et le plus fidèlement possible la réalité clinique en termes d'acquisition, de bruit et de modélisation des tumeurs. La simulation permet de remédier aux limitations des images cliniques, notamment l'absence de vérité terrain sur la tumeur et permet de disposer rapidement d'un nombre suffisant d'images pour l'évaluation des systèmes de classification. Si de tels simulateurs IRM existent pour d'autres applications [10], aucun modèle n'a encore été développé pour l'étude du cancer de la prostate. La limitation majeure de ces outils concerne le réalisme des données de synthèse. En IRM, la modélisation complète d'un organe, incorporant le niveau de détail anatomique souhaité et la variabilité des paramètres physiques (T2, ...), constitue un thème de recherche à part entière.

4.6.2 Partitionnement des images disponibles pour l'entraînement du système

Une fois le type d'images défini, il convient de les séparer en différents groupes pour la mise en place d'une stratégie de construction et d'évaluation du CAD. Idéalement, il est souhaitable de séparer en deux groupes le jeu de données à disposition : un partie sert de base d'apprentissage, pour fixer le modèle prédictif, l'autre de base de test, pour en évaluer les performances (capacités de généralisation). Cette méthode est désignée sous le nom de *hold-out*. En effet, la qualité réelle d'un modèle ne peut pas être objectivement estimée par les performances mesurées sur des données qui ont servi à le construire. Néanmoins, dans la pratique, les études reposent sur peu d'exemples (i.e. chaque exemple compte) et cette séparation en deux n'est pas souhaitable, l'évaluation des performances étant alors trop fortement dépendante du partage réalisé.

Plusieurs méthodes de partitionnement (ou 'ré-échantillonnage') des bases d'images permettant de simuler les conditions réelles d'utilisation à partir des données disponibles sont données dans la littérature, parmi lesquelles les méthodes de :

Validation croisée (k-fold). Technique particulière de validation d'une architecture de modèle dans laquelle plusieurs modèles de même architecture sont construits sur k sous-ensembles disjoints des données disponibles. La performance de chacun des modèles est alors estimée sur la partie des données qui n'a pas été utilisée lors de sa construction. Chaque exemple sert à l'apprentissage $k-1$ fois et est testé 1 fois. Les résultats obtenus pour toutes les combinaisons (échantillons d'apprentissage/test) sont ensuite synthétisés pour donner une estimation du pouvoir de généralisation de l'architecture testée.

Leave-One-Out (LOO) / Leave-one-Patient-Out (LOPO). Cas particulier de la validation croisée. Lorsqu'un seul échantillon sert au test du modèle de classification et tous les autres à l'apprentissage, la méthode est appelée leave-one-out. Dans la littérature et plus loin dans cette thèse, on verra souvent utilisé le terme LOPO (pour *leave-one-patient-out*). L'idée est de considérer les données d'un même patient comme un tout indissociable, évitant ainsi d'entraîner et de tester l'algorithme sur des données issues d'un même patient (*clustered data*).

Resubstitution. La méthode de resubstitution, quant à elle, utilise toutes les données disponibles pour l'apprentissage et teste le modèle du classifieur sur chacune de ces mêmes données (les ensembles d'apprentissage et de test ne sont pas disjoints). Cette méthode, simple à mettre en œuvre, est souvent critiquée parce qu'elle ne prend pas en considération le comportement de l'algorithme lorsqu'un cas de figure inconnu est rencontré.

Boostrap. La méthode par bootstrap [33] permet, lorsque la base d'exemples est trop restreinte, de répliquer les images disponibles. Soit N la taille de la base de données initiale, notée $\mathcal{B}_{init}$. Le principe est alors d'effectuer plusieurs tirages avec remise pour constituer de nouveaux échantillons de données de taille N (i.e. un élément de

$\mathcal{B}_{init}$ peut ne pas appartenir à l'ensemble d'apprentissage, ou y figurer plusieurs fois), utilisés comme ensembles d'apprentissage. L'ensemble de test est toujours $\mathcal{B}_{init}$. Les performances obtenues sur chaque échantillon sont alors moyennées. A noter que la méthode du *ordinary bootstrap* consiste à estimer le biais de resubstitution sur des tirages obtenus par bootstrap, que l'on soustrait alors à l'estimation de performance de classification réalisée sur $\mathcal{B}_{init}$.

Des détails théoriques sur ces méthodes et leur comparaison peuvent être trouvés dans l'article de Fukunaga *et coll.* [39]. Comme le soulignent Sahiner *et coll.* [107, 108] et Li *et coll.* [60], le choix du partitionnement doit être pris en compte lors de la comparaison de différentes études puisqu'il peut en effet expliquer des différences de performances entre elles.

Les études de Sahiner *et coll.* [108], Li *et coll.* [60] et Dundar *et coll.* [29] ont démontré que les méthodes de *hold-out* et de *leave-one-out* (LOO) peuvent être considérées comme non-biaisées, même si elles ont plutôt tendance à sur-estimer l'erreur de classification (méthodes pessimistes). Dans le cas de la méthode de LOO (ou k-fold avec k très grand), la répétition de la procédure pour les différents échantillons amène à une erreur moyenne estimée proche de l'erreur théorique du classifieur pour l'ensemble des échantillons disponibles. De plus, elle a démontré une bonne capacité de généralisation. Le point faible du LOO est son coût en temps qui peut s'avérer prohibitif pour de larges bases de données. La méthode de resubstitution a, quant à elle, tendance à sous-estimer fortement l'erreur de classification et à fournir une variance sur la mesure de l'erreur plus grande ; elle favorise le sur-ajustement aux données d'apprentissage.

4.7 Les CAD du cancer de la prostate : une application en développement

Comme nous l'avons explicité dans le chapitre 2, la détection et la localisation du cancer de la prostate à un stade précoce sont cruciales pour permettre une prise en charge (traitement ou surveillance active) efficace. Jusqu'à présent, les biopsies "pseudo-aléatoires" (ou *randomisées*, puisqu'il s'agit du terme consacré) restent la méthode de référence pour le diagnostic des adénocarcinomes. Néanmoins, en plus d'être un acte invasif et possiblement dangereux, le diagnostic par biopsies ne permet pas une exploration exhaustive de la glande. Il peut donc être source de sur-traitement (à cause d'une détection d'un foyer de taille minime) ou de sous-traitement (dû à des tirages de biopsies à côté d'une cible maligne) selon les cas. Les radiologues se tournent donc de plus en plus vers l'imagerie IRM et en explorent en particulier les performances pour l'aide au ciblage des biopsies vers des régions visuellement suspectes (signal IRM singulier). L'IRM, en permettant d'apprécier l'extension des foyers malins, pourrait également permettre la mise en place de traitements focaux moins radicaux que la prostatectomie. Nous avons vu, section 3.8, que bien qu'aucune séquence IRM n'ait à ce jour permis de discriminer de manière précise et fiable les zones de cancer des tissus bénins [55], de nombreuses études évaluant les performances

d'observateurs humains ont montré que la précision du diagnostic (évaluée dans la zone périphérique ZP, zone de prédominance de la carcinogenèse prostatique) peut être significativement améliorée en combinant différentes séquences IRM [19, 42, 45, 57, 82, 123, 140].

Néanmoins, fusionner et analyser un nombre croissant d'informations visuelles devient une tâche complexe et fastidieuse, en particulier pour des radiologues peu expérimentés. C'est une tâche d'autant plus difficile que, comme le montrent les figures 3.9 et 3.10 (page 48), les tissus malins et bénins peuvent présenter un signal similaire (du moins à l'œil...) dans une séquence et différent dans une autre ; et qu'il n'existe à ce jour aucun abaque précis pour analyser les images IRM et en particulier le cas d'informations contradictoires. L'objectif des systèmes d'aide au diagnostic est d'assister le radiologue durant sa tâche de diagnostic en lui fournissant un indice de confiance sur des zones suspectes.

Quelques méthodes d'aide au diagnostic du cancer de la prostate, reposant sur des algorithmes de classification supervisée ou non-supervisée, ont été proposées ces dix dernières années dans la littérature.

Dans leur article datant de 2006, Zhu *et coll.* [144] dressent un état de l'art des études émergentes sur l'aide au diagnostic du cancer de la prostate, mais aussi sur les techniques de simulation et de guidage des biopsies ainsi que de segmentation de la glande, en imagerie échographique et IRM. Ils soulignent que si l'utilisation des systèmes CAD s'est largement développée et a démontré son utilité dans certaines applications telles que la mammographie ou le cancer du poumon, leur application au cancer de la prostate reste alors trop limitées.

Six ans plus tard, nous proposons dans cette section une mise à jour de cet état de l'art en faisant le point sur les méthodes CAD récemment proposées pour l'imagerie de la prostate.

Ces méthodes peuvent être organisées selon plusieurs critères : le type de données utilisées (IRM mono ou multi-séquence, échographie, etc), la région prostatique étudiée (ZP, ZT, glande entière), le type d'algorithme de classification mis en œuvre (classification supervisée ou non-supervisée, etc) ou encore l'approche envisagée (CADx versus CADe).

On notera que la comparaison des méthodes proposées reste très difficile puisqu'aucune d'entre elles ne repose sur la même base de données (nombre de patients, procédure d'annotation, modalité d'acquisition d'images a priori différents, vérité terrain histologique ou radiologique...) et que les méthodes d'évaluation mises en œuvre ne sont pas standardisées (partitionnement différent de la base de données, évaluation à l'échelle du pixel, d'une région, d'une coupe, d'un patient...).

Les tableaux 4.3 et 4.4 présentent une synthèse de cet état de l'art que nous détaillons ci-après.

4.7.1 Systèmes d'aide à la décision CADx versus aide à la détection CADe

Deux types d'études CAD sont à distinguer dans la suite de cet état de l'art : celles se focalisant sur la conception de systèmes d'aide à la détection (CADe) et celles préférant l'approche de l'aide à la décision (CADx). Les études se focalisant sur l'établissement de systèmes système d'aide à la détection (CADe) proposent de calculer une carte de probabilité de présence du cancer, il s'agit notamment de celles de Chan *et coll.* [18], Madabushi *et coll.* [67,68], Langer *et coll.* [59], Viswanath *et coll.* [131], Ozer *et coll.* [83], Artan *et coll.* [5] et Lopes *et coll.* [65].

Les études portant sur des systèmes d'aide à la décision (CADx) proposent d'attribuer un score de suspicion de malignité pour une région d'intérêt (ROI) suspectée par le radiologue, il s'agit notamment de celles de Puech *et coll.* [93], de Tiwari *et coll.* [124] et Vos *et coll.* [132,133].

Si ces deux approches CAD n'abordent pas le même problème, elles reposent sur les mêmes algorithmes de discrimination.

4.7.2 Les méthodes supervisées utilisant l'IRM multi-paramétrique

Dans leur étude publiée en 2003, Chan *et coll.* [18] combinent trois types de caractéristiques (ou*features*) : (1) de l'information anatomique, traduisant la localisation (coordonnées cylindriques du voxel) au sein de la zone périphérique (ZP), (2) des valeurs d'intensités mesurées sur les images en T2-pondérée (T2-w), densité de protons et sur les cartographies T2 (T2-map) et cartes ADC (obtenues à 1.5 T), (3) des caractéristiques de texture extraites des différentes séquences (à partir de la matrice de co-occurence GLCM en particulier). Ces attributs sont utilisés pour construire une carte de probabilité de présence du cancer dans la zone périphérique (ZP) en appliquant des classifieurs de type maximum de vraisemblance, séparateur à vaste marge (SVM) et analyse discriminante linéaire (ADL). En l'absence d'analyse histologique, l'apprentissage se fait sur une base de données de 15 patients annotées par un radiologue expert (guidé par les résultats des biopsies). L'apprentissage est réalisé par validation croisée de type LOPO (*leave-one-patient-out*, apprentissage sur n-1 patients et test sur le dernier, répété n fois). Les performances maximales sont obtenues pour l'ADL utilisant tous les attributs (AUC=0.83) ; faute de convergence à l'apprentissage, les performances obtenues avec le SVM et tous les attributs (notamment les textures) ne sont pas quantifiées ($AUC_{SVM,intensité}$=0.64, $AUC_{SVM,intensité+anatomie}$=0.76). Les performances obtenues avec une classification au maximum de vraisemblance (utilisant un seul attribut à la fois) atteignent seulement $AUC_{SVM,T2w}$=0.6. Ils montrent que l'utilisation conjointe de tous les attributs aboutit à des performances de classification statistiquement bien meilleures qu'en utilisant uniquement les valeurs de niveaux de gris directement extraites des images ($AUC_{ADL,intensité}$=0.62). En conclusion, les auteurs soulignent l'apport de l'approche multi-séquence et de l'extraction d'attributs images. Il s'agit de la première étude proposant un schéma automatique de discrimination des tissus prostatiques malins/bénins à partir d'images IRM.

Madabushi *et coll.* [67] proposent, en 2005, d'extraire différents paramètres : (1) statistiques (médiane, moyenne locales, etc), (2) de gradients, et (3) de texture (paramètres de Gabor ou issus de la GLCM) à partir d'images IRM T2-w haute résolution (4 T) acquises *ex vivo*. Ils utilisent un classifieur Bayésien qui fournit, pour chacun des attributs pris individuellement, une carte de vraisemblance de l'appartenance à la classe maligne. Ces cartes sont ensuite fusionnées par différentes méthodes de combinaisont de vraisemblance : vote à la majorité, moyenne, adaboost et méthode de l'ensemble général (GEM, *general ensemble method*, qui construit la règle de décision à partir d'une combinaison linéaire des probabilités estimées sur chacun des attributs, [91]). L'évaluation repose sur les données de 5 patients parmi lesquelles seules 33 coupes IRM axiales sont considérées et annotées suivant la vérité histologique. L'apprentissage est effectué sur 5 coupes. Les auteurs montrent que la combinaison des différents attributs extraits permet d'obtenir des performances supérieures à celles obtenues en utilisant chaque attribut de manière indépendante. La comparaison des performances individuelles de chacune des caractéristiques montre que celles issues de la GLCM sont les plus discriminantes, suivies des attributs statistiques du premier ordre et de type gradient. Les paramètres de Gabor sont de loin les moins discriminants. La méthode de combinaisons de type GEM est la plus performante avec VPP=30% (valeur prédictive positive). Les résultats sont comparés aux performances de 4 experts mais aucune conclusion générale sur les différences de performances n'a pu être mise en évidence. En conclusion, les auteurs soulignent les limites de leur approche *ex vivo* mono-séquence (qui ne peut, par nature, pas être étendue à l'imagerie de diffusion ou de perfusion) et du faible nombre de cas d'apprentissage. En 2006, Madabushi *et coll.* [68] étendent leur étude à la comparaison des performances réalisées par un classifieur naïf de Bayes (CNB), un classifieur de type k-plus proches voisins (k-PPV), un algorithme de Boosting et de Bagging. Les auteurs montrent que les performances réalisées par le k-PPV sont les meilleures en termes de précision et soulignent ainsi qu'un classifieur non-paramétrique simple qui requiert un apprentissage minimal obtient de meilleures performances (AUC=0.94) que le classifieur de Bayes (AUC=0.93) et que des méthodes plus sophistiquées de Boosting (AUC=0.93) ou Bagging (0.92). Dans cette étude, Madabushi *et coll.* remarquent également que la variabilité entre les différents classifieurs est significativement plus faible que celle mesurée sur 5 experts humains.

En 2009, Viswanath *et coll.* [131] poursuivent le travail réalisé par Madabushi. Ils construisent un système d'aide à la détection (CADe) basé sur des forêts aléatoires d'arbres de décision (RDF) qui combinent les prédictions réalisées par un classifieur naïf de Bayes (CNB). Leur système intègre des attributs statistiques, de gradients et de textures (issus de la GLCM) extraits de l'image T2-w et les valeurs du signal mesuré sur séquence la DCE, acquises cette fois *in vivo* sur un imageur à 3 T. Là encore, la combinaison de l'information T2-w et DCE et des différents attributs extraits augmente de manière significative les performances de détection (AUC_{T2w}=0.7, $AUC_{T2w+attributs,DCE}$=0.81). A noter qu'un point faible de cette étude est le jeu de données très restrictif composé uniquement de 18 coupes axiales issues des acquisitions réalisées sur 6 patients.

Puech *et coll.* [93], proposent un système d'aide à la décision pour l'analyse des images DCE, utilisant des paramètres semi-quantitatifs [1] extraits des courbes de réhaussement en produit de contraste (temps et valeurs remarquables, pentes de WI, WO). Ils évaluent un algorithme de *scoring* sur 10 points, basé sur un ensemble de critères expérimentaux organisés sous forme d'arbre. Cette heuristique est évaluée sur 121 ROI extraites de 84 patients (zones périphériques et transitionnelles confondues). La vérité terrain est construite soit à partir de la vérité histologique lorsqu'elle est disponible, soit par un radiologue expert au vu des résultats des biopsies. Avec une AUC de 0.77, les performances du système d'aide à la décision (CADx) sont meilleures que celles d'un radiologue junior (AUC=0.57, $p < 0.0001$). Les auteurs soulignent néanmoins le manque de spécificité de la méthode proposée et le biais introduit par une validation par re-substitution. Les perspectives concernent l'utilisation de paramètres pharmacocinétiques issus de la modélisation des courbes de réhaussement DCE et de l'imagerie de diffusion.

Langer *et coll.* [59], réalisent en 2009 une analyse par régression logistique des niveaux de gris des images de type T2-map et ADC, et de paramètres pharmacocinétiques Ktrans et Ve calculés à partir de la DCE, obtenus à 1.5 T. L'évaluation repose sur une base d'images issues de 29 patients, pour lesquelles la vérité terrain a été reporté de l'histologie vers les images par un radiologue expert seul. L'apprentissage est réalisé sur un ensemble de ROI malignes/bénignes extraites des images. L'étude des performances réalisées par chacun des paramètres pris individuellement place l'ADC comme attribut le plus discriminant (AUC_{ADC}=0.68, AUC_{T2}=0.67, AUC_{Ktrans}=0.59 et , AUC_{ve}=0.54). Le modèle optimal consiste en une combinaison des valeurs d'ADC, de T2-map et de Ktrans ($AUC_{T2,ADC,Ktrans}$=0.70).

Dans leur étude publiée en 2010, Vos *et coll.* [132] évaluent le pouvoir de discrimination d'un schéma CADx utilisant les données issues des séquences IRM T2-w et DCE combinées avec un classifieur de type SVM. Outre les attributs correspondant aux intensités des images en T1 et T2, ils proposent d'extraire les paramètres pharmacocinétiques Ktrans, Ve, Kep et WO. Leur étude repose sur les données de 29 patients. La vérité terrain est construite à partir des données de l'analyse histologique reportées après consensus radiologue/anatomo-pathologiste sur les images IRM. Ils montrent que l'utilisation du signal de la séquence T2-w peut significativement améliorer les performances obtenues en utilisant uniquement les paramètres pharmaco-cinétiques issus de la DCE ($AUC_{DCE} = 0.84$ versus $AUC_{T2,DCE} = 0.89$), démontrant ainsi une fois de plus le besoin d'une approche multi-séquence. Cette étude fait suite à un premier papier publié en 2008 (Vos *et coll.* [133]) dans lequel seules les données issues de la DCE étaient exploitées. L'originalité de l'approche testée en 2008 est de distinguer dans l'établissement de leur vérité terrain : (1) les tissus malins (notés {M}), (2) les tissus bénins d'apparence normale ({N}) et (3) les tissus bénins d'apparence suspecte à l'IRM ({NS}), faux positifs potentiels. L'analyse des performances de leur système distingue donc deux problèmes de classification : (1) le

1. La section 7.2.3, page 106, sera consacrée à la définition des paramètres de perfusion semi-quantitatifs (WI, WO, etc) et pharmaco-cinétiques (Ktrans, Ve, Kep)

problème "classique" de discrimination des tissus malins et bénins ({M} versus {N, NS}), (2) la discrimination des tissus malins et suspects ({M} versus {NS}), plus difficile mais présentant un plus grand intérêt clinique. Les performances AUC_{DCE} obtenues par *leave-one-patient-out* (LOPO) sur les données de 34 patients sont respectivement de 0.92 et 0.83.

Artan *et coll.* [5] comparent, en 2010, différentes méthodes supervisées de type SVM (SVM ou C-SVM, avec ou sans optimisation du paramètre de coût de mauvaise classification C) et incorporent de l'information spatiale via l'utilisation de champs aléatoires conditionnels (CRF). Ils utilisent les caractéristiques de type T2-w, ADC et Kep obtenues sur 21 patients à 1,5 T. Les paramètres des classifieurs sont optimisés de manière à maximiser l'indice de Dice ; les performances obtenues $(\text{SE, SP, DSC})_{SVM}$= (0.73, 0.67, 0.40) versus $(\text{SE, SP, DSC})_{CRF}$=(0.64, 0.78, 0.46) montrent le potentiel de l'utilisation d'une méthode de régularisation spatiale.

En 2011, Lopes *et coll.* [65] comparent deux approches de classification : SVM et adaboost. Leur étude repose sur les données T2-w acquises sur 27 patients, pour lesquelles la vérité histologique est connue, et desquelles sont extraits des paramètres de texture de type ondelettes, filtre de Gabor et paramètres d'Haralick (issus de la GLCM) ainsi que des caractéristiques fractales. Les performances maximales sont obtenues en utilisant uniquement les attributs fractals ($\text{AUC}_{fractal}$= 0.92 versus $\text{AUC}_{ondelette,Gabor,Haralick}$=0.88). La comparaison des performances des deux algorithmes de classification n'est pas explicitée en termes d'AUC. Les perspectives concernent l'utilisation d'autres séquences IRM dans le schéma de classification.

4.7.3 Systèmes CAD reposant sur des algorithmes de classification non-supervisée

L'utilisation des méthodes de classification non-supervisée pour la localisation du cancer de la prostate est apparue plus tardivement dans la littérature. En 2009, Tiwari *et coll.* [124] proposent d'utiliser une modalité d'imagerie IRM encore peu exploitée : la spectroscopie par résonance magnétique (MRS). Cette modalité fonctionnelle fournit, sous forme de spectres, des informations sur la concentration relative de certains métabolites (citrate, créatine et choline en particulier) sensée révéler la présence de lésion malignes. Elle a déjà fait l'objet d'études impliquant des observateurs humains, avec des conclusions mitigées quant à son intérêt diagnostique (lire section 3.8). L'objectif de Tiwari *et coll.* est de classer automatiquement les spectres de résonance magnétique sans passer par une détection manuelle préalable des pics de métabolites, rendue difficile par le faible rapport signal sur bruit (SNR). Un algorithme hiérarchique de classification non supervisée (les k-moyennes), combiné à différentes approches d'extraction et de sélection de caractéristiques (z-score, analyse en composantes principales et réduction non-linéaire de dimension), est utilisé pour distinguer les ROIs de tissus prostatiques en trois classes de tissus : (1) normal, (2) suspicieux et (3) non déterminé. L'évaluation repose sur une série de 18 examens pour

lesquels la vérité terrain, en l'absence de pièce histologique, est issue de l'analyse, par un expert, des images IRM T2-w et MRS. Les auteurs mesurent une sensibilité de 81.39% et une spécificité de 64.71%. Dans leur conclusion, les auteurs proposent de combiner les données fonctionnelles MRS avec des images morphologiques en T2-w et d'évaluer leur méthode sur une base de données plus riche et plus fiable (position des foyers malins).

Liu *et coll.* [63] proposent, en 2009, une segmentation du cancer prostatique utilisant des images issues d'une acquisition IRM multi-paramétrique. La classification repose sur une modélisation en champs de Markov flous (FMRF). Les auteurs proposent de réaliser simultanément (processus itératif) l'estimation des paramètres du modèle (moyennes et variances des distributions, gaussiennes) et la classification (floues) des données, contrairement aux approches classiques où les valeurs de paramètres sont choisies arbitrairement ou déterminées par apprentissage. La méthode est testée sur un jeu de données multi-paramétriques obtenu sur 11 patients pour lesquelles la vérité histologique est connue. Les caractéristiques étudiées sont : le signal (SI) mesuré en T2-w et T2-map, les valeurs d'ADC ainsi que les paramètres *wash-in* (WI), *wash-out* (WO) et l'aire sous la courbe de réhaussement (AUGC) calculés à partir de la DCE. Les performances maximales obtenues sont (SP, SE)=(0.89, 0.87) correspondant à un indice de Dice DSC de 0.62. Les auteurs montrent l'apport de leur méthode par rapport à une classification par champs de Markov flous (MRF) classique. En perspective, les auteurs proposent d'ajouter aux mesures brutes d'intensité (qui souffrent d'un fort recouvrement tissus sains/pathologiques) des caractéristiques de texture afin de tester leur pouvoir discriminant.

En 2010, Ozer *et coll.* [83] proposent de comparer la classification non-supervisée par FMRF proposée par Liu *et coll.* [63] à deux approches par classification supervisée : séparateur à vaste marge (SVM) et *relevance vector machine* (RVM) [2]. Les paramètres des SVM et RVM sont optimisés soit de façon à maximiser la précision, soit de façon à garantir une spécificité par patient donnée. Trois types de caractéristiques, issues de trois séquences IRM, sont ici exploitées : le SI T2-w, les valeurs de la carte ADC, et les valeurs de constante de flux Kep, calculées à partir de la série DCE. L'évaluation repose sur les données de 20 patients pour lesquels les données histologiques sont accessibles (prostatectomie radicale). Les auteurs soulignent qu'il n'y a pas de différence significative entre les résultats de classification obtenus par les méthodes supervisées SVM et RVM ($Acc \simeq 0.85$ et AUC$\simeq$0.82). Ces résultats sont en revanche meilleurs que ceux obtenus par la méthode FMRF (non-supervisées) développée dans Liu *et coll.* [63]. Ils soulignent également que l'écart-type dans les performances des FMRF est plus important que ceux des méthodes supervisées (segmentations moins 'robustes'). Outre la puissance de l'algorithme, une explication envisagée est la plus grande sensibilité des méthodes non-supervisées à un recalage imparfait des images issues des différentes séquences IRM. Pour contourner ce problème de recalage inter-séquences, l'étape d'apprentissage des méthodes supervisées n'est pas effectuée à l'échelle du voxel mais sur les valeurs moyennes calculées sur les régions d'intérêt ROIs.

2. Relevance Vector Machine : classifieur supervisé adaptant au problème de classification un apprentissage bayésien par régression sur des points à étiquettes réelles.

Publication	CAD	Données	Acquisition	Vérité terrain	Caractéristiques	Classifieur	Evaluation	Résultats	Commentaires
Chan [18], 2003	CADe	15 patients	IRM 1,5T, T2-w, T2-map, ADC, DP	biopsies + signal IRM	SI T2-w, T2-map, ADC, DP et T2-w, coord. cylindriques, texture (GLCM, DCT)	ADL, SVM	LOPO, courbe ROC	AUC_{ADL} = 0.83, AUC_{SVM} = 0.76	Pas de convergence avec les caractéristiques de texture pour le SVM (AUC_{SVM} obtenue uniquement sur mesures SI et coord. cylind.)
Madabushi [67], 2005	CADe	5 patients (33 coupes IRM)	IRM 4T, T2-w *ex-vivo*	histologie partielle	stats de 1erordre, gradients, Gabor, GLCM	Bayes + {Vote à la majorité / Adaboost / GEM}	*hold-out*	VPP=30%	Classification de chaque carte de caractéristique indépendamment puis combinaison. Base de données réduite à 5 patients. Vérité terrain partiellement annotée.
Madabushi [68], 2006	CADe	5 patients (33 coupes IRM)	IRM 4T, T2-w *ex-vivo*	histologie partielle	stats de 1er ordre, gradients, Gabor, GLCM	Bayes + k-PPV + Bagging + Boosting	courbes ROC	AUC_{Bayes}=0.93, AUC_{kPPV}=0.94, AUC_{Bag}=0.92, AUC_{Boost}=0.93	Variabilité inter-classifieur significativement plus faible que variabilité inter-expert. Meilleures performances + robustesse du k-PPV.
Viswanath [131], 2009	CADe	6 patients (18 coupes IRM)	IRM 3T, T2-w, DCE	histologie partielle	stats de 1er ordre, gradients, GLCM, SI DCE	Bayes + forêt aléatoire de décision (RDF)	3-fold avec re-substitution	AUC=0.81	Procédure de validation avec re-substitution. Jeu réduit de 18 coupes.
Vos [133],2008	CADx	34 patients	IRM 1,5T, DCE	histologie	SI T1(t=0), WO, Ve, Ktrans, Kep	SVM	LOPO, courbe ROC	AUC = 0.83	Pas de traitement d'image. Dissociation des performances de discrimination des tissus sains d'apparence normale et d'apparence suspecte à l'IRM
Puech [93], 2009	CADx	84 patients (121 ROIs)	IRM 1,5T, DCE	biopsie + signal IRM	WI, WO	Heuristique, arbre de décision	courbe ROC	SE_{ZP}=100%, SP_{ZP}=45%, AUC_{ZP}=0.77	Analyse de toute la prostate (ZP et ZT) ; prédit un score de suspicion $\in$ {1,2,3,4,5}
Langer [59], 2009	CADe	29 patients	IRM 1,5T, ADC, T2-map, DCE	histologie	SI T2-map, ADC, Ve, Ktrans	Régression logistique	bootstrap + apprentissage / test sur même données + courbe ROC	AUC = 0.70	Ve qualifié de non-informatif ; comparaison voxel à voxel (résolution dégradée) ; ADC le plus discriminant
Vos [132], 2010	CADx	29 patients	IRM 1,5T, T2-w, DP, DCE	histologie	SI T2-map, SI T1(t=0) , WO, Ve, Ktrans, Kep	SVM	LOPO, courbe ROC	AUC = 0.89	Pas de traitement d'image. Quantification de l'apport du T2-map par rapport à la DCE seul [133].
Ozer [83], 2010	CADe	20 patients	T2-map, DCE, DWI	histologie	T2-map, ADC, Kep	SVM, RVM	LOPO, courbe ROC	AUC_{SVM} $\simeq$0.82, AUC_{RVM} $\simeq$0.82	-
Artan [5], 2010	CADe	21 patients (21 coupes)	IRM 1,5T, T2-w, DCE, ADC	histologie	SI T2-w et ADC, Kep	SVM, CRF	LOPO, courbe ROC, DSC	SE_{SVM}=0.73 SP_{SVM}=0.67, DSC_{SVM}=0.40, SE_{CRF}=0.64 SP_{CRF}=0.78, DSC_{CRF}=0.46	DSC faible. Utilisation d'une 1 seule coupe IRM jugée significative par patient
Lopes [65], 2011	CADe	27 patients	T2-w	histologie	Fractal (FD, mBm), GLCM, ondelettes	SVM, Adaboost	courbe ROC, 4-fold (échelle voxel)	AUC=0.92	Données du même patient utilisées pour apprentissage/test

Table *4.3* – *Etat de l'art des schémas CAD reposant sur des méthodes de classification supervisée*

Publication	**CAD**	**Données**	**Acquisition**	**Vérité terrain**	**Caractéristiques**	**Classifieur**	**Evaluation**	**Résultats**	**Commentaires**
Tiwari [124], 2009	CADx (meta-voxels)	18	IRM 1,5T, MRS	signal IRM	valeurs du spectre	sélection des *features* via z-score, ACP, réduction non-linéaire (LLE, GE) + k-moyenne	ROC	SE=81%, SP=65%	Evaluation sur les 6 exemples annotés (parmi les 18 jeux de données)
Liu [63], 2009	CADe	11	IRM 1,5T, T2-w, ADC, T2-map, DCE	histologie	SI T2-w, T2-map, ADC ; WI, WO, AUGC	champs de Markov flous (FMRF)	-	SE= 87%, SP = 89%, DSC=0.62	DSC faible. Perspective sur l'utilisation de caractéristiques de texture.
Ozer [83], 2010	CADe	20	T2-map, DCE, DWI	histologie	T2-map, ADC, Kep	MRF	ROC	AUC_{MRF} ≃0.76	Sensibilité au mauvais recalage des séquences

Table 4.4 – *Etat de l'art des schémas CAD reposant sur des méthodes de classification non-supervisée*

4.7.4 Discussion

Les publications concernant le développement de prototypes CAD pour le cancer de la prostate en IRM se sont multipliées ces 10 dernières années.
Les études testent différentes approches de classification, supervisée et non-supervisée, mais reposent sur des bases de données différentes, rendant difficile toute comparaison d'autant plus que les mesures d'évaluation ne sont pas homogènes. En effet, si dans certaines études l'unité de base (pour la comptabilisation de VP/FP etc) est une région d'intérêt (ROI), dans d'autres, c'est le pixel ou la coupe IRM, etc. De plus, les performances sont, selon l'étude, évaluées par le biais de la courbe ROC, par la mesure de Dice, par un couple sensibilité/spécificité etc. Enfin, le protocole de validation aussi est souvent différent. Certaines études réalisent une validation croisée de type *leave-one-patient-out* (LOPO), *leave-one-out* (LOO ; par opposition au LOPO, elle implique que l'apprentissage et le test sont réalisés sur les données d'un même patient), ou *k-fold* ; d'autres évaluent leur algorithme par resubstitution, etc.
On notera également que les jeux de données exploités sont parfois très restreints [67, 68] rendant difficile la généralisation des résultats.
On remarque que, de manière générale, les méthodes de classification supervisée semblent engendrer de meilleures performances que les approches non-supervisées.
Enfin, si deux études [67, 68], proposent une comparaison des performances de leur prototype CAD avec celles d'experts, aucune étude ne propose une évaluation de l'apport clinique du prototype présenté. En effet, l'objectif d'un système CAD n'est pas de se substituer à l'expert mais plutôt de l'assister dans sa pratique clinique.

4.8 Conclusion

Ce chapitre nous a permis de faire une présentation des grandes étapes de construction des systèmes d'aide au diagnostic (CAD), des algorithmes les plus couramment exploités et des méthodes d'évaluation existantes.
Dans une dernière section, nous avons fait l'état de l'art des prototypes CAD développés dans le cadre du diagnostic du cancer de la prostate par imagerie IRM. Au regard de cet état de l'art, nous avons proposé notre contribution, détaillée dans la partie II suivante.
Le chapitre 5 présente le contexte et l'objectif de notre étude ainsi que nos choix méthodologiques, faits en s'appuyant notamment sur les limites observées dans cet état de l'art.

II Diagnostic assisté par ordinateur du cancer de la prostate par analyse des images IRM multi-paramétrique

Chapitre 5

Choix méthodologiques

5.1 Introduction

Ce chapitre précise le contexte, les motivations et objectifs de cette première partie de thèse. Nous présentons les choix méthodologiques réalisés après étude de l'état de l'art.

5.2 Contexte et motivations

Ce travail de thèse s'inscrit dans un contexte clinique fort puisqu'il est motivé et co-encadré par le professeur Olivier Rouvière, praticien hospitalier et chef du service de radiologie urinaire et vasculaire de l'hôpital Edouard Herriot (Lyon, France).

Le cancer de la prostate est l'un des sujets d'étude prioritaires de ce service et a d'ores et déjà fait l'objet de nombreuses publications médicales, concernant notamment :

- le guidage des biopsies par les données de l'IRM [19],
- l'étude des récidives après traitement HIFU par imagerie IRM [79, 99, 101],
- la possibilité d'un diagnostic fiable, voire d'une cartographie du cancer, par IRM multi-paramétrique (IRM-mp) [7, 42, 78, 103]. Le Pr. Rouvière est d'ailleurs porteur d'un projet INCa sur cette thématique (projet Cartographix, 2010-2013).

C'est dans ce dernier axe de recherche que s'inscrit cette thèse.

Comme nous l'avons déjà souligné, si le développement du traitement focal du cancer de la prostate, alternatif à la prostatectomie radicale, est devenu un enjeu clinique majeur, il ne peut être envisagé que si les foyers malins peuvent être parfaitement localisés dans la glande, i.e. si une technique d'imagerie en permet l'identification selon des critères fiables

et reproductibles. L'IRM-mp est aujourd'hui la technique la plus prometteuse en ce sens.

Néanmoins, ainsi que l'illustrent les figures 3.9 et 3.10, présentées page 49, l'analyse des données IRM-mp est une tâche difficile. Elle nécessite en effet de synthétiser l'information portée par des images de nature différente et la prise de décision face à des signaux souvent contradictoires est très subjective. La variabilité des diagnostics intra- et inter-expert est telle [38, 103] qu'il n'est aujourd'hui pas possible d'envisager une cartographie du cancer (diagnostic et localisation des foyers malins) reposant sur l'imagerie IRM-mp, qui permettrait, à terme, un traitement focal des tumeurs. De plus, la difficulté à discriminer les foyers suspects des foyers réellement malins sur les images (faible spécificité) rend l'utilisation de l'IRM dans le cadre de re-biopsies encore trop inefficace (en comparaison de biopsies pseudo-aléatoires quadrillant la prostate). Se limiter à quelques foyers permettrait en outre d'effectuer plusieurs prélèvements sur une même cible et ainsi de mieux appréhender son extension et son score de Gleason.

5.3 Objectifs

Le besoin clinique initial et prioritaire souligné lors de la mise en place de ce projet de thèse est celui d'un système automatique permettant une aide à la décision (CADx) i.e. proposant un "deuxième avis" au lecteur sur une cible suspectée.
En effet, la difficulté n'est pas tant de détecter des anomalies de signal sur les images IRM que de parvenir à discriminer, parmi ces foyers, les zones suspectes (mais saines) des zones réellement malignes.
Comme l'illustre la figure 5.1, l'objectif est d'aider le radiologue à poser un diagnostic correct en proposant un score de malignité fiable, objectif et reproductible sur des cibles suspectées.
L'intérêt d'un tel système automatique est notamment sa reproductibilité. En fournissant un score de suspicion de malignité utilisé comme second avis, il pourrait non seulement être une aide permettant au clinicien d'améliorer son diagnostic mais pourrait aussi permettre de réduire la variabilité inter- et intra-observateur, en particulier entre praticiens juniors et seniors.
Comme nous l'avons présenté au chapitre précédent, un système CAD repose sur un ensemble de choix méthodologiques standardisés, que nous définissons et justifions ci-après. Ces choix seront ensuite re-précisés et détaillés dans les chapitres 6 et 7 suivants.

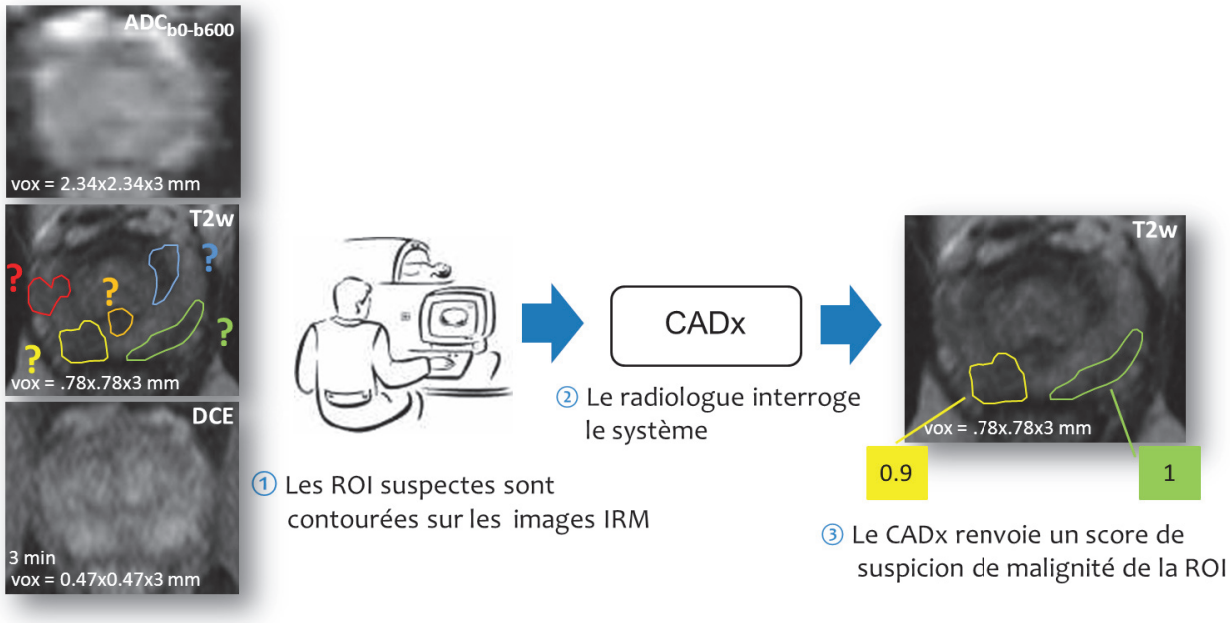


Figure 5.1 – *Schéma d'utilisation du système d'aide à la décision (CADx) :*
(1) le radiologue contoure les zones qu'il suspecte d'être malignes (éliminant au passage les zones qu'il sait, par expérience, identifier comme bénignes, telles que les zones de stroma fibromusculaire, la présence de sang suite à biopsies, etc) ;
(2) il interroge le système qui applique son modèle de prédiction aux données testées ;
(3) le CADx propose au radiologue un score de malignité (entre 0 et 1 sur l'exemple) pour chaque ROI, qu'il peut alors utiliser comme un "second avis".

5.4 Choix méthodologiques

5.4.1 Base de données

A ce jour, aucune base de données d'images IRM-mp de la prostate n'est disponible au partage par la communauté scientifique.
Cette étude repose donc sur une base de données cliniques que nous avons collectées à l'hôpital Edouard Herriot (Lyon, France). La base CLARA-P (pour *corrélations anatomo-radiologiques en IRM de prostate*) a été initiée en 2009 et nous l'avons enrichie de nouveaux cas tout au long de cette thèse. Elle comprend les données radiologiques et histologiques de patients pour lesquels la présence du cancer a été vérifiée par biopsie et qui ont subi un traitement par prostatectomie radicale. La base de données utilisée ne contient donc que des sujets pathologiques.

Nous avons choisi de baser notre étude sur une imagerie IRM multi-paramétrique (IRM-mp). En effet, comme détaillé dans la section 4.7, dans leurs papiers respectifs, Chan *et coll.* [18], Viswanath *et coll.* [131], Langer *et coll.* [59], Puech *et coll.* [93], Ozer *et coll.* [83], Artan *et coll.* [5], Vos *et coll.* [132], et beaucoup d'autres, soulignent que la combinaison de plusieurs séquences IRM améliore systématiquement les performances de

discrimination des tissus.
Dans notre étude, trois types d'images IRM, les plus exploitées dans la littérature, ont été acquises : la séquence "morphologique" T2-pondérée (T2-w) et les séquences "fonctionnelles" de diffusion (DWI) et de perfusion (DCE). Elles ont été réalisées sur l'IRM 1.5 T de l'hôpital Edouard Herriot. Les données devant être obtenues en routine clinique, il était nécessaire de réaliser l'examen sur un temps restreint et donc de limiter notre choix à quelques séquences. Les séquences de spectroscopie ou de cartographie T2 notamment, plus coûteuses en temps, n'ont pas été retenues.

Une faiblesse soulignée dans beaucoup d'études [18, 59, 75, 93, 124] est le manque d'une vérité terrain histologique fiable. Beaucoup de travaux reposent soit sur une vérité histologique partielle, soit sur des résultats de biopsies *randomisées* ou sur une "vérité expert" issue de l'analyse *en aveugle* des images IRM par un radiologue expert.
Au contraire, dans notre étude, les données de la pièce de prostatectomie sont systématiquement analysées de manière exhaustive par deux anatomo-pathologistes et utilisées comme référence pour établir une cartographie précise des lésions malignes sur les images IRM-mp.
A noter que la construction d'une telle base de données est un travail de longue haleine. D'une part, elle nécessite la mobilisation de radiologues, d'anatomo-pathologistes et d'un chercheur pour des séances de travail (mise en corrélation anatomo-radiologique) longues et fastidieuses (compter 1 heure minimum par cas annoté). D'autre part, le délai entre l'examen IRM-mp et la revue effective en consensus des pièces histologiques issues de la prostatectomie est de 6 mois environ. Enfin, un certain nombre de données ont dû être exclues de l'analyse, soit parce que les coupes histologiques étaient trop endommagées (lors de la découpe au microtome, cf. section 6.4), soit parce que les artefacts étaient trop nombreux sur les images IRM (mouvement du patient, prothèse, etc), rendant dans les deux cas les données illisibles.

Notre étude, que nous présentons dans le chapitre suivant, repose sur les données de 30 patients. Elle représente une base riche, comparée à la majorité des études de la littérature. Notre banque de corrélations anatomo-radiologiques constitue un socle solide pour la mise en place de cette étude et un atout pour la réussite du projet.

5.4.2 Choix des caractéristiques descriptives

Beaucoup d'études [5, 59, 83, 124] se sont limitées à l'utilisation directe des valeurs du signal IRM acquis (niveaux de gris des images) comme caractéristiques pour la discrimination des tissus sains et pathologiques. Or, de nombreux autres travaux [18, 65, 67, 131, 133] ont souligné l'importance de ne pas se limiter à l'étude des signaux bruts mais, au contraire, d'extraire à partir des images en niveaux de gris des caractéristiques descriptives de différents types, permettant souvent d'améliorer les performances de discrimination.

Après avoir fait la synthèse des paramètres descriptifs proposés dans la littérature, nous choisissons d'extraire des caractéristiques aussi bien structurelles (statistiques locales, gradients et textures) que fonctionnelles (attributs descriptifs de la cinétique de réhaussement en agent de contraste) des différentes séquences IRM. La majorité de ces caractéristiques descriptives (ou *features*) ont été sélectionnées, après analyse de la littérature, en fonction de leur pouvoir démontré à discriminer les tissus sains des tissus malins. Nous les présentons de manière détaillée dans la section 7.2, page 98.

Nous incluons dans notre méthodologie une étape de sélection des caractéristiques afin de ne conserver que les caractéristiques les plus informatives dans la tâche diagnostique parmi l'ensemble de celles extraites. Cette étape de pré-sélection des caractéristiques n'a pas été beaucoup étudiée dans la littérature des CAD de prostate.

Notons que nous avons choisi de ne pas nous intéresser aux caractéristiques de forme dans cette approche par ROI. D'une part parce que le cancer de la prostate est de forme très irrégulière et peut aussi bien se présenter sous une forme parfaitement sphérique, bien définie ou au contraire être de type "pieuvre", infiltrant et de fait difficile à délimiter (et dans ce cas sa forme dépendant donc du contourage expert). La forme n'apparaît donc pas être un critère discriminant [103]. D'autre part, les approches de classification retenues dans cette étude impliquent de délimiter des régions saines sur la prostate. Ceci étant réalisé de manière automatique et aléatoire, la forme de la région contourée n'est pas porteuse d'information.

5.4.3 Choix des classifieurs

Ainsi que nous l'avons présenté dans le chapitre 4 dédié aux systèmes CAD, de nombreux algorithmes de classification ont été proposés et utilisés dans la littérature des CAD de prostate.

Ayant à notre disposition une base de données cliniques annotées, notre travail s'est naturellement orienté vers la conception d'un système de classification supervisée afin d'exploiter au maximum l'information à disposition. Cette orientation a été confirmée par l'analyse de l'état de l'art dans lequel la comparaison des méthodes de classification supervisée/non supervisée est favorable aux méthodes supervisées (lire section 4.7, page 69).

La comparaison des performances des algorithmes proposés dans les diverses études de la littérature est difficilement réalisable puisqu'ils sont évalués sur des bases de données différentes (populations et modalités d'imagerie), selon des protocoles de validation différents (validation croisée type *leave-one-patient-out*, *leave-one-out*, *hold-out*, etc ; resubstitution) et selon des critères variables (analyse ROC, indice de Dice, couple spécificité/sensibilité etc ; unité de comptabilisation : pixel, ROI, coupe, etc).
En nous appuyant sur notre base de données CLARA-P, nous proposons donc de comparer de manière objective quatre algorithmes de classification parmi les plus cités dans la littérature CAD : le séparateur à vaste marge (SVM), l'analyse discriminante linéaire (ADL),

le classifieur naïf de Bayes (CNB) et les k-plus proches voisins (k-PPV) afin d'identifier l'approche optimale. Ces classifieurs reposent sur des approches d'apprentissage différentes mais toutes relativement rapides. Nous les présentons de manière détaillée dans la section 7.5, page 121.

5.4.4 Evaluation

Aucune des études présentées dans l'état de l'art, section 4.7, ne propose une évaluation clinique des prototypes CAD conçus. Deux études [67, 68], de la même équipe, proposent une comparaison des performances de leur prototype CAD avec celles d'experts. Néanmoins, l'objectif d'un système CAD n'est pas de se substituer à l'expert mais plutôt de l'assister dans sa pratique clinique.

Dans notre étude, il nous a semblé important de déterminer l'intérêt pratique de notre système en l'évaluant dans un contexte clinique réel. Nous proposons de quantifier son apport, en termes de performances diagnostiques, auprès de douze radiologues d'expérience variable entre 6 mois et 7 ans. Outre l'évolution des performances globales, c'est l'évolution des disparités entre radiologues juniors et seniors ainsi que la confiance dans le système qui sont étudiées. Cette étude est présentée dans le chapitre 9, page 135.

5.5 Conclusion

Dans cette partie II de la thèse, nous tentons de répondre aux deux questions suivantes :

- quel couple Caractéristiques/ Classifieur est le plus pertinent dans la tâche de discrimination des tissus sains et malins ?
- quelle est l'utilité clinique du meilleur couple sélectionné ?

Chapitre 6

Base de données cliniques d'apprentissage

6.1 Introduction

Afin d'étudier les caractéristiques du cancer de la prostate sur l'imagerie IRM et, à terme, d'en améliorer le diagnostic, la constitution d'une base de données prospective (base "CLARA-P" pour CorréLations Anatomo-RAdiologiques en IRM de Prostate) à l'initiative du Pr. Olivier Rouvière, a débuté en septembre 2008 après déclaration au CPP (comité de protection des personnes, Sud-Est IV, Référence : L 09-04) et à la CNIL (commission nationale de l'informatique et des libertés, traitement n° 08-06). Dans le cadre de ce projet, trois types d'imagerie IRM (T2-w, de diffusion et de perfusion) sont réalisés de façon systématique sur les patients de l'hôpital Edouard Herriot (Lyon, France) au préalable à une prostatectomie radicale lorsque la présence de cancer est avérée (par biopsies) (lire section 6.2). Une première analyse en aveugle, décrite en section 6.3, avec contourage de tous les foyers suspects, est réalisée par deux radiologues (Olivier Rouvière _ senior _ et Flavie Bratan _ junior). Les pièces de prostate sont ensuite reçues à l'état frais par l'équipe d'Anatomie et Cytologie Pathologique de l'hôpital Edouard Herriot, fixées et détaillées selon un protocole qui sera présenté dans la section 6.4. La lecture histologique se fait de façon standard par une anatomo-pathologiste (Florence Mège-Lechevallier _ senior _ assistée d'Anne-Laure Chesnais _ junior) avec contourage, sur les lames, des foyers carcinomateux. Une deuxième lecture est programmée pour une corrélation anatomo-radiologique afin d'établir une vérité diagnostique (lire section 6.5). Les coupes histologiques réalisées sur les prostates prélevées sont mises en correspondance avec les images IRM et servent de référence pour la localisation *a posteriori* des tumeurs. C'est à ce moment que sont revus les faux positifs et les faux négatifs des radiologues pour obtenir une explication. La vérité

terrain histologique est alors reportée sur l'imagerie, constituant la base de l'apprentissage de nos classifieurs pour l'établissement d'un système d'aide au diagnostic qui sera présenté dans le chapitre 7 suivant.

6.2 Acquisition des données IRM-mp

6.2.1 Déroulement de l'examen

Tous les examens IRM de la prostate sont exécutés selon un protocole standardisé. Les examens IRM sont réalisés sur un scanner clinique à 1.5 T (Siemens Magnetom Symphony MR 2004A, Siemens Medical Systems, Erlangen, Allemagne) en utilisant une antenne pelvienne multi-canaux (voir illustrations 3.5a et 3.5b, page 38). Les paramètres des séquences sont ceux utilisés en routine clinique à l'hôpital Edouard Herriot de Lyon (France) par le Pr. Rouvière pour l'analyse post-biopsies (et dans certains cas pré-biopsies) de la prostate. Dans l'ordre, ce sont d'abord les images en T2-pondérées (T2-w) qui sont acquises, suivies des images pondérées en diffusion (DWI), pour finir avec les images de perfusion (DCE). Le temps total de l'examen est approximativement de 30 minutes.

6.2.2 Paramètres des séquences

Pour commencer, les images T2-w turbo-spin-écho sont obtenues dans les plans axial, sagittal et coronal. Les 24 coupes axiales contiguës couvrent entièrement la glande [temps de répétition (T_R) : 7750 ms ; temps d'écho (T_E) : 109 ms ; champ de vue (FOV) : 200×200 mm ; matrice : 256×256 ; nombre de coupes : 24 ; épaisseur des coupes : 3 mm, sans espacement].

Les données de la séquence DWI sont ensuite acquises [T_R : 4800 ms ; T_E : 90 ms ; FOV : 300×206 mm ; matrice : 128×88 ; nombre de coupes : 24 ; épaisseur des coupes : 3 mm, sans espacement ; angle de bascule (flip angle) : 90° ; valeurs de b : 0 and 600 s/mm^2].

Enfin, une séquence DCE acquise en T1-pondérée (T1-w) et écho de gradient 3-D *fat-saturated* est obtenue dans le plan axial avant et après injection de produit de contraste (bolus injecté en intra-veineuse à une vitesse de 3 ml/s, à une dose de 0.1 mmol/kg de gadotérate méglumine _ Gd-DOTA, Dotarem, Guerbet, Roissy, France) [T_R : 5.38 ms ; T_E : 2.73 ms ; angle de bascule : 10° ; FOV : 210×240 mm ; matrice : 448×512 ; nombre de coupes : 24 ; épaisseur des coupes : 3 mm, sans espacement ; temps d'acquisition : 15 s, répétée 12 fois, pour une durée totale de 3 min].

Notons que l'épaisseur de coupe de 3 mm a été choisie de manière identique pour toutes les séquences IRM réalisées pour permettre une comparaison directe coupe à coupe entre séquences.
Les paramètres des différentes séquences utilisées sont synthétisés dans le tableau 6.1.
Une description détaillée des spécificités de chaque séquence est donnée au chapitre 3, partie I.

Séquence	T_R **(ms)**	T_E **(ms)**	**Angle bascule (°)**	**Epaisseur coupe (mm)**	**Matrice (vox)**	**Nb. coupes**	**FOV (mm)**	**dimension voxel (mm)**	**temps échantillonnage (s)**
T2-w	7750	109	180	3	256×256	24	200×200	.78x.78x3	-
DCE	5.38	2.73	10	3	448×512	24	210×240	.47x.47x3	15
DWI	4800	90	90	3	128×88	24	300×206	2.34x2.34x3	-

Table *6.1* *– Paramètres utilisés pour l'imagerie de la prostate par IRM à 1.5 T*

6.2.3 Proportionnalité entre le signal mesuré et la concentration en agent de contraste

Nous proposons, section 7.2.3, de modéliser la cinétique de l'agent de contraste (AC), capturée par la séquence DCE, par le biais d'un modèle pharmaco-cinétique permettant d'extraire différents paramètres traduisant les propriétés des milieux traversés. Cette modélisation repose sur la mesure de concentration en AC, à laquelle nous n'avons pas directement accès [1]. Il est alors d'usage de faire l'hypothèse de proportionnalité entre le signal T1-w mesuré (qui se traduit directement par les valeurs de niveaux de gris de l'image) et les valeurs de concentrations en AC [116, 138].

Un protocole permettant de vérifier si cette hypothèse est bien raisonnable a été mis en place. Une gamme de 15 tubes témoins a été réalisée avec les concentrations en gadolinium suivantes :

$$0,\ 0.05,\ 0.1,\ 0.2,\ 0.3,\ 0.5,\ 0.75,\ 1,\ 1.5,\ 2,\ 2.5,\ 3,\ 4,\ 7.5,\ 10\ \text{mM/L}.$$

La dilution en cascade de l'AC (Gd-DOTA, Dotarem 0.5 mmol/L, laboratoire Guerbet, France) a été réalisée dans du sérum physiologique.

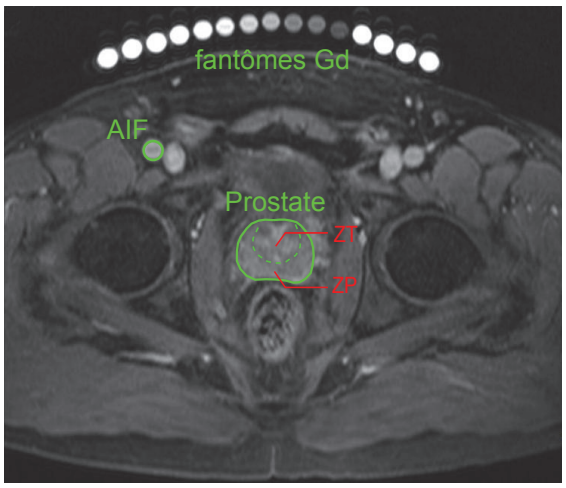
La ceinture contenant les 15 tubes témoins de concentration croissante en AC est placée sur le pelvis d'un échantillon de 10 patients test (voir illustration 6.1a) pour identifier l'intervalle de linéarité entre le signal mesuré et la concentration réelle en AC. Le signal mesuré sur ces fantômes (voir illustration 6.1b) suggère une relation linéaire (coefficient de corrélation : 0.98) entre la concentration en gadotérate méglumine (Gd-DOTA) et le signal IRM, pour l'ensemble des valeurs mesurées dans la glande prostatique et dans l'artère fémorale (on appelle fonction d'entrée artérielle _AIF_ cette dernière valeur, voir l'illustration 6.1c). Ceci nous encourage ainsi à utiliser directement le signal IRM mesuré

1. Les valeurs de concentrations en AC peuvent être calculées à partir des cartes T1 réalisées pour différents angles de bascule. On a :

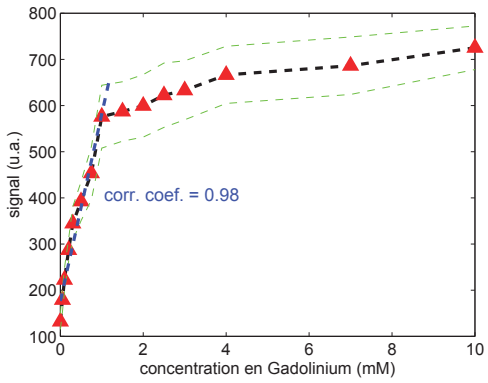
$$1/\mathrm{T}_1 = 1/\mathrm{T}_{1,0} + \mathrm{R}_1.C_{[AC]}$$

avec $\mathrm{R}_1 = 3.5$ L/mmol.s, la relaxivité du Gadolinium et $\mathrm{T}_{1,0} = 3$ s, le T_1 natif correspondant au temps de relaxation sans AC. Néanmoins, la cartographie T1 n'est pas réalisée en routine dans notre étude.

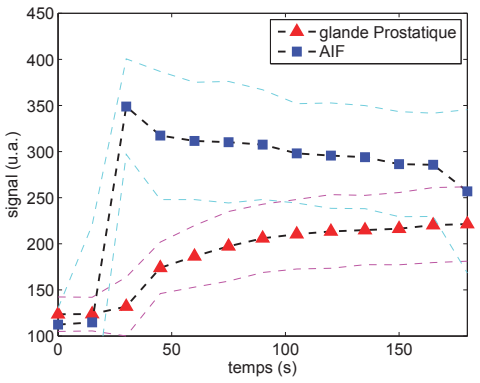
pour la modélisation pharmaco-cinétique.



(a) Patient de contrôle



(b) Courbe signal/concentration



(c) Signal mesuré sur la prostate et l'AIF

Figure *6.1* – *(a) Image T1-w MR axiale acquise sur un patient de contrôle de 56 ans, 3 minutes après l'injection de Gd-DOTA. Les contours de la glande prostatique ainsi que les zones périphériques et de transition (ZP et ZT) sont délimitées. Une ROI est dessinée sur l'artère fémorale pour mesurer la fonction d'entrée artérielle (AIF). (b) Tracé du signal moyen mesuré sur les 15 tubes de concentration témoins placés sur les patients de contrôle. (c) Valeurs moyennes du signal mesuré sur la glande prostatique et dans l'AIF sur les patients de contrôle, avec les intervalles de confiance $\pm\sigma$ correspondants.*

6.3 Analyse des images IRM

Deux radiologues, ayant respectivement 2 et 15 ans d'expérience dans l'analyse des images IRM du cancer de la prostate, analysent en aveugle les images IRM des patients et contourent les régions jugées suspectes. Lorsqu'un tissu d'intérêt est visible sur plusieurs coupes consécutives (mêmes caractéristiques image), il est contouré sur chacune des coupes sur lesquelles il est visible. Dans la suite, le terme ROI se rapporte donc à l'ensemble de ces contours 2-D réalisés pour délimiter un tissu cible. Un score correspondant au degré

de suspicion de malignité est affecté à chaque ROI mise en évidence [27] :

- 0 : bénignité certaine,
- 1 : probablement bénin,
- 2 : intermédiaire,
- 3 : probablement malin,
- 4 : malignité certaine.

6.4 Analyse des données histologiques

Préparation des pièces de prostatectomie

Tous les patients considérés dans cette étude ont subi une prostatectomie radicale. Chaque pièce de prostatectomie est traitée selon la technique standard dite de Stanford schématisée sur la figure 6.2c et dont les étapes sont détaillées ci-dessous.

Orientation de la pièce de prostatectomie : la pièce est orientée vésicules séminales en haut et en arrière, face rectale plane, face antérieure convexe et apex en pointe vers le bas. Le lobe droit et le lobe gauche sont encrés respectivement en jaune et noir, la face antérieure en rouge (voir illustration 6.2b).

Fixation de la pièce de prostatectomie : la pièce est fixée pendant 24 heures au minimum dans un fixateur formolé (formol tamponné à 10).

Séparation des vésicules séminales et de la base : les vésicules séminales (en bleu et rouge sur l'illustration 6.2c) sont coupées au niveau de la base selon un axe horizontal et séparées de la prostate.

Conisation : les conisations transversales des portions distales de l'apex (en rose sur l'illustration 6.2c) et du col vésical (en vert) d'épaisseur comprise entre 3 et 5 mm sont pratiquées.

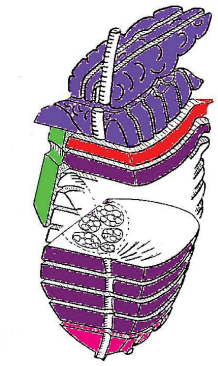
Découpe de la pièce de prostatectomie : le reste de la glande (en violet sur l'illustration 6.2c) est ensuite coupé en totalité en tronçons de 6 mm d'épaisseur (illustration 6.2d) selon un plan perpendiculaire à la surface rectale, correspondant au plan utilisé pour l'imagerie IRM, à l'aide d'une machine dédiée (illustration 6.2d) conçue en interne. Cette tâche est réalisée au service d'histopathologie de l'hôpital Edouard Herriot.

Les tranches sont mises dans des cassettes et refixées pendant au moins 48 heures avant de subir une étape de déshydratation (dans des bains de formol et de méthyl) puis d'être incluses en paraffine.

Enfin, les blocs de paraffine qui en résultent sont coupés au microtome (illustration 6.2f). Les rubans obtenus sont étalés sur lames et colorés à l'Hématoxyline-Eosine-Safran (HES) (Hématoxyline-Eosine-Safran) (voir figure 6.2g).

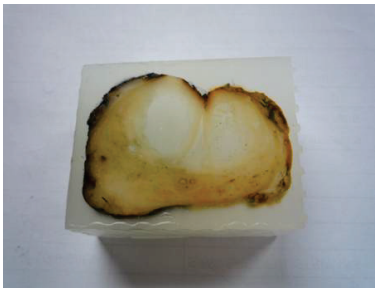
(a) Prostate fraîche

(b) Prostate encrée

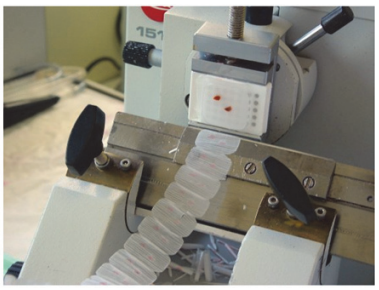


(c) Schéma de Stanford

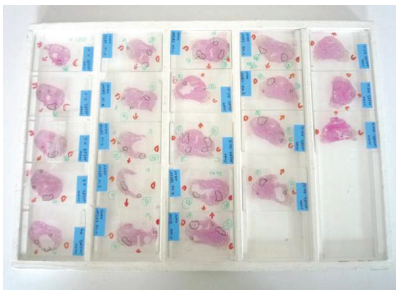
(d) appareil de découpe



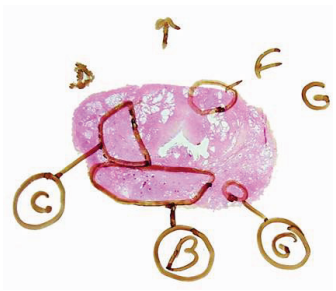
(e) Tranche de 6-mm



(f) Microtome



(g) Caissette



(h) Coupe histologique annotée

Figure *6.2* – *Différentes étapes de la préparation de la pièce de prostatectomie. (a) Prostate fraîche. (b) Prostate après orientation et encrage (face antérieure en rouge, lobe droit en jaune, lobe gauche en noir). La prostate est ensuite fixée dans du formol et conisée en suivant (c) le schéma de coupe de Stanford. (d) Un appareil dédié est utilisé pour découper la prostate perpendiculairement à la surface rectale de l'apex à la base (e) en tranches de 6 mm d'épaisseur. Les tronçons sont ensuite mis en cassettes et inclus dans la paraffine. Ces blocs de paraffine sont découpés au (f) microtome tous les 500 microns jusqu'à épuiser le bloc au moins de moitié. On étale alors les rubans obtenus sur lames. Après coloration au HES, on obtient ainsi un (g) ensemble de coupes (au moins tous les 3 mm). (h) Chaque coupe histologique est alors analysée au microscope et annotée par un anatomo-pathologiste.*

Analyse des coupes histologiques

Une anatomo-pathologiste, ayant 10 ans d'expérience, analyse au microscope l'ensemble de la pièce histologique issue de la prostatectomie radicale. Notons que les résultats de

l'imagerie pré-chirurgicale ne sont pas communiqués à ce stade. Sur chacune des lames histologiques où elles sont visibles, toutes les zones malignes sont contourées (voir illustrations 6.2h, et 2.4, page 21). Les tumeurs ne sont considérées comme telles que si leur surface dans le plan est au minimum de 2 mm x 2 mm, si elles sont visibles sur au moins 2 coupes et ont un score de Gleason ≥ 5 (lire section 2.4). Les lésions malignes séparées de moins d'1 mm l'une de l'autre dans le même plan, avec la même architecture et le même score de Gleason ont été considérées comme faisant partie de la même tumeur. Les régions de tissus remarquables et leurs caractéristiques (prostatite, inflammation, néoplasie intraépithéliale de la prostate, etc) sont également mises en évidence.

A la fin de l'analyse, un jeu de coupes séparées de 3 mm est sélectionné parmi l'ensemble des lames réalisées, permettant ainsi une comparaison directe avec les coupes IRM pour l'analyse de la corrélation anatomo-radiologique (voir photo 6.2g).

6.5 Corrélation anatomo-radiologique

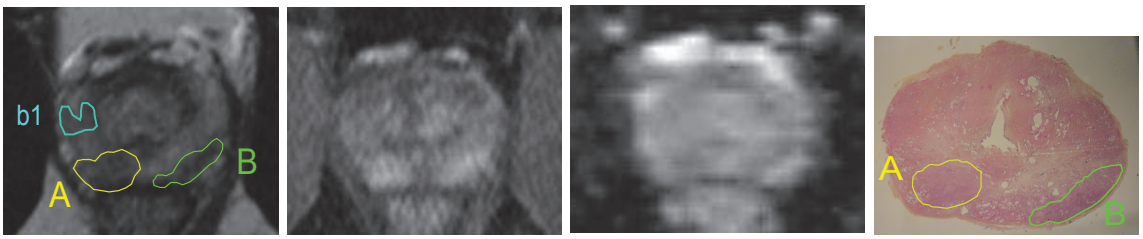
Les données issues de l'analyse en aveugle des images IRM (section 6.3) et celles issues de la lecture des lames histologiques (section 6.4) sont revues lors d'une troisième séance de lecture par les deux radiologues, un anatomo-pathologiste et moi-même, travaillant en consensus. En s'aidant notamment de repères anatomiques (urètre, capsule, canaux éjaculateurs, kystes, etc), toutes les zones malignes repérées sur les coupes histologiques sont reportées précisément sur les images IRM-mp. On note que, suivant la référence histologique, la lésion entière est contourée et pas seulement la région de forte anormalité (pic de réhaussement sur la DCE par exemple).
L'analyse de la pièce de prostatectomie, utilisée comme "vérité terrain", permet la délimitation *a posteriori*, sur les images IRM (voir la figure 6.3), des zones de :

- tissu malin (VP ou FN de l'analyse radiologique),
- tissu bénin présentant un signal anormal sur les images IRM (FP ou VN de l'analyse radiologique),
- tissu bénin présentant une apparence normale sur les images IRM (VN de l'analyse radiologique).

Comme cela a été précédemment réalisé par Vos *et coll.* [133], nous ferons référence à ces trois différentes classes de tissus en utilisant les abréviations :

- **{M}** : pour tissu **M**alin,
- **{NS}** : pour **N**on malin mais **S**uspect et,
- **{N}** : pour **N**ormal.

La classe {NS} correspond aux régions bénignes mais qui présentent un signal IRM suspect : réhaussement précoce sur la DCE, hyposignal significatif sur les images T2-w et/ou ADC, asymétrie gauche/droite notable etc. Il peut s'agir de faux positifs (FP) d'un ou des deux radiologues lors de l'analyse en aveugle (section 6.3) ou de tissus pour lesquels ils se sont longuement interrogés (et qui pourraient notamment être sources de FP pour d'autres radiologues, en particulier juniors). La classe {NS} comprend des régions tota-



(a) T2-w **(b)** DCE **(c)** ADC **(d)** coupe histologique

Figure *6.3* – *IRM de Prostate : (a) T2-w axial, (b) DCE (45 s après injection de Gd-DOTA) et (c) carte ADC obtenue sur un patient âgé de 53 ans avec (d) la coupe histologique numérisée correspondante. Les cancers vérifiés à l'histologie (lettres majuscules A et B) sont reportés sur les images IRM ainsi que les tissus jugés visuellement suspects à l'analyse radiologique (lettre minuscule b1). Des régions d'intérêt (ROIs) de tissus sains sont extraites automatiquement pour couvrir le reste de la zone périphérique (ZP).*

lement saines pour lesquelles l'histologie n'apporte aucune explication et des régions de tissus bénins correspondant par exemple à des foyers d'inflammation, d'atrophie, de néoplasie intraépithéliale de bas grade, de glandes kystisées, de remaniements myomateux (etc) ou à des traces de sang suite à des biopsies pré-IRM.
Lors de cette séance de corrélation, un total de 42 ROIs malignes {M} a été contouré dans la ZP ainsi que 49 ROIs normales mais d'apparence suspecte {NS} ; 124 cibles bénignes et d'apparence normale {N} ont été aléatoirement contourées dans la ZP saine restante (voir tableau 6.2).

La segmentation manuelle a été réalisée par les radiologues en utilisant la station de visualisation d'images open-source OsiriX® (Genève, Suisse).

Notation	**Nb. d'entités**	**Description**
M (Malin)	42	Régions malignes. L'histologie a révélé la présence de lésions cancéreuses sur toute la surface (Gleason ≥ 5)
N (Normal)	124	Régions saines et visuellement normales à l'IRM. Aucune trace de cellules cancéreuses n'est visible à l'analyse histologique
NS (Non malin mais Suspect)	49	Régions bénignes mais qui présentent un signal suspect à l'IRM (réhaussement précoce, hyposignal T2-w ou ADC, asymétrie etc). Il peut s'agir de régions totalement saines pour lesquelles l'histologie n'apporte aucune explication ou de tissus bénins correspondant par exemple à des foyers d'inflammation, de néoplasie intraépithéliale de bas grade, de glandes kystisées, de remaniements myomateux etc ou à des traces de sang suite à biopsies.

Table *6.2* – *La base de données contient les images IRM-mp de 30 patients sur lesquelles ont été annotées, dans la ZP, 215 ROIs réparties en : 42 ROIs malignes {M}, 49 ROIs normales mais d'apparence suspecte {NS} et 124 cibles bénignes et d'apparence normale {N}.*

6.6 Conclusion

L'étude proposée repose sur une base de données cliniques riche, annotée de manière fiable et exhaustive. Elle est composée des images IRM multi-paramétrique (IRM-mp) T2-pondérées (T2-w), de diffusion (carte de coefficient apparent de diffusion, ADC) et de perfusion (*dynamic contrast enhanced*, DCE) acquises sur 30 patients au préalable à une prostatectomie radicale. Les pièces anatomiques issues de la prostatectomie sont préparées et découpées de façon à permettre une mise en correspondance avec les images IRM (plan de coupe et espacement entre coupes similaires). Les coupes histologiques obtenues sont analysées afin de localiser précisément les foyers malins. Anatomo-pathologistes, radiologues et chercheur se réunissent enfin pour constituer la vérité terrain. Cette séance de corrélation anatomo-radiologique consiste à reporter chacun des foyers malins repérés à l'histologie sur les images IRM-mp. Un ensemble de foyers bénins est également contouré afin de permettre l'apprentissage sur les deux classes de tissus (sain/pathologique). On distinguera dans la suite les foyers sains d'apparence normale à l'IRM (notés {N}) des foyers sains d'apparence suspecte (notés {NS}), sources de faux positifs (FP).

Chapitre 7

Description du système de classification

7.1 Introduction

Comme nous l'avons souligné chapitre 4, les systèmes CAD reposent sur un schéma de principe standardisé que nous rappelons ci-dessous (voir figure 7.1). Nous présentons dans ce chapitre les choix effectués à chacune des étapes.

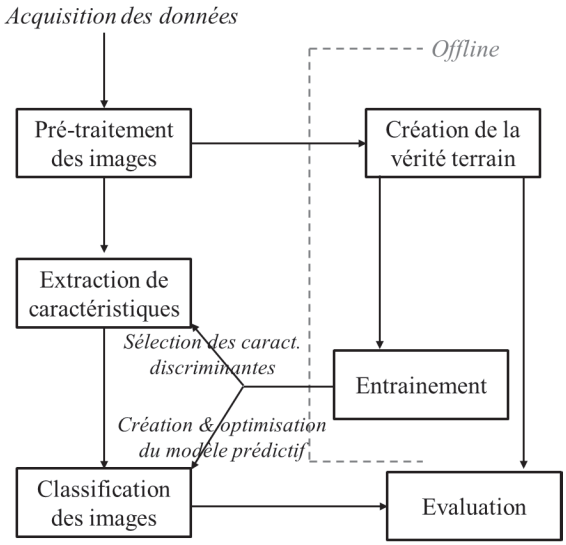


Figure 7.1 – *Schéma de principe des CAD supervisés*

7.2 Extraction des caractéristiques descriptives

L'étape d'extraction des caractéristiques est destinée à traduire de manière synthétique l'information pertinente contenue dans les ROIs. Le but est d'extraire les propriétés caractéristiques du signal mesuré sur les ROIs et de les exprimer sous la forme d'un vecteur, appelé vecteur de caractéristiques. La représentation obtenue servira de base aux étapes ultérieures : sélection des paramètres discriminants, section 7.3, et classification des données, section 7.5.

Faisant ainsi la synthèse des paramètres décrits dans la littérature (lire section 4.7), nous proposons d'extraire un ensemble de caractéristiques images, traduisant les propriétés structurelles des images en niveaux de gris, ainsi que des paramètres fonctionnels, extraits de la séquence DCE.

7.2.1 Pré-traitement des images

Correction des images

Les images IRM peuvent être entachées de nombreux artefacts liés aux imperfections du système d'acquisition. Il est souvent préférable de corriger ces artefacts afin d'obtenir des informations quantitatives avec une précision acceptable. On remarque que les artefacts sont souvent plus gênants dans le cas des antennes endo-rectales que dans celui des antennes de surface. Les inhomogénéités de la radiofréquence (RF) sont la source principale de variation artefactuelle du signal. En effet, les inhomogénéités spatiales de la RF émise, couplées avec des phénomènes d'off-résonance liés aux inhomogénéités de champ de B_0, induisent des variations spatiales des impulsions émises qui à leur tour influencent l'intensité du signal. Un certain nombre de pré-traitements des images IRM peuvent être réalisés directement lors de l'examen, grâce à des algorithmes implémentés par le constructeur (Siemens Medical Systems, Erlangen, Allemagne) sur la console d'acquisition de l'IRM 1.5 T de l'hôpital Edouard Herriot (Lyon, France). C'est le cas de la correction de l'inhomogénéité de champ (qui se traduit par des variations lisses d'intensité sur l'image), utilisée en routine clinique dans notre étude.

Normalisation des images

Pour permettre une comparaison entre les données des différents patients, le signal T2-w moyen de l'urine présente dans la vessie (en hyper-signal) est mesuré sur l'image et c'est le rapport "signal brut T2-w / signal moyen de l'urine" qui est calculé et utilisé [97]. De plus, hormis pour les caractéristiques dérivées de la modélisation pharmaco-cinétique du signal DCE proposée section 7.2.3, c'est le réhaussement relatif du signal, décrit par l'équation suivante, qui est préféré au signal T1-w brut [138] :

$$\Delta S = \frac{SI(t) - SI(t=0)}{SI(t=0)}, \tag{7.1}$$

où "t" fait référence au temps après l'injection de l'agent de contraste (AC), SI(t) au signal au temps t et SI(t=0) au signal avant l'injection de l'AC.

L'utilisation du réhaussement relatif nous affranchit des variations de sensibilité de l'antenne au cours des différentes acquisitions.

Notons que les cartes d'ADC sont déjà des données quantitatives.

7.2.2 Caractéristiques descriptives des images

Les caractéristiques descriptives des images extraites sont de quatre types : analyse directe des niveaux de gris, calcul de paramètres statistiques de premier ordre, de texture (statistique du second ordre) et de gradients sur ces niveaux de gris. On notera que pour la séquence DCE, seul un temps d'acquisition (t = 1 min) est utilisé pour l'extraction de ces caractéristiques [1]. Sur une acquisition totale de 3 minutes, ce temps correspond au temps moyen d'arrivée au pic de réhaussement ($SI_{95\%}$) pour les tissus malins (et présente donc un meilleur contraste).

Niveaux de gris des images

Les valeurs d'intensité des images mesurées pour chacune des trois séquences IRM sont utilisées en tant que telles.

Paramètres statistiques du premier ordre

Les caractéristiques statistiques traduisent des informations globales sur le signal de la région étudiée, qui est considéré alors comme une variable aléatoire réelle. Ces informations sont indépendantes de la géométrie et de la répartition spatiale des valeurs des pixels. Ces mesures sont calculées, pour chaque pixel de l'image, sur une fenêtre locale glissante de dimension 9x9-pixels centrée sur le pixel d'intérêt. Ces caractéristiques statistiques (détaillées dans le tableau 7.1) incluent la moyenne, la médiane, l'écart type et la déviation absolue moyenne.

Paramètres de texture

Des mesures de texture du second ordre sont également extraites. Contrairement aux mesures du premier ordre décrites ci-dessus, les caractéristiques de texture reflètent la distribution spatiale des niveaux de gris au sein de la région considérée. Nous avons considéré ici la famille de caractéristiques de texture issue de la matrice de co-occurence GLCM (de l'anglais *Grey Level Co-occurence Matrix*). Ces paramètres permettent de prendre en compte la relation entre les groupes de deux pixels voisins dans une région particulière.

Le calcul des paramètres repose sur celui de la matrice de co-occurrence des niveaux de gris. C'est une matrice de dimension N_g x N_g où N_g correspond au nombre de niveaux de

1. afin de limiter les temps de calcul et l'encombrement de la mémoire

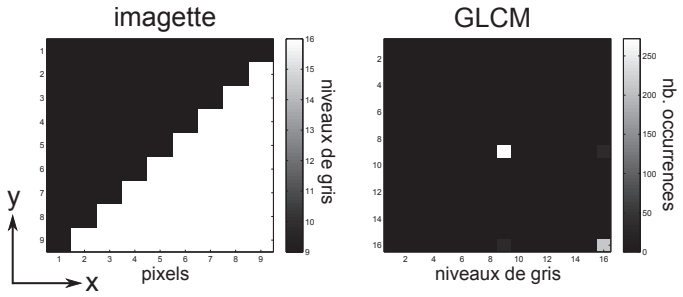
gris présents dans l'image. Chaque élément $GLCM_{\theta,d}(i,j)$ $(i,j = 1 \dots N_g)$ de la matrice GLCM est obtenu en dénombrant le nombre de fois qu'un pixel de niveau de gris i est voisin d'un pixel de niveau de gris j à la distance $d = (dx, dy) = (d\ cos\theta\ ,\ d\ sin\theta\ )$. Notons que cette valeur est normalisée par le nombre total de comparaisons. Appelons $p_{\theta,d}(i,j)$ cette valeur normalisée. L'élément $p_{\theta,d}(i,j)$ représente l'estimation de la probabilité de transition d'un niveau de gris à un autre selon la direction donnée par l'angle θ, avec une distance inter-pixels d bien définie. On peut construire autant de matrices de co-occurrence qu'il y a de définitions d'adjacence (ordonnée/non ordonnée, horizontale/verticale/diagonale etc).

Dans notre travail, le choix de θ s'est limité aux directions de valeurs 0, 45, 90 et 135 degrés et celui de la distance d à 1 pixel. Les couples (i,j) ne sont pas orientés (la matrice de co-occurrence est symétrique). Nous proposons d'annuler la dépendance en direction par la combinaison de toutes les matrices $p_{\theta,d}$, pour la distance donnée d, en une seule matrice normalisée p_d :

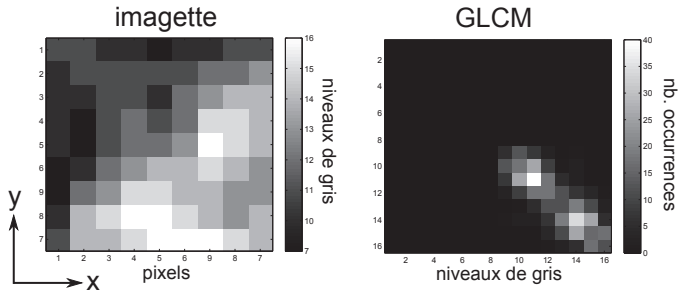
$$p_d = \frac{\sum_\theta p_{\theta,d}}{4} \text{ avec } \theta\ = 0, 45, 90 \text{ et } 135°. \tag{7.2}$$

Une première étape de réduction des niveaux de gris (ici à 16 niveaux de gris) a été opérée afin de réduire les temps de calcul, l'occupation mémoire et de limiter l'influence du bruit.

Haralick [47] propose un ensemble de 19 attributs de texture décrivant la répartition des niveaux de gris dans l'image, calculés à partir de la matrice GLCM. D'autres auteurs [20, 113] ont également étudié et étendu certains de ces attributs. Les 22 paramètres finalement exploités dans cette étude sont présentés dans le tableau 7.2. Comme l'illustre la figure 7.2, ils caractérisent l'homogénéité, la transition des niveaux de gris et la présence de structures organisées.



(a) Imagette présentant 2 plages homogènes et matrice GLCM correspondante



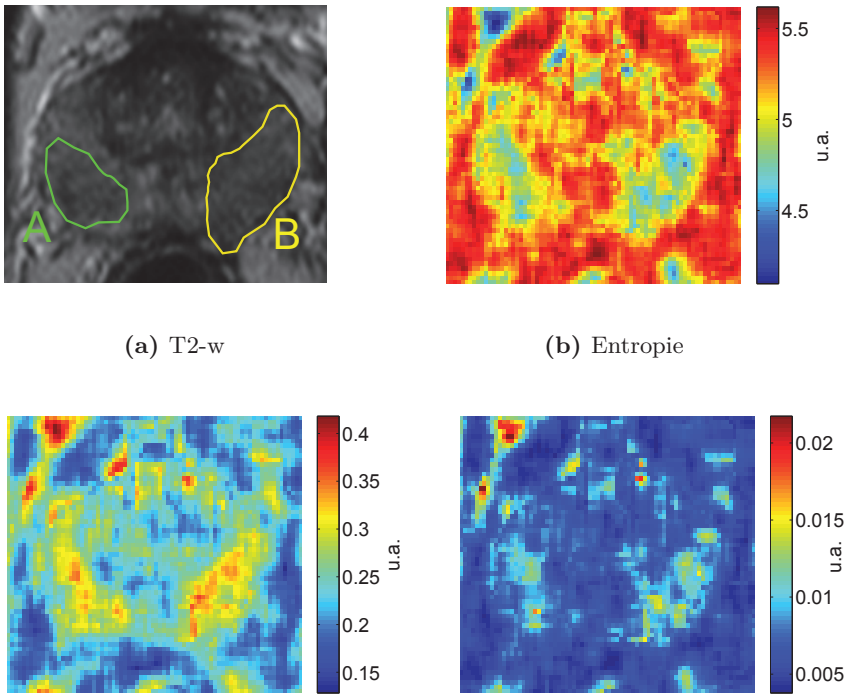
(b) Imagette bruitée et matrice GLCM correspondante

Figure 7.2 – *Exemples de deux imagettes présentant des structures (contraste, bruit, etc) différentes.*
On représente, sous forme d'images, les matrices de coocurrences (GLCM) associées (obtenues pour une décomposition en N_g=16 niveaux de gris, d=1, moyennées sur 4 directions θ=0, 45, 90, 135°).
Les paramètres extraits sont listés dans les tableaux ci-dessous (la signification des abréviations est donnée dans le tableau 7.2).

image	***autoc***	***contr***	***corrm***	***corrp***	***cprom***	***cshad***	***dissi***	***energ***	***entro***	***homo***
(a)	*155*	*5.6*	*0.77*	*0.77*	*2167*	*64*	*0.8*	*0.4*	*1*	*0.88*
(b)	*161*	*1.9*	*0.77*	*0.77*	*367*	*0.4*	*1*	*0.03*	*3.6*	*0.57*

image	***maxpr***	***sosv***	***savg***	***svar***	***sent***	***dvar***	***dent***	***inf1***	***inf2***	***inv***	***indn***	***indmn***
(a)	*0.5*	*158*	*24*	*583*	*0.96*	*5.6*	*0.4*	*-0.5*	*0.7*	*0.9*	*1*	*1*
(b)	*0.07*	*162*	*25*	*522*	*2.6*	*1.9*	*1.2*	*-0.2*	*0.8*	*0.6*	*0.9*	*1*

Exemple : L'entropie (entro) est une mesure du désordre, elle vaut 1 pour l'image (a) et 3.6 pour l'image (b), plus bruitée ; au contraire, la mesure de contraste (contr) vaut 5.6 pour l'image (a) qui présente une frontière marquée entre deux plages homogènes de valeurs éloignées, contre 1.9 pour l'image (b) qui présente beaucoup de faibles variations locales.



(a) T2-w **(b)** Entropie

(c) *Inverse difference* **(d)** Energie

Figure 7.3 – *Exemples de cartes de caractéristiques de texture : (a) Image T2-w normalisée avec les cancers A et B contourés par les experts, (b) Carte des mesures d'entropie (u.a.), (c) Carte d'Inverse difference (u.a.), (d) carte des mesures d'energie (u.a.). Les régions malignes A et B sont mises en évidence par les caractéristiques extraites.*

Caractéristique	Description
Moyenne (moy)	$\frac{1}{\lvert\mathcal{V}(x)\rvert}\sum_{y\in\mathcal{V}(x)}\mathcal{I}(y)$
Médiane (med)	valeur m telle que le nombre de pixels appartenant à $\mathcal{V}(x)$ de valeur $\geq$ m est égal au nombre de pixels de valeur $\leq$ m
Ecart type (std)	$\sqrt{\frac{1}{\lvert\mathcal{V}(x)\rvert}\sum_{y\in\mathcal{V}(x)}(\mathcal{I}(y)-\mathcal{I}(x))^2}$
Déviation absolue moyenne (var)	$\lvert\,\mathcal{I}(x)-(\frac{1}{\lvert\mathcal{V}(x)\rvert}\sum_{y\in\mathcal{V}(x)}\mathcal{I}(y))\,\rvert$

Table 7.1 – *Liste des 4 paramètres statistiques du 1er ordre extraits sur une fenêtre $\mathcal{V}(x)$ de dimension 9x9 pixels centrée sur le pixel d'intérêt x. La notation $\mathcal{I}(y)$ fait référence à la valeur du pixel y de l'image $\mathcal{I}$.*

Caractéristique	**Description**
Autocorrelation *(autoc)*	$\sum_{i,j} ijp(i,j)$
Contrast *(contr)*	$\sum_{i,j} \lvert i-j\rvert^2 p(i,j)$
Correlation *(corrm)* [20]	$\frac{\sum_{i,j}(i-\mu_x)(j-\mu_y)p(i,j)}{\sigma_x\sigma_y}$
Correlation *(corrp)* [47]	$\frac{\sum_{i,j} ijp(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$
Cluster Prominence *(cprom)*	$\sum_{i,j}(i+j-\mu_x-\mu_y)^4 p(i,j)$
Cluster Shade *(cshad)*	$\sum_{i,j}(i+j-\mu_x-\mu_y)^3 p(i,j)$
Dissimilarity *(dissi)*	$\sum_{i,j} \lvert i-j\rvert p(i,j)$
Energy *(energ)*	$\sum_{i,j} p(i,j)^2$
Entropy *(entro)*	$-\sum_{i,j} p(i,j)\log(p(i,j)+\epsilon) = \mathrm{H_{XY}}$
Homogeneity *(homo)*	$\frac{\sum_{i,j} p(i,j)}{1+\lvert i-j\rvert^2}$
Maximum probability *(maxpr)*	$\max(p(i,j))$
Sum of squares : Variance *(sosv)*	$\sum_{i,j}(i-\mu)^2 p(i,j)$
Sum average *(savg)*	$\sum_{k=2}^{2N} kp_{i+j}(k)$
Sum variance *(svar)*	$\sum_{k=2}^{2N}(k-\mathrm{H_{X+Y}})p_{i+j}(k)$
Sum entropy *(sent)*	$-\sum_{k=2}^{2N} p_{i+j}(k)\log(p_{i+j}(k)+\epsilon) = \mathrm{H_{X+Y}}$
Difference variance *(dvar)*	$\sum_{k=0}^{N-1} k^2 p_{i-j}(k)$
Difference entropy *(dent)*	$-\sum_{k=0}^{N-1} p_{i-j}(k)\log(p_{i-j}(k)+\epsilon)$
Information measure of correlation1 *(inf1)*	$\frac{\mathrm{H_{XY}}-\mathrm{H_{XY1}}}{\max(\mathrm{H_X},\mathrm{H_Y})}$
Information measure of correlation2 *(inf2)*	$\sqrt{1-\exp^{-2(\mathrm{H_{XY2}}-\mathrm{H_{XY}})}}$
Inverse difference *(inv)*	$\sum_{i,j} \frac{p(i,j)}{1+\lvert i-j\rvert}$
Inverse difference normalized *(indn)*	$\sum_{i,j} \frac{p(i,j)}{1+\lvert i-j\rvert/N}$
Inverse difference moment normalized *(indmn)*	$\sum_{i,j} \frac{p(i,j)}{1+(\lvert i-j\rvert/N)^2}$

Table 7.2 – *Liste des 22 caractéristiques de texture issues de la GLCM, où :*
$p(i,j)$ désigne la probabilité de transition d'un niveau de gris i à un autre j, c'est-à-dire que p est la GLCM normalisée.
μ_x, μ_y, σ_x et σ_y désignent les moyennes et écarts-type des densités partielles de probabilité p_x et p_y.
p_{i+j} (respectivement p_{i-j}) désigne la somme des probabilités $p(i,j)$ dont la somme (respectivement la différence) des coordonnées vaut $i+j$ (respectivement $i-j$).
$\mathrm{H_X}$ et $\mathrm{H_Y}$ désignent les entropies des densités de probabilité partielles p_x et p_y.
$\mathrm{H_{XY1}} = \sum_{i,j} p(i,j)\log(p_\mathrm{x}(i)p_\mathrm{y}(j))$ et $\mathrm{H_{XY2}} = \sum_{i,j} p_\mathrm{x}(i)p_\mathrm{y}(j)\log(p_\mathrm{x}(i)p_\mathrm{y}(j))$ sont deux estimations d'entropie jointe (i.e. d'information mutuelle).
Du fait de la précision limitée des calculs numériques, il est utile d'introduire une valeur infinitésimale $\epsilon > 0$ afin de garantir la stricte positivité de l'argument du log.

Paramètres de Gradients

Trois différents opérateurs de gradients ont été utilisés : les filtres de Sobel et de Kirsch (dont les noyaux sont détaillés dans le tableau 7.3) et un gradient numérique (approche par différences finies). Les gradients directionnels sont évalués dans les trois directions principales (horizontal, vertical, diagonal). Des images caractérisant la magnitude et l'amplitude du gradient sont également calculées. Le tableau 7.3 présente les 16 paramètres extraits.

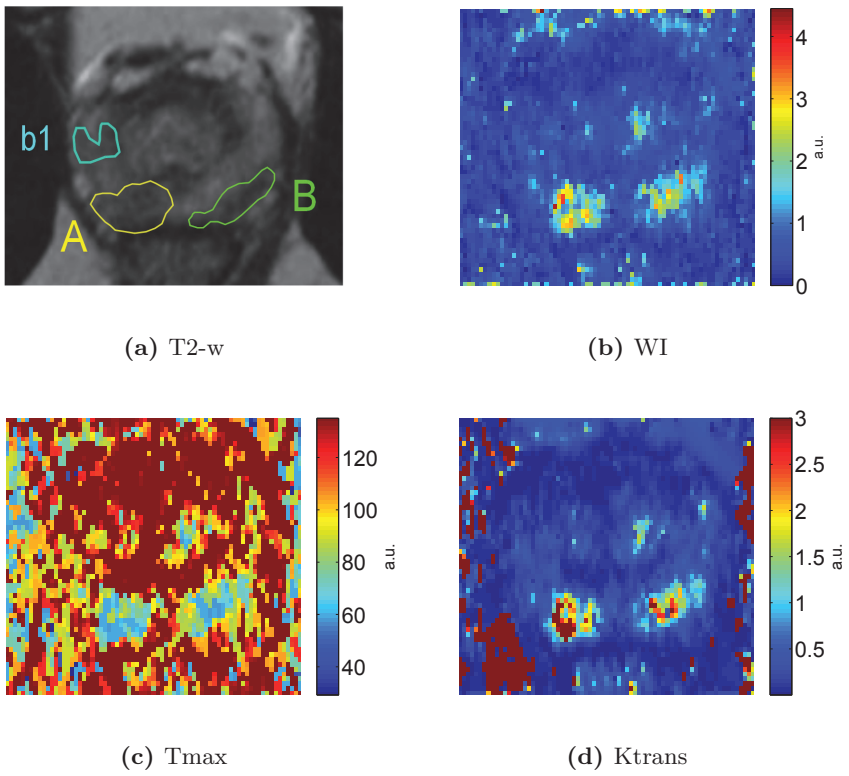
Gradient	**Description**
Sobel	$G_1 = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}$, $G_2 = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$, $G_3 = \begin{bmatrix} 0 & +1 & +2 \\ -1 & 0 & +1 \\ -2 & -1 & 0 \end{bmatrix}$, $G_4 = \begin{bmatrix} -2 & -1 & 0 \\ -1 & 0 & +1 \\ 0 & +1 & 2 \end{bmatrix}$,
Kirsch	$G_1 = \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix}$, $G_2 = \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}$, $G_3 = \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}$, $G_4 = \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix}$,
	$G_5 = \begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix}$, $G_6 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}$, $G_7 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix}$, $G_8 = \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix}$,
Différences finies centrées	$\delta_x \mathcal{I}(x,y) = \frac{\mathcal{I}(x+1,y) - \mathcal{I}(x-1,y)}{2}$ $\delta_y \mathcal{I}(x,y) = \frac{\mathcal{I}(x,y+1) - \mathcal{I}(x,y-1)}{2}$

Table *7.3* – *Les images de gradient de Sobel et de Kirsch $\mathcal{G}_i$ sont obtenues par filtrage de l'image source $\mathcal{I}$ par les noyaux donnés ci-dessus.*
Pour chaque noyau G_i, l'image résultante $\mathcal{G}_i$ est obtenue par :
$\mathcal{G}_i$=$G_i \star \mathcal{I}$ où $\star$ est l'opérateur de convolution.
Des images caractérisant la magnitude et l'amplitude du gradient sont également calculées :
$\mathcal{G}_{norm} = \sqrt{\sum_{i \in Sobel} \mathcal{G}_i^2}$ et $\mathcal{G}_{max}$ telle que $\mathcal{G}(y)_{max} = \max_{i \in Kirsch} \mathcal{G}(y)_i$, pour tout pixel $y \in \mathcal{I}$.

7.2.3 Caractéristiques fonctionnelles

Différents types de caractéristiques "fonctionnelles", caractérisant la structure et la fonction vasculaire des tissus, peuvent être extraites de la séquence dynamique DCE. Comme nous l'avions vu au chapitre 3.5, page 42, dans les "cas d'école", les tissus malins se distinguent *a priori* des tissus bénins par une pente de réhaussement raide, un temps d'arrivée au pic rapide, un pic de réhaussement important et un lavage décroissant. Quelques travaux [93, 130, 133, 138] ont d'ores et déjà souligné l'intérêt de certains paramètres issus de la DCE pour la discrimination des tissus, nous encourageant dans cette voie.

Nous nous sommes intéressés dans cette étude à l'extraction 1) de caractéristiques "semi-quantitatives", obtenues par analyse directe de la forme de la courbe de réhaussement en agent de contraste (courbe tracée à partir des images DCE) et 2) de caractéristiques pharmaco-cinétique "quantitatives", obtenues par modélisation de la cinétique de réhaussement des agent de contraste (AC) en termes de paramètres physiologiques. La liste des paramètres extraits est donnée dans le tableau 7.4. Nous les décrivons ci-après.



(a) T2-w **(b)** WI

(c) Tmax **(d)** Ktrans

Figure 7.4 – *Exemples de cartes de caractéristiques fonctionnelles : (a) Image T2-w normalisée avec les cancers A et B contourés par les experts ainsi qu'un tissu bénin d'aspect suspect b1, (b) Carte du wash-in (u.a.), (c) carte de Tmax (u.a.), (d) carte Ktrans (u.a.). Les régions malignes A et B sont mises en évidence par les caractéristiques extraites.*

Caractéristiques semi-quantitatives de la cinétique de distribution de l'AC

L'analyse de la courbe de réhaussement dynamique du signal permet une première description simplifiée de la distribution de l'AC [84]. Elle fournit des paramètres empiriques tels que la pente, le temps d'arrivée de l'AC, sa pente d'élimination, le temps où le réhaussement est maximal. On s'intéresse ici directement aux variations du signal T1-w après injection du bolus. On extrait de la courbe de réhaussement (voir figure 7.5) les 13 paramètres suivants :

Valeurs de signal remarquables :

- signal de base, "baseline" (SI_0, acquis au temps 0),
- pic du contraste (réhaussement maximal, SI_{max}),
- valeur à 95% du pic ($SI_{95\%}$),
- valeur à 5% du pic (SI_{onset}, début du réhaussement),
- amplitude ($A = SI_{pic} - SI_0$),
- signal de fin d'acquisition (SI_{end}).

Temps remarquables :

- temps d'arrivée au pic (T_{pic}),

– temps d'arrivée à 95% du pic (T_{max}),
– temps à 5% du pic (T_{onset}),
– temps de perfusion ($T_{95\%} - T_{onset}$).

Pentes :

– pente de réhaussement, "wash-in" (WI(% s^{-1}) = $\frac{SI_{95\%}-SI_{onset}}{T_{95\%}-T_{onset}}$),
– pente du lavage, "wash-out" (WO(% s^{-1}) = $\frac{SI_{end}-SI_{95\%}}{T_{end}-T_{95\%}}$).

Aire sous la courbe : aire sous la courbe de réhaussement en produit de contraste (AUGC, de l'anglais *Area Under the Gadolinium Curve*).

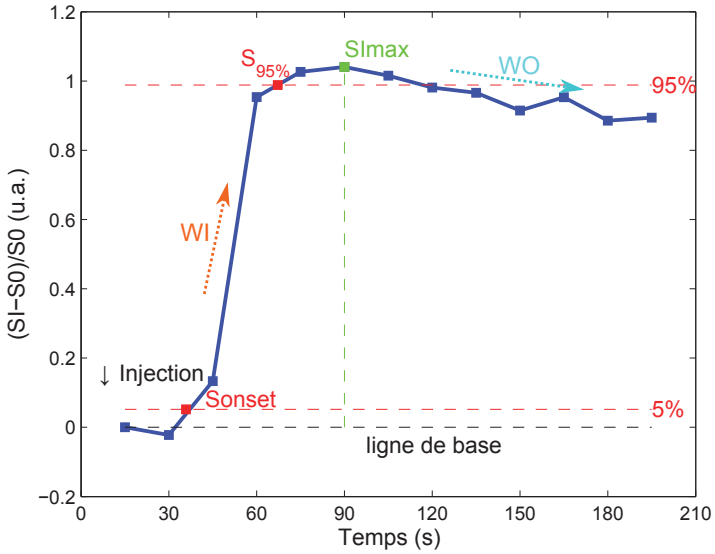


Figure 7.5 – *Un exemple de courbe temps-intensité du signal d'un pixel en IRM DCE. Quelques paramètres fonctionnels semi-quantitatifs extraits de la courbe de réhaussement de l'AC sont représentés : S_0, S_{onset}, $SI_{95\%}$, SI_{max}, WI et WO.*

On notera que ces valeurs sont à normaliser par la valeur correspondante calculée sur le signal de la fonction d'entrée artérielle (AIF, de l'anglais *Arterial Input Function*) afin de permettre une comparaison inter-patient. On normalise donc SI_0, SI_{pic}, $SI_{95\%}$, A, SI_{end}, WI, WO par S^{AIF}_{max} et AUGC par $AUGC^{AIF}$.

Modélisation pharmaco-cinétique

Pour quantifier la cinétique des AC en termes de paramètres physiologiques, il faut définir les éléments de la tumeur ou les structures du tissu et le processus fonctionnel qui affectent la distribution du traceur. Le processus de modélisation compartimentale consiste à représenter les différentes structures du tissu par un ensemble de compartiments et à définir des lois d'échanges entre ces différents compartiments. Ces échanges sont régis par un ensemble d'équations différentielles paramétrées par des constantes d'échanges. Il est classique de diviser le tissu en quatre espaces liquidiens : l'espace plasmatique vasculaire,

l'espace des éléments figurés du sang (globules rouges et blancs), l'espace extravasculaire/extracellulaire (EEE) et l'espace intra-cellulaire. Un modèle à quatre compartiments est cependant très difficile à résoudre, en pratique on cherche donc un modèle simplifié. Notre étude fait l'hypothèse d'un modèle à deux compartiments décrit sur l'illustration 7.6.

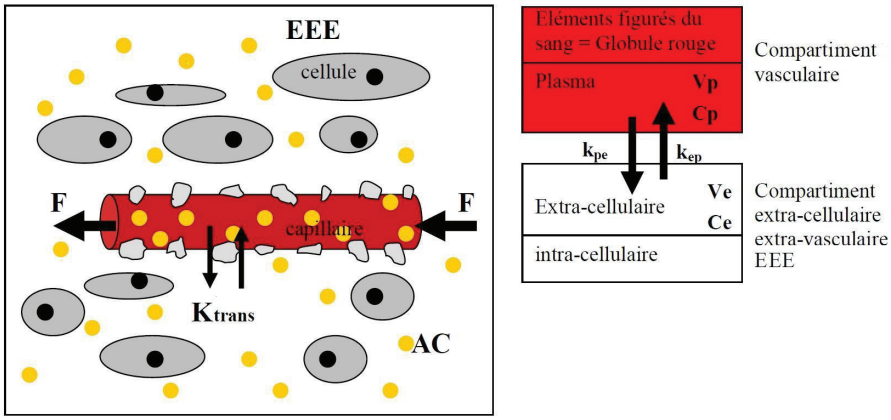


Figure 7.6 – *Représentation schématique du modèle bi-compartiments, composé du compartiment plasmatique vasculaire (en rouge) et du compartiment EEE avec les cellules (espace intra-cellulaire, en gris) et l'interstitium (en blanc). Les molécules d'AC sont représentées en jaune. Source : Calmels [14].*
Les variables impliquées dans la modélisation de l'échange de l'AC entre ces deux compartiments sont :
- v_e et v_p, les fractions volumiques de l'EEE et du plasma respectivement, comprises entre 0 et 1. Dans le modèle utilisé, la contribution du compartiment vasculaire au signal est considérée négligeable, v_p n'est pas estimée.
- k_{pe} et k_{ep}, constantes de flux du plasma vers l'EEE et de l'EEE vers le plasma respectivement, en min^{-1} (supposés identiques : $k_{pe} = k_{ep}$).
- $Ktrans = k_{pe}v_e$ constante de transfert, en min^{-1}.

Nous avons choisi d'utiliser la modélisation introduite par Tofts [125], la plus répandue dans la littérature, qui propose de décrire l'évolution de la concentration $C(t)$ de l'AC en fonction du temps par la formule ci-dessous :

$$C(t) = Ktrans.\mathrm{AIF}(t) \star \exp\left(\frac{-Ktrans.t}{V_e}\right) \tag{7.3}$$

où :
t = temps
C(t) = concentration en AC du tissu au temps t
AIF(t) = fonction d'entrée artérielle (concentration en AC dans l'artère au temps t)
Ktrans = constante de transfert entre le compartiment vasculaire et l'EEE (min^{-1})
Ve = volume de l'EEE (%)
$\star$ = opérateur de convolution

Les figures 7.7b et 7.7c illustrent cette modélisation.

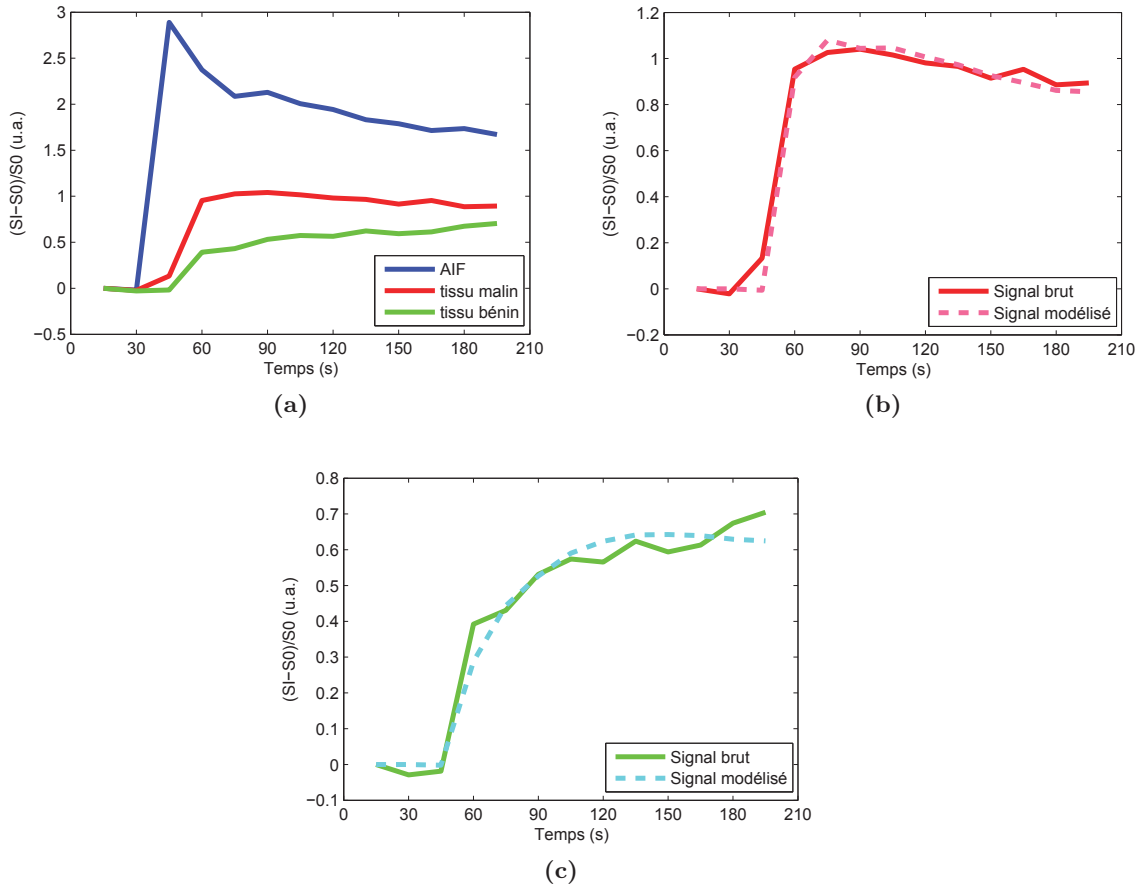


Figure 7.7 – *Modélisation pharmaco-cinétique du signal DCE. (a) Exemples de courbes de réhaussement mesurées sur trois ROIs : l'AIF (en bleu), une région pathologique (en rouge) et une région bénigne (en vert). Le réhaussement mesuré sur la région maligne est plus rapide et plus important que celui mesuré dans le tissu sain. Contrairement au tissu malin, la prise de contraste est continue dans le tissu sain (pas de wash-out). Les figures (b) et (c) illustrent l'ajustement du modèle de Tofts (en pointillés) sur les courbes correspondant au signal mesuré (en traits pleins).*

L'estimation de la perméabilité tissulaire Ktrans et du volume extracellulaire Ve peut alors être réalisée par minimisation de l'erreur quadratique entre les données mesurées $C(t)$ et le signal modélisé, le calcul s'effectuant itérativement à partir de la formule 7.3. Notre implémentation repose sur la Toolbox Optimization de Matlab® (MATLAB version 7.9.0, 2009, The MathWorks Inc., Natick, Massachusetts).
D'autres paramètres peuvent être évalués, notamment la constante de flux entre l'EEE et le compartiment vasculaire (Kep) que nous considérons dans notre étude :

$$\text{Kep} = \frac{Ktrans}{V_e} \text{ (en min}^{-1}\text{)}. \tag{7.4}$$

On note que dans la formule 7.3, il apparaît nécessaire d'estimer la concentration $C(t)$

en AC dans le tissu prostatique. Or nous n'avons pas directement accès à ces valeurs de concentrations. Nous avons donc fait l'hypothèse de proportionnalité entre la concentration en AC dans le tissu et le signal mesuré (qui se traduit directement par les valeurs de niveaux de gris de l'image). Afin de nous assurer de la validité de cette hypothèse sur nos données, nous l'avons vérifiée à l'aide d'une ceinture de tubes remplis de Gd-DOTA aux concentrations variables, installée sur un échantillon de 10 patients volontaires. Le détail de ce protocole de validation est donné section 6.2.3, page 89.

Remarque : cette modélisation ne prétend pas fournir la meilleure estimation possible des paramètres pharmaco-cinétiques et nous restons conscients des limitations induites par nos hypothèses. Notre objectif est de calculer, en utilisant au mieux les données cliniques à disposition, des caractéristiques descriptives nous permettant à terme de discriminer les tissus cancéreux des tissus bénins.

Caractéristique	**Description**
SI_0	signal de base, acquis à t=0
SI_{max}	pic de réhaussement
$SI_{95\%}$	valeur à 95% du pic
SI_{onset}	valeur à 5% du pic
A	amplitude du réhaussement
SI_{end}	valeur du signal en fin d'acquisition
T_{pic}	temps d'arrivée au pic
T_{max}	temps d'arrivée à 95% du pic
T_{onset}	temps d'arrivée à 5% du pic
T_{perf}	temps de perfusion
WI	pente de réhaussement (*wash-in*)
WO	pente de lavage (*wash-out*)
AUGC	aire sous la courbe de réhaussement
K_{trans}	constante de transfert (perméabilité)
v_e	volume de l'EEE
K_{ep}	constante de flux entre l'EEE et le compartiment vasculaire

Table 7.4 – *Liste des 16 paramètres de perfusion semi-quantitatifs et quantitatifs extraits des images DCE*

7.2.4 Modélisation des ROIs

Chacune des 145 caractéristiques descriptives décrites précédemment est extraite pour chacun des pixels de l'image (obtention de cartes de caractéristiques). Elles sont ensuite résumées par ROI. On calcule une série de mesures statistiques du premier ordre sur les valeurs obtenues sur la ROI. Il s'agit de la moyenne, de la médiane et des 1er ou 3^{e} quartiles. Ces deux dernières mesures sont particulièrement adaptées pour les ROIs présentant un motif hétérogène ou un spot (fort hypo ou hyper-signal) localisé dans la région ; elles sont, de plus, moins sensibles aux valeurs extrêmes (*outliers*).

7.3 Sélection des caractéristiques discriminantes

7.3.1 Introduction

On rappelle que notre problème initial se présente sous la forme d'une base d'apprentissage constituée de n données étiquetées $(\mathbf{x}_i, l_i)_{i=1...n}$ où $\mathbf{x}_i = [x_i^1, x_i^2 \dots x_i^d] \in \mathbb{R}^d$ est le vecteur de caractéristiques correspondant à la donnée i, x_i^j la j^{e} caractéristique de la données i, et $l_i \in \{-1, 1\}$ la classe de cette donnée (sain/malin).

On s'intéresse ici au vecteur $[X^1, X^2 \dots X^d]$, où X^j représente la j^{e} caractéristique, et on cherche à réduire sa dimension d en éliminant les variables non-informatives afin : 1) d'éviter le sur-apprentissage, 2) de diminuer la complexité algorithmique et accélérer le processus d'apprentissage et 3) d'améliorer l'interprétabilité du modèle.

En effet, parmi cet ensemble de caractéristiques, il se peut que certaines soient redondantes voire corrélées. Or les performances des classifieurs sont fortement affectées par le choix des caractéristiques ; en effet, submerger le classifieur de caractéristiques non-informatives peut engendrer deux problèmes : d'une part, le classifieur peut devenir trop spécifique à la base qui lui a servi d'apprentissage (problème d'*overfitting*) en perdant sa capacité de généralisation et d'autre part il peut ne pas donner de l'importance à l'information réellement significative tant celle-ci est noyée dans la masse de données.

7.3.2 Méthodologie

Après une première étape qui permet d'éliminer les caractéristiques de variance nulle (et donc non-informatives) et les doublons (vecteurs de caractéristiques identiques), nous avons dans cette étude, procédé à une sélection *a priori* des caractéristiques par approche de type *filter* (lire la section 4.3, page 53). Notre méthode consiste tout d'abord à évaluer les caractéristiques de manière individuelle selon quatre critères (test statistique, information mutuelle, mRMRd et mRMRq) indépendants du classifieur et mesurant leurs performances intrinsèques. Ces différents critères peuvent alors être utilisés pour ordonner les caractéristiques selon le score obtenu. Une fois les caractéristiques ordonnées, plusieurs choix sont possibles. On peut par exemple sélectionner les caractéristiques en fonction d'un seuil s sur le critère considéré (pour le test statistique, p-valeur < 0.05 par exemple) ou se limiter aux k premières variables ordonnées, avec s et k à fixer.

Pour éviter d'avoir à fixer un seuil de manière arbitraire, nous avons choisi de sélectionner les caractéristiques, à partir de cette liste ordonnée, en fixant le seuil d'acceptation sur le critère de manière empirique. Une technique de recherche avant séquentielle est utilisée avec un classifieur pour sélectionner le sous-ensemble final de caractéristiques. En d'autres termes, on estime les performances du classifieur sur des sous-ensembles comportant un nombre croissant de caractéristiques (ordonnées selon le critère considéré) jusqu'à trouver le sous-ensemble qui maximise les performances de discrimination du classifieur, mesurées en terme d'aire sous la courbe ROC (AUC) (on commence par la caractéristique qui a le meilleur score, les deux meilleures, etc.)

Les critères évalués sont explicités dans les sections 7.3.3, 7.3.4 et 7.3.5 ci-après.

7.3.3 Test statistique (test *t*)

Une approche simple pour extraire les caractéristiques les plus significatives est de faire l'hypothèse d'indépendance des variables et de réaliser un test de Student (test t ou t-test). Une façon d'évaluer si une caractéristique descriptive est suffisamment informative pour la tâche de discrimination des tissus sains et malins est en effet de réaliser un test t permettant de quantifier la significativité (p-valeur) de la différence entre les valeurs moyennes prises par cette caractéristique sur ces deux populations. Les caractéristiques peuvent alors être ordonnées en fonction de la valeur absolue de la p-valeur en sortie du test statistique. Dans notre étude, nous utilisons le test t de Welch [137], adaptation du test de Student dans le cas où les variances des deux populations sont différentes.

7.3.4 Information Mutuelle (IM)

L'information mutuelle est une mesure de dépendance entre les distributions de deux populations (Fraser et Swinney [1986]). Soit deux variables aléatoires X et Y. Leur information mutuelle $\mathcal{I}(X;Y)$ est définie en fonctions des densités de probabilités $\mathbb{P}(X)$, $\mathbb{P}(Y)$ et $\mathbb{P}(X,Y)$ par :

$$\mathcal{I}(X;Y) = \int\int \mathbb{P}(X=x, Y=y) \log \frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(X=x)\mathbb{P}(Y=y)} dxdy.$$

Dans notre étude, on considère que les instances des variables aléatoires X et Y sont respectivement les valeurs de la i$^{\text{e}}$ caractéristique, $\mathbf{x}^i = [x_1^i, x_2^i, \dots x_n^i]^\top$, et les étiquettes de classe, $\mathbf{l} = [l_1, l_2, \dots l_n]^\top$. Nous proposons d'estimer la pertinence d'une caractéristique via l'information mutuelle $\mathcal{I}(\mathbf{x}^i;\mathbf{l})$, reflétant la dépendance entre la variable de classe et la valeur que prend la caractéristique. Plus l'information mutuelle est grande, "meilleure" est la caractéristique. Les probabilités $\mathbb{P}(x^i)$, $\mathbb{P}(l)$ et $\mathbb{P}(x^i, l)$ sont estimées par les fréquences des différentes valeurs possibles.

Notre implémentation s'est appuyée sur la toolbox MutualInfo (version 0.9, [88]).

7.3.5 *minimum-Redundancy, Maximum-Relevancy* (mRMRd, mRMRq)

La plupart des méthodes de type *filter* reposent sur le concept d'un ordonnancement simple, dans lequel les caractéristiques sont ordonnées selon un critère (information mutuelle, test t etc) puis sélectionnées en fonction de leur rang. Néanmoins, ce faisant, il est possible que des caractéristiques fortement corrélées, et donc redondantes, soient sélectionnées.

L'algorithme de *minimum-Redundancy, Maximum-Relevancy* (mRMR) proposé par Peng *et coll.* [88] vise à résoudre ce problème. L'idée de base est de sélectionner le jeu de caractéristiques de façon à :

- minimiser la redondance (**mR**) d'information entre les caractéristiques, en sélectionnant les caractéristiques qui sont les plus dissimilaires aux autres.
- maximiser la pertinence (**MR**), en sélectionnant les caractéristiques les plus corrélées à l'étiquette de classe.

Les auteurs utilisent l'information mutuelle pour calculer ces deux critères. La redondance et la pertinence de la caractéristique X^i sont évaluées par :

$$\text{Redondance}(i) = \frac{1}{d^2}\sum_{i,j=1}^{d} \mathcal{I}(\mathbf{x}^i;\mathbf{x}^j) \text{ et } \text{Pertinence}(i) = \frac{1}{d}\sum_{i=1}^{d}\mathcal{I}(\mathbf{x}^i;\mathbf{l}), \tag{7.5}$$

où $\mathcal{I}(\mathbf{x}^i;\mathbf{x}^j)$ est l'information mutuelle entre la i^e et la j^e caractéristique et $\mathcal{I}(\mathbf{x}^i;\mathbf{l})$ est l'information mutuelle entre la i^e caractéristique et l'ensemble des étiquettes de classes $\mathbf{l}$.

Le score d'une caractéristique est obtenu par combinaison de ces deux critères, par soustraction (**mRMRd**) ou par division (**mRMRq**) :

$$\text{mRMRd}(i) = \text{Pertinence}(i) - \text{Redondance}(i) \text{ et } \text{mRMRq}(i) = \frac{\text{Pertinence}(i)}{\text{Redondance}(i)}. \tag{7.6}$$

Plus les mRMRd et mRMRq sont grands, plus la caractéristique est pertinente (et moins elle est redondante).

7.4 Choix des classifieurs

7.4.1 Introduction

L'étape de classification vise à déterminer la *classe* d'une donnée (ROI ou pixel).

L'approche retenue est de type apprentissage supervisé : à partir d'une base d'exemples préalablement annotés (i.e. dont la classe $l = \pm 1$ est connue), on réalise une étape d'apprentissage qui va permettre de dégager une règle empirique donnant la classe l_i (ou une probabilité d'appartenance à cette classe) d'un exemple i à partir de ses caractéristiques $\mathbf{x}_i$. Il suffira alors, lorsque l'on veut connaître la classe d'une nouvelle donnée test, d'appliquer la règle empirique engendrée par apprentissage.

Au regard de l'état de l'art présenté section 4.7, nous avons choisi de tester et comparer les quatre algorithmes d'apprentissage suivants :

- le séparateur à vaste marge (SVM) ,
- l'analyse discriminante linéaire (ADL),
- les k-plus proches voisins (k-PPV),
- le classifieur naïf de Bayes (CNB),

qui, comme nous l'avons souligné section 4.4.2, reposent sur des approches d'apprentissage différentes, et présentent donc des avantages et des inconvénients spécifiques. L'idée est de pouvoir comparer de façon non biaisée, sur une même base de données, ces algorithmes largement répandus dans la littérature des CAD et de finalement sélectionner le plus performant pour notre système CADx.

Les différents algorithmes d'apprentissage sont décrits ci-dessous.

Rappel des notations :
On considère l'ensemble des couples $(\mathbf{x}_i, l_i)_{i=1,\dots n}$ où les $(\mathbf{x}_i)_{i=1,\dots n}$ sont les n données d'apprentissage décrites chacune par d caractéristiques, $\mathbf{x}_i = [x_i^1, x_i^2, \dots x_i^d] \in \mathbb{R}^d$, et les $(l_i)_{i=1,\dots n}$, avec $l_i \in \{-1, +1\}$, sont les étiquettes de classe correspondantes.

7.4.2 Normalisation des données

Les valeurs des caractéristiques descriptives extraites précédemment sont normalisées au moyen d'une transformation affine. Cette normalisation est telle que les distributions de valeurs de chacune des caractéristiques ont pour moyenne 0 et pour écart type 1 :

$$\hat{x}_i^j = \frac{x_i^j - \mu_j}{\sigma_i}$$

où x_i^j est la caractéristique j de l'exemple i, $\hat{x}_i^j$ est la version normalisée de x_i^j et μ_j, σ_i sont respectivement la valeur moyenne et l'écart type de la distribution de valeurs prises par la caractéristique j. Ceci permet de ramener les valeurs des différents types de caractéristiques dans le même intervalle, améliorant ainsi la stabilité du classifieur.

7.4.3 Le séparateur à vaste marge (SVM)

Soit $\mathcal{H}$ un hyperplan dans l'espace des caractéristiques $\mathbb{R}^d$. $\mathcal{H}$ peut s'exprimer de la façon suivante :

$$\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^d | \mathbf{w}^\top.\mathbf{x} + b = 0\}$$

avec $\mathbf{w}$ un vecteur normal à l'hyperplan, b un réel tel que $\frac{b}{\|\mathbf{w}\|}$ le décalage de l'hyperplan par rapport à l'origine (voir figure 7.8b). On cherche $\mathbf{w}$ et b tels que $\mathcal{H}$ soit un hyperplan séparateur de l'espace des caractéristiques permettant de classer les données $(\mathbf{x}_i)_{i=1...n}$ d'étiquettes $(l_i)_{i=1...n} \in \{-1, 1\}$ correctement selon leur position par rapport au séparateur, i.e. tels que, pour tout i, $l_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 0$.
Dans le cas de données linéairement séparables, il n'est pas possible de déterminer de manière unique un hyperplan séparateur en se basant sur le seul critère de minimisation du nombre d'observations mal classées (voir figure 7.8a).
V. Vapnik [128] a proposé, en 1998, un critère d'optimalité basé sur la "marge" pour séparer des classes linéairement séparables (généralisable à des frontières non linéaires *via* un changement d'espace). Maximiser la marge entre les classes revient à maximiser la plus petite distance séparant un point de l'espace des observations à l'hyperplan séparateur (voir figure 7.8b).

La distance d'un point $\mathbf{x}_i$ à l'hyperplan séparateur est donné par :

$$\frac{|\ \mathbf{w}^\top \mathbf{x}_i + b\ |}{||\ \mathbf{w}\ ||},$$

ce qui implique que la demi-marge (i.e. la distance entre l'hyperplan séparateur et les échantillons les plus proches) vaut :

$$\min_{1 \leq i \leq n} \frac{|\ \mathbf{w}^\top \mathbf{x}_i + b\ |}{||\ \mathbf{w}\ ||} = \min_{1 \leq i \leq n} \frac{l_i(\mathbf{w}^\top \mathbf{x}_i + b)}{||\ \mathbf{w}\ ||},$$

la dernière égalité étant vraie car l'hyperplan $\mathcal{H}$ est séparateur.

Afin de faciliter l'optimisation, quitte à remplacer $\mathbf{w}$ et b par $\frac{\mathbf{w}}{\min\limits_{1 \leq i \leq n}(|\mathbf{w}^\top \mathbf{x}_i + b|)}$ et $\frac{b}{\min\limits_{1 \leq i \leq n}(|\mathbf{w}^\top \mathbf{x}_i + b|)}$ respectivement, on peut supposer que

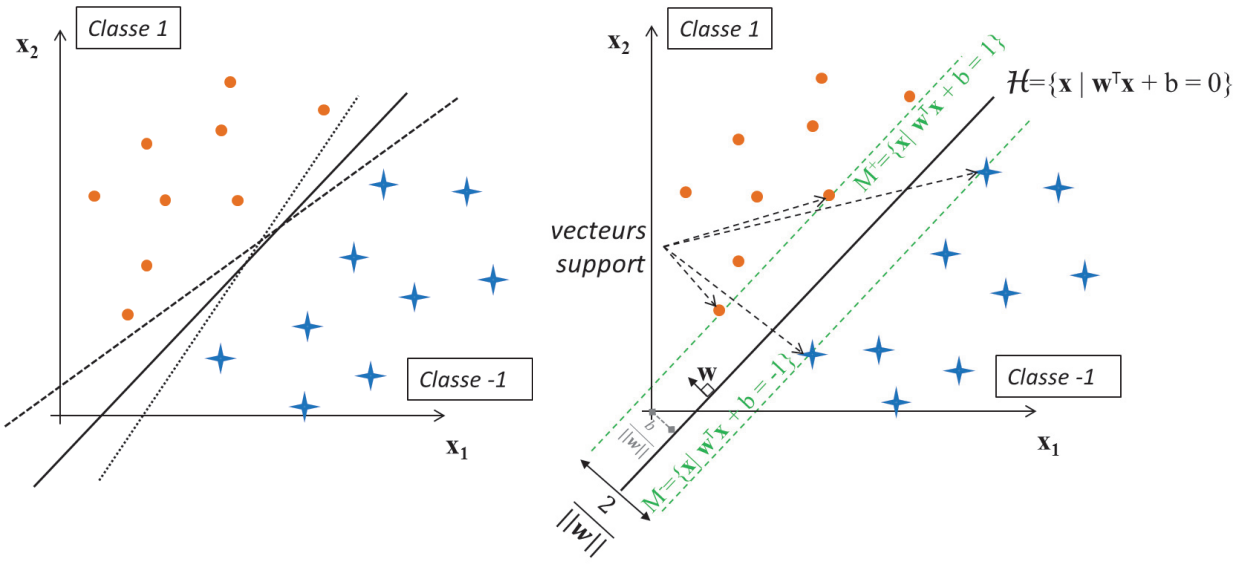
$$\min_{1 \leq i \leq n}(|\ \mathbf{w}^\top \mathbf{x}_i + b\ |) = 1,$$

c'est-à-dire que la fonction discriminante vaut +1 pour les points situés sur la marge et d'étiquette "+1" et -1 pour les points situés sur la marge et d'étiquette "-1" : ces points sont appelés les vecteurs support.

Désormais, avec ce choix de "normalisation" sur $\mathbf{w}$ et b, la marge vaut $\frac{1}{||\mathbf{w}||}$.

Notons que maximiser la marge $\frac{1}{\|\mathbf{w}\|}$ est équivalent à minimiser $\|\ \mathbf{w}\ \|$.

Dans le cas où les exemples sont linéairement séparables, la recherche de l'hyperplan



(a) Hyperplans séparateurs **(b)** Hyperplan séparateur maximisant la marge

Figure 7.8 – *Hyperplans séparateurs et marge maximale (cas linéairement séparables). (a) Exemples d'hyperplans séparant les données des classes "-1" (oranges) et "+1" (bleues). (b) Hyperplan séparateur optimal au sens de Vapnik [128], i.e. maximisant la marge entre les données à séparer.*

optimal est formulée sous la forme du problème d'optimisation suivant :

$$\begin{cases} \text{minimiser} & \frac{1}{2} \parallel \mathbf{w} \parallel^2 \\ \text{sous les contraintes} & l_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad \forall i \in 1 \dots n \end{cases} \tag{7.7}$$

$$\begin{aligned} \text{avec} \quad & \mathbf{x}_i : \text{vecteur de caractéristiques du } i^{\text{e}} \text{ exemple,} \\ & l_i : \text{classe du } i^{\text{e}} \text{ exemple } \in \{-1, 1\}. \end{aligned} \tag{7.8}$$

L'objectif de maximisation de la marge permet d'assurer une bonne généralisation à de nouveaux points test.
On peut montrer [2] que la fonction de décision f du séparateur à vaste marge (SVM) s'écrit :

$$f(\mathbf{x}) = sign(\sum_{i \in \text{supports}} \alpha_i l_i \mathbf{x}_i^\top \mathbf{x} + b)$$

où les α_i sont des coefficients appelés multiplicateurs de Lagrange.

Cependant, dans la majorité des cas pratiques, les classes ne sont pas linéairement séparables. Ce problème est résolu par l'autorisation d'une erreur ξ_i (variable ressort ou *slack variable* en anglais) que l'on cherchera à minimiser (on dit dans ce cas que l'on a une "marge souple"). La contrainte de bonne classification $l_k(\mathbf{w}^\top.\mathbf{x}_k + b) \geq 1$ introduite dans le problème 7.7 devient $l_k(\mathbf{w}^\top.\mathbf{x}_k + b) \geq 1 - \xi_k$. On introduit également une constante de

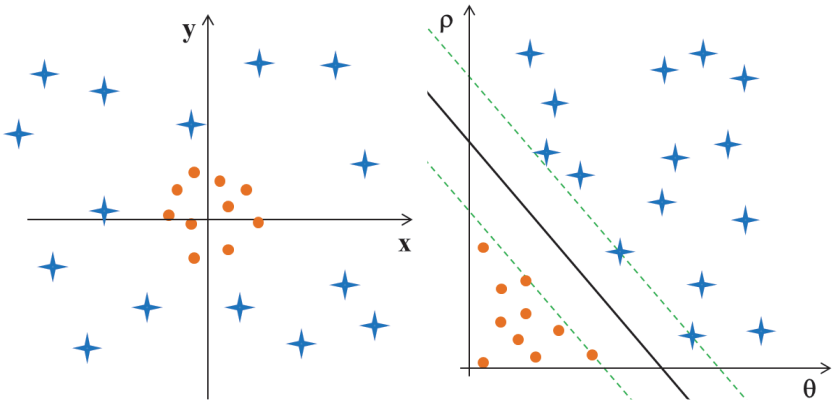
2. Le séparateur à vaste marge (SVM) sera au cœur de la partie III de cette thèse. Les éléments présentés dans cette section seront explicités chapitre 11.

coût, notée C, utilisée pour pénaliser plus ou moins l'erreur. Le problème d'optimisation devient :

$$\begin{cases} \text{minimiser} & \frac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_{i=1}^{n} \xi_i \\ \text{sous les contraintes} & l_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i \in 1 \dots n, \\ & \xi_i \geq 0, \quad \forall i \in 1 \dots n \end{cases} \tag{7.9}$$

$$\begin{aligned} \text{avec} \quad & \mathbf{x}_i : \text{vecteur de caractéristiques du } i^{\text{e}} \text{ exemple}, \\ & l_i : \text{classe du } i^{\text{e}} \text{ exemple} \in \{-1, 1\}, \\ & C : \text{coût de mauvaise classification}, \geq 1, \\ & \xi_i : \text{variable ressort}. \end{aligned} \tag{7.10}$$

Dans le cas de données non linéairement séparables dans leur espace d'origine, un changement d'espace, généralement de dimension plus grande, peut les rendre séparables. A une frontière linéaire dans l'espace transformé, correspond une frontière non linéaire dans l'espace d'origine (voir figure 7.9). L'hyperplan optimal dans l'espace transformé s'écrit :



(a) Coordonnées cartésiennes
Problème non linéairement séparable

(b) Coordonnées polaires
Problème linéairement séparable

Figure 7.9 – *Illustration du changement de base : projection par une fonction ϕ dans la transformée de l'espace des caractéristiques. Des données (a) non linéairement séparables dans le repère original peuvent devenir (b) séparables en utilisant des fonctions noyaux. Dans ce cas, le passage des coordonnées cartésiennes aux coordonnées polaires permet de rendre le problème original linéairement séparable.*

$$f(\mathbf{x}) = \sum_{i \in \text{supports}} \alpha_i l_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle + b$$

où ϕ est la fonction de projection dans le nouvel espace.

En pratique, le passage des points dans l'espace de dimension supérieure n'est jamais réalisé explicitement car, lors de la résolution du système d'équations, le changement de dimension est utilisé uniquement lors de la comparaison (produit scalaire) entre deux points $\mathbf{x}_i$ et $\mathbf{x}_j$. On utilise donc des fonctions de comparaison modifiées, appelées fonctions noyau

(ou *kernels*), qui se comportent comme un produit scalaire et réalisent cette transformation de manière implicite :

$$\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = K(\mathbf{x}_i; \mathbf{x}_j)$$

où K est une fonction noyau.
L'une des fonctions noyaux les plus communément utilisées est la *fonction à base radiale* (RBF, aussi appelée noyau gaussien) :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}) \tag{7.11}$$

où σ est un paramètre permettant de définir l'écart type (i.e. largeur) de la gaussienne associée. C'est celle que nous utiliserons dans cette étude.

Les valeurs optimales des paramètres C (constante de coût de violation de la contrainte de bonne classification) et σ (utilisée dans la formulation du noyau fonction à base radiale (RBF)) sont déterminées par recherche exhaustive sur grille (*grid search* en anglais, voir le tableau 7.5). La meilleure combinaison (C, σ) est choisie comme étant celle maximisant l'aire sous la courbe ROC (cf. section 4.5.2, page 62).

Paramètre	**Valeur**
Fonction noyau	*Radial Basis Function* (équation 7.11)
γ	déterminé par *grid search* : 21 valeurs entre 2^{-10} et 2^{10} (échelle logarithmique)
C	déterminé par *grid search* : 21 valeurs entre 2^{-5} et 2^{15} (échelle logarithmique)

Table 7.5 – *Choix des paramètres du SVM.*
C désigne le coût d'erreur de classification.
$\gamma = \frac{1}{2\sigma^2}$ est le coefficient de l'exponentielle de la RBF.

La probabilité d'appartenance à l'une ou l'autre des deux classes peut être estimée à partir de la distance à la marge. Nous utilisons pour cela l'algorithme de Platt [92] (décrit section 11.6).

Notre implémentation repose sur l'utilisation de la toolbox SVM-KM [15].

7.4.4 L'analyse discriminante linéaire (ADL)

L'analyse discriminante linéaire (ADL) fait l'hypothèse que les données issues des deux classes sont linéairement séparables. Géométriquement, elle cherche dans l'espace des caractéristiques $\mathbb{R}^d$ l'axe de projection optimal (de direction ω^*) des vecteurs de caractéristiques $(\mathbf{x}_i)_{i=1...n}$ de sorte que, dans ce nouveau repère, les points appartenant à deux classes différentes soient aussi distants les uns des autres que possible et que les données issues d'une même classe soient aussi proches que possible.

En pratique, on construit donc une fonction de discrimination canonique f qui s'ex-

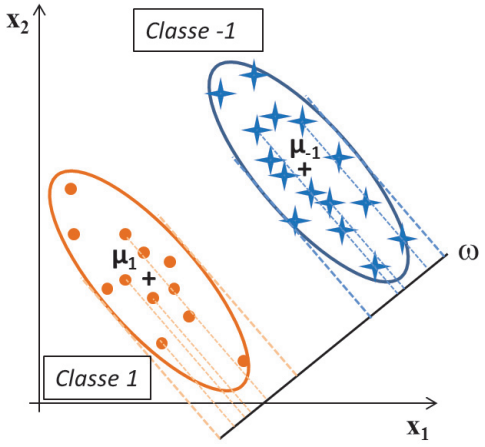


Figure *7.10* – *Illustration de la méthode de classification par ADL reposant sur le calcul d'un axe de projection optimal des classes.*

prime par une combinaison linéaire des variables de caractéristiques : $f(\mathbf{x}) = \omega^\top \mathbf{x}$.
$f(\mathbf{x})$ est le projeté de $\mathbf{x}$ sur l'axe de direction ω.

La projection optimale ω^* est estimée à partir des données d'apprentissage de façon à ce que la séparation entre les distributions des scores de discrimination $f(\mathbf{x}_i)$ des deux groupes, étiquetés ± 1, soit maximale.

L'optimisation de la séparation des classes repose sur la maximisation du critère de Fisher [36] :

$$J(\omega) = \frac{(\hat{\mu}_{\mathbf{1}} - \hat{\mu}_{\mathbf{-1}})^2}{\hat{\mathbf{s}}_{\mathbf{-1}}^2 + \hat{\mathbf{s}}_{\mathbf{1}}^2}$$

où $\hat{\mu}_{\mathbf{-1}}$, $\hat{\mu}_{\mathbf{1}}$, $\hat{\mathbf{s}}_{\mathbf{-1}}$, $\hat{\mathbf{s}}_{\mathbf{1}}$ sont respectivement les projetés des vecteurs moyens et les variances des projections de chacune des deux classes. On maximise le rapport de la variance inter-classes sur la variance intra-classe des scores calculés.
En écrivant que :

$$\hat{\mathbf{s}}_{\mathbf{k}}^2 = \sum_{\mathbf{x}|l=k} (\omega^\top \mathbf{x} - \omega^\top \mu_k)^2 \text{ et } (\hat{\mu}_{\mathbf{1}} - \hat{\mu}_{\mathbf{-1}})^2 = (\omega^\top \mu_{\mathbf{1}} - \omega^\top \mu_{\mathbf{-1}})^2,$$

on ré-exprime le problème d'optimisation à résoudre :

$$\omega^* = \arg\max_{\omega} J(\omega) = \arg\max_{\omega} \left(\frac{\omega^\top \mathbf{S_b} \omega}{\omega^\top \mathbf{S_w} \omega}\right)$$

où :

- la dispersion inter-classe $\mathbf{S_b} = (\mu_{\mathbf{-1}} - \mu_{\mathbf{1}})(\mu_{\mathbf{-1}} - \mu_{\mathbf{1}})^\top$,
- la dispersion intra-classe $\mathbf{S_w} = \sum_{k=-1,1} (\mathbf{x} - \mu_{\mathbf{k}})(\mathbf{x} - \mu_{\mathbf{k}})^\top$.

On peut montrer (en dérivant par rapport à ω et en annulant la dérivée) que la solution s'exprime simplement :

$$\omega^* = \mathbf{S_w}^\top (\mu_{\mathbf{-1}} - \mu_{\mathbf{1}}).$$

7.4.5 Le classifieur naïf de Bayes (CNB)

La technique du classifieur naïf de Bayes (CNB) est basée sur une approche probabiliste reposant sur le théorème de Bayes. Soit $\mathbb{P}(L = l_i)$ la probabilité *a priori* de la classe l_i, $\mathbb{P}(X = \mathbf{x})$, la probabilité d'observer une donnée $\mathbf{x}$, et $\mathbb{P}(X = \mathbf{x}|L = l_i)$ la probabilité d'observer le vecteur $\mathbf{x}$ sachant que la classe est l_i. La règle de Bayes permet alors de calculer la probabilité *a posteriori* de la classe l_i quand $\mathbf{x} = [x^1, \dots x^j, \dots x^d]$ est observé :

$$\mathbb{P}(L = l_i|X = \mathbf{x}) = \frac{\mathbb{P}(X = \mathbf{x}|L = l_i)\mathbb{P}(L = l_i)}{\sum_j \mathbb{P}(X = \mathbf{x}|L = l_j)\mathbb{P}(L = l_j)}.$$

Puisque le dénominateur de la formule de Bayes ne dépend pas de l_i, nous ne nous intéressons qu'au numérateur. Sous l'hypothèse naïve que chacune des variables X^j est conditionnellement indépendante de chacune des autres variables X^k pour $k \neq j$ (où X^j, X^k correspondent respectivement aux caractéristiques j et k), la classe de l'observation $\mathbf{x} = [x^1, \dots x^j, \dots x^d]$ est donnée par :

$$\underset{l}{\operatorname{argmax}}\ \mathbb{P}(L = l) \prod_{j=1}^{d} \mathbb{P}(X^j = x^j|L = l)$$

L'*a priori* de classe $\mathbb{P}(L = l)$, et les distributions de probabilité des variables de caractéristiques $\mathbb{P}(X^j|L)$ sont estimés à partir de la fréquence relative des données de la base d'apprentissage.

7.4.6 k-plus proches voisins (k-PPV)

L'algorithme des k-plus proches voisins (k-PPV) [26] est une méthode permettant de classer les objets en fonction des exemples d'entraînement étiquetés les plus proches dans l'espace des caractéristiques. Elle se base sur une comparaison directe entre le vecteur de caractéristiques de l'entité $\mathbf{x}$ à classer et les vecteurs de caractéristiques des entités de référence $(\mathbf{x}_i)_{i=1\dots n}$. La comparaison consiste en un calcul de distances entre ces entités. L'entité à classer est assignée à la classe majoritaire parmi les classes des k exemples les plus proches au sens de la distance utilisée.

Dans notre étude, nous utilisons une distance euclidienne. Soit $\mathbf{x}_i$ un point de la base d'apprentissage ($\in \mathbb{R}^d$) : $Dist(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{d} (x^j - x_i^j)^2}$.

On utilise une version probabiliste de l'algorithme des k-PPV [40], dans laquelle la probabilité d'appartenance à une classe est approchée par la proportion des k-plus proches voisins qui appartiennent à cette classe :

$$\mathbb{P}(L = l_i|X = \mathbf{x}) \simeq k_i/k,$$

où k_i représente le nombre d'exemples appartenant à la classe l_i parmi les k-plus proches voisins.

Dans notre étude, la valeur de k optimale est choisie parmi $\{1, 2 \dots 20\}$ comme celle maximisant les performances de classification (mesurées en terme d'AUC).

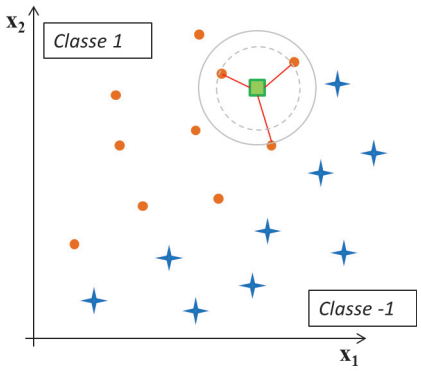


Figure 7.11 – *Illustration de la méthode de classification par k-PPV. Les points d'apprentissage de la classe "1" sont représentés en orange, ceux de la classe "-1" sont en bleus. Sur l'exemple, le nouvel élément (en vert) est classé "1" en fonction de la classe de ses k=3 plus proches voisins.*

7.5 Evaluation des performances

7.5.1 Validation croisée

Etant donnée la taille limitée de notre base de données, les performances des classifieurs sont estimées en utilisant une approche par validation croisée (lire section 4.6) de type Leave-One-Patient-Out (LOPO), évitant ainsi de réaliser l'apprentissage et le test sur les mêmes données. Le score de malignité de chacune des données du N^{e} patient (*left-out*) est calculé à partir de l'apprentissage sur les données extraites des N-1 patients restants. Cette procédure est répétée jusqu'à ce que les N=30 patients soient testés. Cette procédure de validation nous a semblé adaptée à notre problématique puisqu'elle permet de tenir compte du fait que nos données ont une structure en *clusters* (plusieurs ROIs par patient). Il a été montré que cette approche est non biaisée même si elle a tendance à sous-estimer les performances [29, 60, 108].

7.5.2 Problèmes de discrimination considérés

Rappelons que notre base de données est partitionnée en trois classes de tissus, définies lors de la séance de corrélation anato-radiologique (lire section 6.5) :

- les **tissus malins {M}**,
  qui correspondent à des VP ou à des FN de l'analyse radiologique *en aveugle*.
- les **tissus sains {N}** présentant une apparence normale sur les images IRM,
  qui correspondent à des VN de l'analyse radiologique *en aveugle*.
- les **tissus bénins mais suspects {NS}** présentant un signal anormal à l'IRM,
  qui correspondent à des FP ou à des VN de l'analyse radiologique *en aveugle*.

Suivant la démarche introduite par Vos *et coll.* [133], deux problèmes sont considérés pour évaluer les performances de discrimination des classifieurs. La première approche teste le pouvoir de séparation des classes H0={N, NS} versus H1={M} ; la seconde teste la capacité de discrimination des données issues de la classe H0={NS} de celles de type H1={M}. Le second problème est certainement le plus pertinent dans une perspective d'utilisation clinique, et le plus difficile puisqu'il s'agit de discriminer les cas potentiellement sources de FP des cas réellement malins.

Ces deux problèmes de discrimination seront référencés respectivement PB1 et PB2 dans la suite du manuscrit.

Remarquons qu'étant donnée la taille restreinte du jeu de données, et pour garantir une robustesse du schéma de classification, le processus d'apprentissage est systématiquement réalisé sur la base complète composée des cas {M}, {N} et {NS}.

7.5.3 Analyse ROC

Les courbes ROC (pour *receiver operating characteristic*) sont tracées pour chacune des 16 combinaisons "sélection de caractéristiques/ classifieur" testées et pour les deux

problèmes de discrimination considérés (PB1 : {N, NS} versus {M} et PB2 : {NS} versus {M}).

Les performances sont évaluées par le biais de l'aire sous la courbe ROC (notée AUC, pour *Area Under the ROC Curve*) permettant une comparaison non biaisée des différentes approches testées (lire section 4.5.2).
Les performances des différents schémas CAD sont analysées et comparées en utilisant le logiciel ROCKIT® (version 1.1-beta, CE Metz, Université de Chicago) qui fournit une estimation de la valeur AUC ainsi que des intervalles de confiance à 95% et permet une comparaison statistique des différentes approches. ROCKIT est un programme pour l'analyse paramétrique des courbes ROC qui repose sur l'hypothèse d'un modèle binormal (cf. section 4.5.3). [3]

7.6 Conclusion

En nous appuyant sur une base de données d'images IRM multi-paramétriques (IRM-mp) acquises sur 30 patients et desquelles nous avons extrait 140 caractéristiques structurelles et fonctionnelles, nous proposons de tester et comparer les performances de quatre classifieurs combinés avec quatre méthodes de sélection de caractéristiques afin d'isoler la combinaison optimale. La section suivante présente les résultats obtenus par chacun de ces schémas de classification suivant les deux problèmes de discrimination considérés :
"H0={N, NS} versus H1={M} (PB1)" ou "H0={NS} versus H1={M} (PB2)".

3. Les scores de chaque population H0, H1, sont supposés suivre une loi normale $P(\lambda_0, \mu_0)$, $P(\lambda_1, \mu_1)$. Les paramètres du modèle (λ_i, μ_i) sont ajustés par maximisation de la vraisemblance. ROCKIT teste la significativité, au sens statistique, de la différence entre deux courbes ROC appariées (i.e. classifieurs testés sur les même données).

Chapitre 8

Comparaison des performances des différents schémas CAD

8.1 Introduction

Nous présentons dans cette section les résultats obtenus en combinant nos quatre approches de sélection de caractéristiques à nos quatre schémas de classification sur notre base de données de 30 patients (215 ROIs).
La première étape consiste à identifier la combinaison de caractéristiques les plus discriminantes dans la tâche de classification suivant l'approche de sélection employée (test *t*, IM, mRMRd et mRMRq). Il sera intéressant d'observer si une signature dans la liste des caractéristiques sélectionnées peut être identifiée, si les caractéristiques issues d'une séquence (T2-w, ADC, DCE) apparaissent plus discriminantes que les autres, ou si un type de caractéristiques descriptives (signal brut, paramètres statistiques, de gradient, de texture ou pharmacocinétique) est plus discriminant que les autres.
La seconde étape consiste à identifier le schéma de classification (SVM, ADL, CNB et k-PPV) permettant d'obtenir les meilleures performances de discrimination, mesurées par le biais de l'aire sous la courbe ROC (AUC). Il sera intéressant de comparer les performances obtenues par les différentes approches de classification, notamment entre les méthodes "naïves" et les méthodes plus complexes.

Le schéma de classification optimal identifié sera ensuite testé lors d'une évaluation clinique afin de quantifier son apport dans les mesures de performances diagnostiques de

12 radiologues volontaires (cf. chapitre 9).

8.2 Caractéristiques sélectionnées

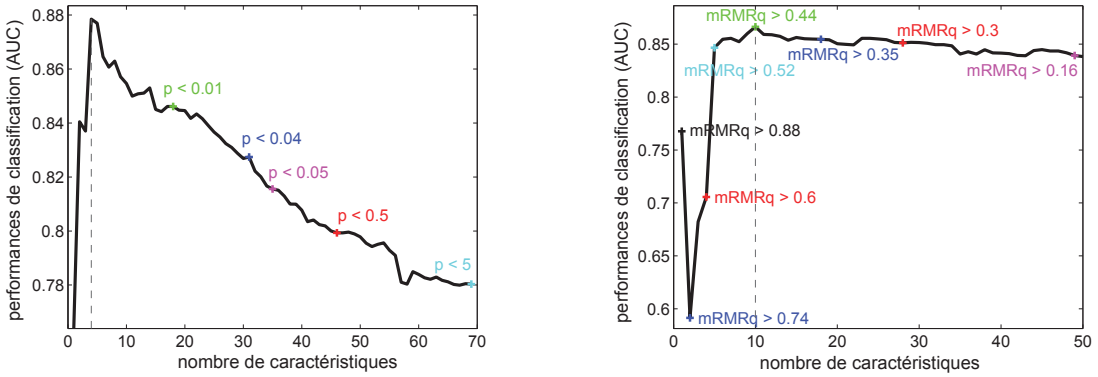
Les quatre critères d'évaluation des caractéristiques décrits dans la section 7.3 sont calculés pour chacun des paramètres descriptifs extraits. On obtient donc quatre listes de caractéristiques ordonnées selon les valeurs croissantes des quatre critères utilisés. Le tableau 8.1 présente les rangs de différentes caractéristiques selon les quatre méthodes de sélection. Seules les caractéristiques les plus discriminantes sont présentées dans ce tableau, correspondant à celles sélectionnées parmi les 25 premières par au moins trois des quatre méthodes en lice.
Certaines caractéristiques sont sélectionnées selon les quatre critères d'ordonnancement ; celles-ci proviennent des trois types de séquences IRM (quatre paramètres issus de l'ADC, deux de la T2-w et deux de la DCE), confirmant a posteriori la légitimité d'une approche multi-séquence mais également de tous les types de caractéristiques descriptives, validant l'intérêt des paramètres statistiques, structurels et fonctionnels dans la discrimination des tissus. La liste des 25 premières caractéristiques sélectionnées par chacune des quatre méthodes de sélection est donnée dans le tableau A.1 de l'annexe A.

	test t	IM	mRMRd	mRMRq	nb. sélections
T2-w gradient δ_y	2	1	6	5	4
ADC med	4	5	8	6	4
DCE Tmax	13	6	4	7	4
ADC Sobel$_{xy}$	17	13	11	10	4
DCE AUGC	9	12	13	19	4
T2-w Sobel$_{xy}$	7	20	14	16	4
ADC 25%	10	10	25	15	4
ADC Sobel$_y$	1	-	1	1	3
T2-w var	-	4	3	3	3
T2-w Sobel$_y$	3	2	-	13	3
DCE moy	5	7	-	12	3
DCE WI	-	3	19	4	3
DCE ve	-	15	7	11	3
T2-w entropy	-	21	12	14	3

Table *8.1* – *Liste des caractéristiques descriptives les plus discriminantes sélectionnées dans les 25 premières par au moins trois des quatre méthodes de sélection. Les abréviations de nom utilisées pour référencer chacune des caractéristiques sont données section 7.2.*

Les résultats des méthodes de sélection sont présentés dans le tableau 8.2 et illustrés par la figure 8.1. La figure 8.1 souligne l'impact qu'a la pré-sélection de caractéristiques sur les performances de l'algorithme de classification. La courbe 8.1a montre les performances de classification obtenues par le classifieur naïf de Bayes (CNB) en fonction du nombre de caractéristiques, celles-ci étant ordonnées selon les p-valeurs dérivées d'une analyse statistique par test t (décrite section 7.3.3). Une valeur d'AUC maximale de 0.88 est obtenue

en utilisant uniquement les quatre premières caractéristiques ainsi ordonnées. L'utilisation d'un nombre supérieur de caractéristiques descriptives conduit à du sur-apprentissage et décroît les performances de discrimination. Ce graphique illustre la variation des performances de classification en fonction du nombre de caractéristiques sélectionnées et montre combien le choix du sous-groupe de caractéristiques à conserver est crucial. La courbe 8.1b montre des résultats similaires obtenus pour le classifieur SVM combiné avec le critère de sélection de type mRMRq (décrit section 7.3.5). La valeur d'AUC optimale de 0.87 est obtenue en utilisant les 10 premières caractéristiques ordonnées selon le critère mRMRq. L'ajout de plus de caractéristiques n'améliore pas (ni ne décroît) les performances de classification. Finalement, le même processus est réalisé pour toutes les autres combinaisons



(a) test t + CNB **(b)** mRMRq + SVM

Figure *8.1* – *Performances de classification (AUC) en fonction du nombre de caractéristiques utilisées. Tâche de discrimination PB1 : H0={N, NS} versus H1={M}.(a) Les caractéristiques sont ordonnées selon les p-valeurs croissantes. Seules les caractéristiques pour lesquelles l'hypothèse nulle (les moyennes de distributions de points pour les deux classes sont identiques) peut être rejetée au seuil de confiance de 95% et plus (p-valeurs ≤ 5%) sont ordonnées. Dans cet exemple évaluant le classifieur naïf de Bayes, la performance optimale (AUC = 0.88) est obtenue en utilisant uniquement les quatre premières caractéristiques. (b) Les caractéristiques sont ordonnées selon les valeurs croissantes du critère mRMRq. La performance maximale (AUC = 0.87) est obtenue lorsque les 10 premières caractéristiques sont utilisées (pour lesquelles mRMRq ≥ 0.44).*

de stratégie de sélection et de classification. Il permet de sélectionner les seuils sur les différents critères correspondant aux meilleures performances de discrimination pour chaque classifieur. Les résultats sont donnés dans le tableau 8.2a pour le premier problème de discrimination (PB1) des cas sains {N, NS} et pathologiques {M} et 8.2b pour le problème plus difficile (PB2) de discrimination des zones suspectes {NS} et cancéreuses {M}.

Par exemple, les 15 caractéristiques sélectionnées en utilisant comme critère la p-valeur obtenue par test t et permettant d'obtenir des performances optimales avec le classifieur SVM sont, pour le T2-w : la moyenne et la médiane du gradient de Sobel-xy, le coefficient de dissimilarité, le gradient numérique-y, le gradient de Sobel-y ; sur l'ADC : le gradient de

	test t	IM	mRMRd	mRMRq
SVM	15	14	8	10
ADL	15	42	8	7
k-PPV	12	6	6	6
CNB	4	6	8	13

(a) PB1 : H0={N, NS} versus H1={M}

	test t	IM	mRMRd	mRMRq
SVM	34	45	9	6
ADL	42	42	8	7
k-PPV	12	23	38	1
CNB	4	7	8	6

(b) PB2 : H0={NS} versus H1={M}

Table *8.2* – *Nombre de caractéristiques sélectionnées pour la discrimination des tissus malins versus bénins en fonction du critère de sélection et de l'algorithme de classification utilisés mais également du problème de classification considéré : (a) PB1 : H0={N, NS} versus H1={M} et (b) PB2 : H0={NS} versus H1={M}. On remarque une réduction importante du nombre de caractéristiques puisqu'on passe systématiquement de plus de 140 à moins de 50 paramètres.*

Sobel-y, la moyenne, la variance, la médiane et le 1er quartile des valeurs ; et sur la DCE : le 1er quartile de l'aire sous la courbe de réhaussement en produit de contraste (AUGC), la valeur moyenne, médiane et le 3e quartile des valeurs de réhaussement relatif (à 1 min après injection de Gd-DOTA), le Tmax.

8.3 Performances de classification

La figure 8.2 montre les courbes ROC expérimentales obtenues sans pré-sélection et avec différentes combinaisons de méthodes de sélection de caractéristiques et d'algorithmes de classification pour la tâche de discrimination des classes {NS, N} versus {M}. La légende indique les valeurs d'AUC correspondantes estimées par le logiciel d'analyse ROCKIT® ainsi que les intervalles de confiance à 95%. La figure 8.3 montre les courbes ROC correspondant à la tâche de discrimination des classes {NS} versus {M} obtenues avec les différents classifieurs sans pré-sélection des caractéristiques (courbes A.1a) et avec sélection de type test t (courbes A.1b). Le tableau 8.3 liste l'ensemble des valeurs d'AUC obtenues pour la tâche de discrimination des classes {NS, N} versus {M}. Le tableau 8.4 fait de même pour la tâche de discrimination des classes {NS} versus {M}. Ces résultats démontrent que la sélection de caractéristiques réalisée préalablement aux étapes d'apprentissage et classification améliore considérablement les performances par rapport à un apprentissage sans pré-sélection (voir la première ligne des tableaux 8.3 et 8.4), et ceci quelle que soit la méthode utilisée.

Les valeurs d'AUC atteintes pour la tâche de discrimination (PB1) des classes {NS, N} versus {M} avec sélection de caractéristiques sont très proches, avec des valeurs s'étalant de 0.83 à 0.89. La meilleure performance (AUC=0.89) est réalisée avec le classifieur SVM combiné avec une sélection basée sur le test t. L'analyse statistique indique qu'il y a une différence significative ($p < 0.05$) entre les paires de courbes ROC suivantes : les

performances atteintes par le SVM sont significativement meilleures que celles obtenues avec les classifieurs ADL, k-PPV et CNB lorsqu'ils sont appliqués sur le jeu de données brut (sans pré-sélection) ; toutes les méthodes de sélection testées (test *t*, information mutuelle (IM), mRMRd et mRMRq) donnent des résultats significativement meilleurs que ceux obtenus sur le jeu de données complet (sans pré-sélection) lorsqu'elles sont couplées avec les classifieurs de type ADL, k-PPV et CNB. L'application d'une méthode de sélection des caractéristiques au préalable à la classification n'améliore pas de manière significative les performances obtenues avec les SVM. La combinaison de la sélection par test *t* et de la classification par SVM réalise des performances significativement meilleures que les combinaisons (que l'on notera de la façon suivante dans la liste ci-après : méthode de sélection/classifieur) : sans sélection/ADL, IM/k-PPV, MRMRq/k-PPV, sans sélection/k-PPV, IM/CNB, sans sélection/CNB. La combinaison IM/SVM est significativement meilleure que les combinaisons : IM/k-PPV et IM/CNB.

Les performances AUC réalisées pour la tâche de discrimination (PB2) des classes {NS} et {M} sont résumées dans le tableau 8.4 et varient de 0.69 à 0.82. Ces résultats soulignent, premièrement, la complexité de cette tâche de discrimination comparée à la tâche PB1 ({NS, N} versus {M}) ; deuxièmement, l'influence plus notable tant du choix du classifieur que de la méthode de sélection de caractéristiques.
La performance maximale (AUC = 0.82) a également été obtenue avec un classifieur SVM associé à une sélection basée sur un test *t*. La performance du SVM est significativement meilleure que celles des ADL, k-PPV et CNB lorsque utilisés sur l'ensemble des caractéristiques (sans pré-sélection). Toutes les méthodes de sélection testées (test *t*, IM, mRMRd et mRMRq) donnent des résultats significativement meilleurs que ceux obtenus en utilisant le jeu de données brut (sans pré-sélection) lorsqu'elles sont couplées avec un classifieur ADL. La méthode de sélection basée sur un test *t* donne des résultats significativement meilleurs que ceux obtenus en utilisant l'ensemble des caractéristiques (sans pré-sélection) lorsqu'elle est couplée avec les classifieurs SVM, ADL, k-PPV et CNB. La combinaison test *t*/SVM est significativement meilleure que les combinaisons suivantes : mRMRd/SVM, t-test/LDA, IM/ADL, mRMRd/ADL, mRMRq/ADL, sans-sélection/ADL, IM/k-PPV, mRMRq/k-PPV, sans-sélection/k-PPV, IM/CNB, mRMRq/CNB, sans-sélection/CNB. La combinaison IM/SVM est significativement meilleure que les combinaisons IM/k-PPV etIM/CNB.

Pour résumer, le classifieur SVM combiné avec une sélection des caractéristiques de type test *t* réalise les meilleures performances pour chacune des deux tâches de discrimination considérées (PB1 et PB2). La figure 8.4 montre les deux courbes expérimentales obtenues avec ce schéma de classification (test *t*/SVM) obtenues pour les deux tâches de discriminations ainsi que les courbes ROC dérivées de l'analyse et de la modélisation réalisée par ROCKIT (en pointillés).

Les performances du classifieur SVM sont significativement meilleures que celles obte-

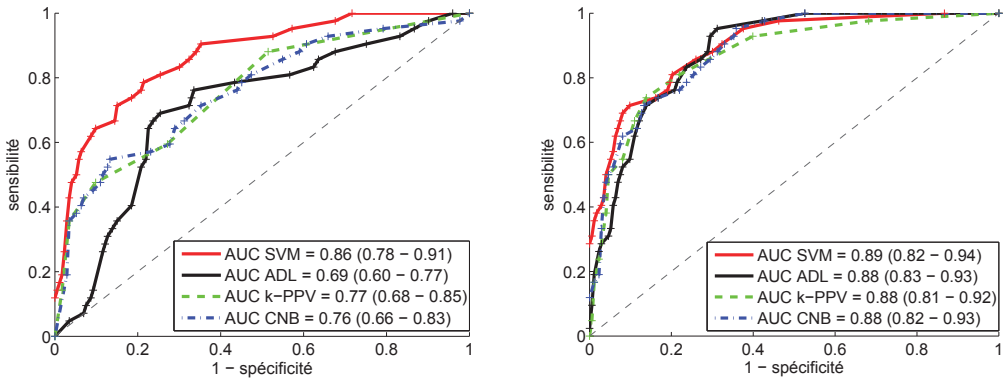
Table *8.3* – *Performance (AUC) des classifieurs en fonction de la méthode de sélection de caractéristiques utilisée. Tâche de discrimination PB1 : H0={N, NS} versus H1={M}.*

mesure AUC	SVM	ADL	k-PPV	CNB
sans sélection	0.86 [0.78 - 0.91]	0.69 [0.60 - 0.77]	0.77 [0.68 - 0.85]	0.76 [0.66 - 0.83]
test t	**0.89** [0.82 - 0.94]	0.88 [0.83 - 0.93]	0.88 [0.81 - 0.92]	0.88 [0.82 - 0.93]
IM	0.87 [0.80 - 0.92]	0.87 [0.82 - 0.92]	0.83 [0.76 - 0.89]	0.84 [0.78 - 0.90]
mRMRd	0.87 [0.81 - 0.92]	0.87 [0.81 - 0.91]	0.87 [0.80 - 0.92]	0.87 [0.81 - 0.92]
mRMRq	0.87 [0.81 - 0.92]	0.87 [0.81 - 0.91]	0.84 [0.77 - 0.90]	0.86 [0.80 - 0.90]

Table *8.4* – *Performance (AUC) des classifieurs en fonction de la méthode de sélection de caractéristiques utilisée. Tâche de discrimination PB2 : H0={NS} versus H1={M}.*

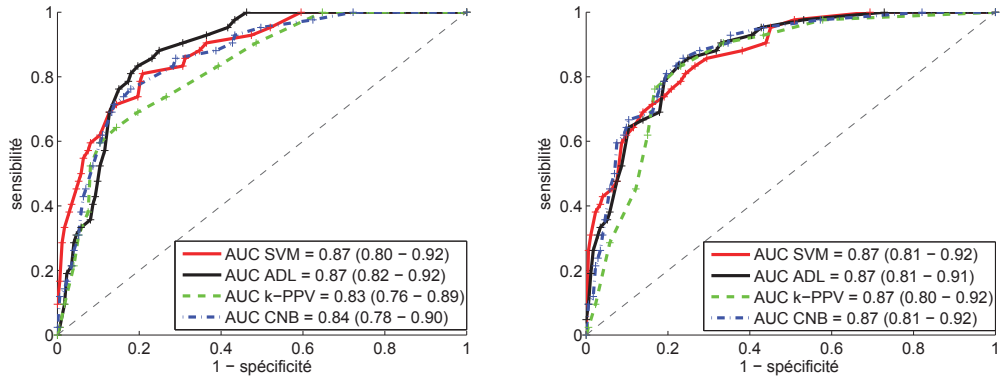
mesure AUC	SVM	ADL	k-PPV	CNB
sans sélection	0.72 [0.61 - 0.82]	0.56 [0.44 - 0.67]	0.66 [0.54 - 0.77]	0.63 [0.51 - 0.74]
test t	**0.82** [0.73 - 0.90]	0.75 [0.64 - 0.84]	0.78 [0.68 - 0.87]	0.77 [0.66 - 0.85]
IM	0.78 [0.67 - 0.86]	0.77 [0.66 - 0.85]	0.70 [0.59 - 0.80]	0.69 [0.57 - 0.79]
mRMRd	0.76 [0.64 - 0.84]	0.72 [0.61 - 0.82]	0.76 [0.65 - 0.86]	0.75 [0.65 - 0.84]
mRMRq	0.76 [0.65 - 0.84]	0.72 [0.61 - 0.82]	0.71 [0.60 - 0.81]	0.73 [0.62 - 0.82]

nues avec les 3 autres classifieurs sur le jeu de paramètres global (sans sélection). Même si la valeur d'AUC obtenue avec le schéma de classification test t/SVM est plus haute que celles obtenues avec les autres combinaisons de méthode de sélection / classifieur, une différence statistiquement significative n'a pas pu être démontrée. Un jeu de données plus important aurait peut-être pu amener à de telles conclusions.



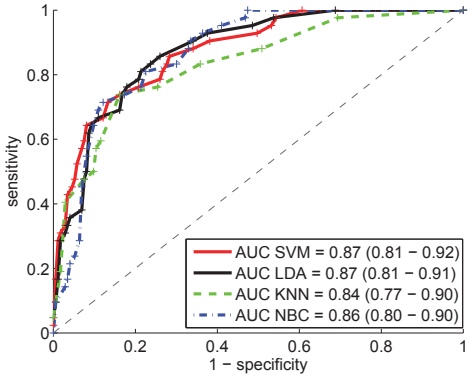
(a) sans selection

(b) test t



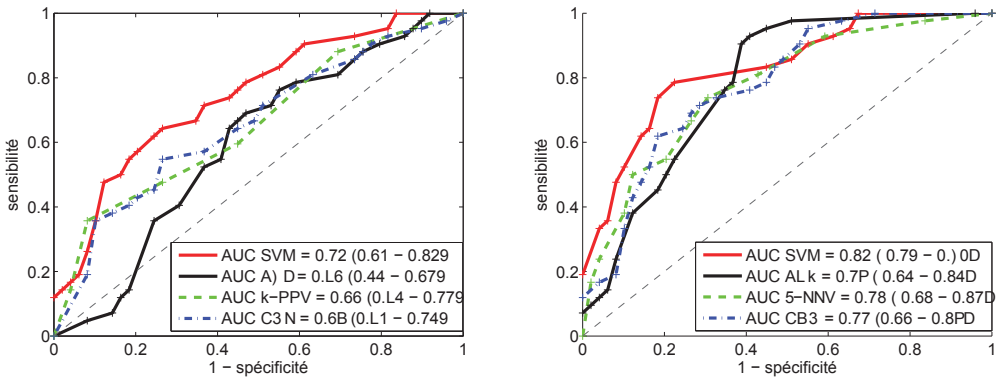
(c) IM

(d) mRMRd



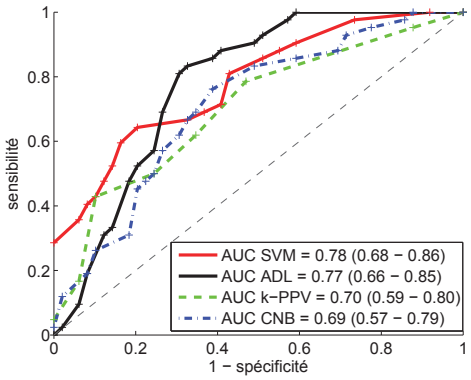
(e) mRMRq

Figure *8.2* – *Performances des SVM, ADL, k-PPV et CNB en fonction de la méthode de sélection des caractéristiques appliquée : (a) en utilisant toutes les caractéristiques extraites (sans sélection), après une sélection par (b) test t, (c) information mutuelle, (d) mRMRd et (e) mRMRq. Ces courbes ROC correspondent à la tâche de discrimination PB1 : H0={N,NS} versus H1={M}.*

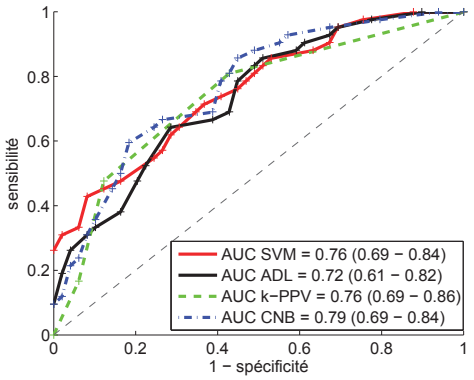


(a) sans selection

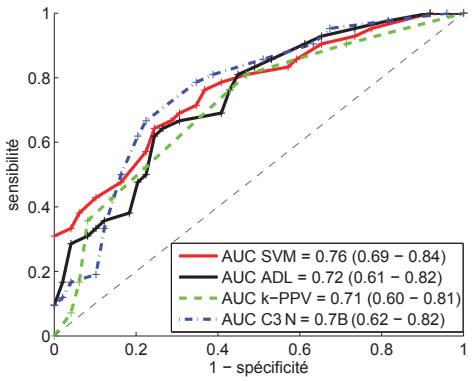
(b) test t



(c) IM



(d) mRMRd



(e) mRMRq

Figure *8.3* – *Performances des SVM, ADL, k-PPV et CNB en fonction du jeu de caractéristiques utilisé : (a) en utilisant toutes les caractéristiques extraites (sans sélection), avec une sélection basée sur (b) un test t, (c) l'information mutuelle, (d) le critère mRMRd et (e) le critère mRMRq. Ces courbes ROC correspondent à la tâche de discrimination PB2 : H0={NS} versus H1={M}.*

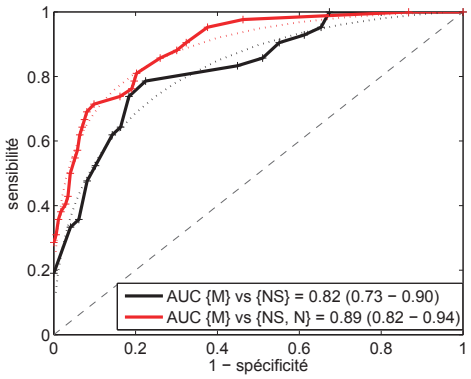


Figure *8.4* – *Performances de classification des SVM après une pré-sélection basée sur un test t pour les deux tâches de discrimination considérées : H0={NS} versus H1={M} et H0={N,NS} versus H1={M}.*

8.4 Discussion

Cette étude démontre la faisabilité du développement d'un système CADx capable de discriminer les tissus malins {M} des tissus bénins {N, NS} avec une AUC de 0.89 (intervalle de confiance à 95% = 0.81-0.94). Cette valeur d'AUC de 0.89, obtenue en utilisant une approche par SVM, peut être comparée favorablement aux résultats de la littérature. Dans leurs études combinant la DCE et la séquence T2-w, Vos *et coll.* [133] et Viswanath *et coll.* [131], obtiennent respectivement des performances diagnostiques de 0.89 et 0.82. Chan *et coll.* [18] estiment la probabilité de présence du cancer à partir de la séquence T2-w, de la cartographie T2-map et des images de diffusion ; ils parviennent à une AUC de 0.84. Dans leurs études combinant les images issues des séquences T2-w, DCE et DWI, Artan *et coll.* [5] et Ozer *et coll.* [83] réalisent des performances de 0.79 et 0.83, respectivement.

Néanmoins, comme nous l'avons déjà souligné, il reste difficile d'établir une comparaison juste entre les études puisque la population testée varie d'une étude à l'autre (notamment en nombre de patients), les données sont différentes (différents types et nombres de séquences exploitées) les annotations ne suivent pas le même protocole (classification par pixels, sextants, ROIs, etc), les tâches de discrimination sont différentes ({M} versus {N} ou {N, NS}) et l'évaluation ne repose pas sur les mêmes critères (validation croisée LOO, LOPO, k-fold, resubstitution etc ; critères de recouvrement ou de dénombrement différents, etc). De ce fait, la comparaison aux autres études doit être considérée avec précaution.

Le second problème de discrimination considéré vise à séparer les tissus malins {M} des régions suspectes {NS} qui présentent un signal anormal à l'IRM. La plupart de ces régions suspectes sont des faux positifs, identifiées comme possiblement malignes lors de la phase d'annotation *a priori* (lire section 6.3). Il s'agit donc d'un problème plus difficile et d'une tâche de discrimination cliniquement plus intéressante. La valeur d'AUC obtenue est de 0.82 (0.73 - 0.9) (voir le tableau 8.3). Ce résultat suggère qu'un système CADx pourrait être un avis additionnel utile pour les radiologues, en particulier non-experts (nous testons cette hypothèse dans le chapitre 9 suivant).

Cette étude nous a également permis de tester, sur un même jeu de données, différentes stratégies de sélection de caractéristiques et de classification des données, rendant ainsi possible une comparaison objective de ces méthodes. De plus, comparé à la majorité des jeux de données exploitées dans la littérature, notre base de données est riche par son nombre de cas (30 patients, 215 ROIs) et par sa vérité terrain construite à partir d'une référence histologique fiable, analysée de manière exhaustive. Bien que les différences ne soient pas toujours statistiquement significatives, la comparaison des résultats des classifieurs, présentée dans les tableaux 8.3 et 8.4, montre que les meilleures performances de discrimination sont obtenues en utilisant le classifieur SVM combiné avec une sélection des

caractéristiques basée sur un test *t*.
Remarquons que l'utilisation de méthodes d'analyse statistique prenant en compte le fait que les données ont une structure de *clusters* (plusieurs ROIs par patient), aurait peut-être permis de montrer des différences plus significatives (d'un point de vue statistique) entre les différentes stratégies de classification. Au moment où cette étude a été réalisée, la tendance générale des différences d'AUC nous semblait suffisante pour identifier le schéma optimal. Néanmoins, une telle méthode (basée sur l'approche statistique proposée par Rutter *et coll.* [106]) a été par la suite mise en œuvre pour l'évaluation clinique de notre prototype CADx. Pour cette étude, l'objectif était effectivement d'étudier si une différence significative entre les performances des lecteurs avec et sans le système CADx pouvait être mise en évidence. Il était donc nécessaire d'utiliser l'approche la plus adaptée d'un point de vue statistique.

Les résultats intéressants de la tâche de sélection des caractéristiques peuvent également être soulignés. L'étude montre qu'une sélection appropriée d'un sous-groupe de caractéristiques est une étape tout aussi cruciale que le choix du classifieur pour améliorer les performances de discrimination et éviter le sur-apprentissage. Comme nous pouvons le voir sur les tableaux 8.3 et 8.4, l'utilisation de l'ensemble des caractéristiques extraites sans distinction conduit à de plus faibles performances, pour un classifieur donné.

On peut également remarquer, dans le tableau 8.1, que chacune des trois séquences IRM (T2-w, DCE, ADC) ainsi que chacun des différents types de caractéristiques (statistique, structurel et fonctionnel) s'avère pertinent dans la tâche de discrimination, confirmant ainsi *a posteriori* la validité d'une approche multi-paramétrique.

Il est important de noter que les performances obtenues dans cette étude dépendent très certainement du type de scanner IRM et des paramètres d'acquisition ainsi que de la capacité des experts à concevoir une base d'entraînement fiable. Néanmoins, le cadre méthodologique et la suite d'algorithmes que nous avons développés permettent d'estimer facilement les performances des différents schémas de classification sur un nouveau jeu de données annotées.

8.5 Conclusion et perspectives

A ce travail s'ouvrent différentes perspectives pour améliorer les performances. D'une part, une augmentation régulière du nombre des cas dans la base de données améliorera la pertinence de l'étude. D'autre part, d'autres méthodes de sélection doivent être testées, en particulier les méthodes de type "enveloppantes" (*wrapper*). Le travail sur l'extraction de caractéristiques peut également être approfondi avec l'inclusion d'autres paramètres quantitatifs (ondelettes, paramètres de Gabor etc) et qualitatifs (critères de localisation de la cible, utilisation d'un atlas). Nous pouvons également nous pencher sur la problématique de la combinaison des sorties des classifieurs afin de tester si elle permet d'améliorer les

performances globales comme cela a été suggéré par Niemeijer *et coll.* [77]. Une autre approche à tester pourrait être de réaliser l'apprentissage de d classifieurs sur chacune des d caractéristiques (ou sur des sous-ensembles de caractéristiques, groupées selon la séquence ou le type) suivi d'une combinaison des sorties ; dans cet objectif et pour ajouter de la cohérence dans l'étape d'apprentissage, l'utilisation d'approches de type *Multiple Kernel Learning* (MKL) [58] pourrait être intéressante.

Chapitre 9

Evaluation en conditions cliniques

9.1 Introduction

Nous avons souhaité réaliser une évaluation complète de notre prototype CADx en mesurant ses performances en conditions cliniques. Le schéma CADx optimal, basé sur un SVM couplé à un test t, a ainsi été évalué en pratique clinique auprès de douze radiologues, de six mois à sept ans d'expérience, pour lesquels nous avons mesuré les performances diagnostiques sans et avec l'outil CADx utilisé en "second avis".

L'objectif était d'analyser comment l'outil CADx affectait la décision du radiologue et son niveau de confiance dans l'interprétation des images IRM-mp.
Le protocole d'évaluation est détaillé section 9.2. Les résultats obtenus sont détaillés sections 9.3.2 et 9.3.3.

Notons que l'analyse statistique présentée sections 9.3.2 et 9.3.3 a été réalisée en collaboration avec L. Roche et M. Rabilloud du service de Biostatistiques des Hospices Civils de Lyon (Lyon, France).

9.2 Protocole

9.2.1 Participants

Douze radiologues, d'expérience variable entre six mois (pour les six internes) et sept ans, se sont portés volontaires pour cette étude (voir tableau 9.1). On distinguera dans la suite les 6 radiologues juniors des 6 radiologues séniors (i.e. ayant une expérience significative dans l'analyse des images IRM de prostate).

Lecteur n°	1	2	3	4	5	6	7	8	9	10	11	12
Expérience (années)	0.5	3	7	1	1	0.5	0.5	0.5	0.5	0.5	1	2

Table *9.1* – *Expérience des participants à l'étude clinique*

9.2.2 Matériel

Les données utilisées sont celles correspondant aux 30 patients, précédemment utilisées pour la comparaison des différents schémas de classification. Les protocoles d'acquisition et d'annotations sont détaillés dans le chapitre 6. Pour chaque patient, on considère :

les images IRM-mp, comprenant la séquence T2-w, la cartographie ADC, et les images de la séquence DCE.

un jeu de contours (ROIs), correspondant :

- à des zones de tissus malins, précédemment notées {M},
- à des zones saines mais qui présentent un signal pouvant être considéré comme suspect sur tout ou partie des séquences , précédemment notées {NS}.

Il s'agit des contours réalisés lors de la corrélation anatomo-radiologique (lire section 6.5) pour la constitution de la vérité terrain. Le nombre de ROIs est variable suivant le patient considéré. A noter que les ROIs correspondant à des tissus sains qui ne présentent pas un signal suspect à l'IRM (précédemment notées {N}) ne sont pas considérées dans cette étude.

Ces données sont rendues anonymes préalablement à l'étude.
Pratiquement, un Macintosh®, sur lequel les données IRM-mp des patients sont chargées, est fourni à chaque lecteur. La visualisation des données se fait via le logiciel OsiriX® (Genève, Suisse). L'utilisateur est libre d'organiser sa fenêtre de visualisation, et en particulier de sélectionner les temps d'acquisition souhaités pour la séquence dynamique DCE. Les contours peuvent facilement être supprimés/ajoutés sur la fenêtre de visualisation.

9.2.3 Déroulement du protocole

L'expérience se déroule en trois lectures, organisées au sein du pavillon "Radiologie urinaire et vasculaire, diagnostique et interventionnelle" de l'hôpital Edouard Herriot (Lyon, France).

Les deux premières lectures sont réalisées sans utilisation du système CADx et permettent d'évaluer la variabilité intra-observateur ; elles sont séparées de 5 semaines. La troisième lecture, réalisée directement après la deuxième, permet, elle, d'évaluer l'apport du CADx.
Le protocole d'évaluation est présenté au préalable aux lecteurs lors d'une séance dédiée. Le principe du prototype CADx est expliqué, ses performances obtenues en validation-croisée présentées.

9.2.4 Première et deuxième lectures

Lors de la première lecture, chacun des douze lecteurs doit analyser les données des 30 patients. Pour chacun des patients, la série de ROIs préalablement tracées est soumise à l'évaluation du lecteur. Pour chacune d'entre elles, le lecteur doit fournir un score de suspicion de malignité, ainsi qu'il le ferait en routine clinique. Ce score, détaillé dans le tableau 9.2, s'échelonne entre 0 (bénignité certaine) et 4 (malignité certaine).

Score	Signification
0	bénignité certaine
1	probablement bénin
2	intermédiaire
3	probablement malin
4	malignité certaine

Table *9.2* – *Score de suspicion de malignité*

La même procédure est suivie lors de la deuxième lecture. Les scores affectés lors de la première lecture ne sont pas communiqués.

9.2.5 Analyse avec CADx

Une troisième (et dernière) lecture est réalisée directement après la deuxième lecture. Pour chaque ROI de chaque patient, le score de malignité calculé par le CADx est présenté au lecteur. Le lecteur a alors la possibilité d'amender le score précédemment affecté (en deuxième lecture) en fonction de celui proposé par le système.

Aucun *a priori* sur la manière de prendre en compte le score fourni par le système n'a été donné. Les performances (courbes ROC) du prototype CADx ont été communiquées et discutées, laissant le lecteur seul juge de la confiance qu'il pouvait accorder au système. Chacun des participants peut choisir librement d'utiliser ou non ce second "avis" et de modifier son jugement.

Les scores attribués lors de ces trois lectures sont enregistrés et analysés *a posteriori*.

9.2.6 Analyse des résultats

L'objectif de cette étude est multiple. Il s'agit d'une part d'évaluer la variabilité inter- et intra- expert de l'analyse diagnostique des images et d'en étudier l'évolution lorsque l'avis CADx est utilisé. D'autre part, nous souhaitons mesurer l'impact du CADx sur la confiance des lecteurs et leurs performances diagnostiques. Après avoir estimé les tendances générales de l'évolution des scores attribués au cours des lectures, on réalise une analyse statistique des résultats par le biais de modèles de régression. Cette partie de l'analyse a été réalisée par L. Roche et M. Rabilloud du service de Biostatistiques des Hospices Civils de Lyon (Lyon, France). Nous avons contribué activement à ce travail en motivant le choix des modèles, des variables et des paramètres à mesurer. Les détails méthodologiques des différentes analyses présentées dans la suite de ce chapitre sont donnés dans l'annexe B.

9.3 Résultats

9.3.1 Variabilité intra- et inter-expert

Les lectures 1 et 2 nous permettent d'étudier la variabilité intra-expert. Au cours de ces deux lectures, les mêmes cibles sont étudiées et annotées, par le lecteur, d'un score variant entre 0 et 4. Le tableau 9.3 détaille, pour chaque lecteur, les valeurs du coefficient de corrélation entre les scores attribués en lecture 1 et ceux attribués en lecture 2 (5 semaines plus tard). On observe une forte variabilité dans l'attribution des scores diagnostics (coefficient de corrélation variant entre 0.39 et 0.9), en particuliers pour les lecteurs juniors (1, 6, 7, 8, 9, 10). On remarque la très forte variabilité du lecteur 11 (coefficient de corrélation = 0.39), que nous n'avons pas pu expliquer.

Nous étudions également la variabilité inter-experts pour les lectures 1 et avec CADx (les résultats obtenus pour la lecture 2 étant sensiblement équivalents à ceux de la lecture 1). Le tableau 9.4 détaille les valeurs du coefficient de corrélation entre les scores attribués par les différents lecteurs en lecture 1. Le tableau 9.5 détaille, lui, les valeurs du coefficient de corrélation entre les scores attribués par les différents lecteurs lors de la lecture avec le CADx. On observe une augmentation notable du coefficient de corrélation moyen avec l'utilisation du CADx puisqu'il passe de 0.63 en lecture 1 à 0.78 lors de la lecture avec le CADx. Les scores attribués par les experts, indépendamment de leur niveau d'expertise, tendent à s'homogénéiser lorsque le score de malignité calculé par le CADx est pris en compte.

Lecteur	1	2	3	4	5	6	7	8	9	10	11	12
Corrélation	0.87	0.81	0.90	0.81	0.76	0.83	0.72	0.69	0.74	0.79	0.39	0.79

(a) Corrélation intra-lecteur

Lecteur	senior	junior	tous
Corrélation	0.74	0.77	0.76

(b) Corrélation intra-expert suivant l'expérience

Table *9.3* – *Etude de la corrélation intra-experts mesurée entre les lectures 1 et 2 (a) pour chaque lecteur (b) en moyenne pour les lecteurs juniors et seniors, et pour tous les lecteurs confondus.*

Lecteur	2	3	4	5	6	7	8	9	10	11	12
1	0,77	0,76	0,68	0,73	0,72	0,61	0,65	0,73	0,79	0,46	0,70
2		0,72	0,68	0,67	0,71	0,62	0,61	0,68	0,73	0,47	0,71
3			0,64	0,77	0,67	0,62	0,63	0,71	0,77	0,34	0,74
4				0,68	0,58	0,59	0,68	0,70	0,74	0,40	0,60
5					0,66	0,63	0,68	0,70	0,79	0,28	0,71
6						0,60	0,64	0,65	0,67	0,58	0,61
7							0,66	0,55	0,61	0,28	0,64
8								0,71	0,65	0,39	0,66
9									0,70	0,38	0,68
10										0,33	0,72
11											0,25

(a) Corrélation inter-lecteur

Lecteur	senior	junior
senior	0.58	0.64
junior		0.68
moyenne	0.63	

(b) Corrélation inter-expert suivant l'expérience

Table *9.4* – *Etude de la corrélation inter-experts mesurée sur la lecture 1 (a) entre chaque lecteur (b) en moyenne entre les lecteurs juniors et seniors et sur l'ensemble des lecteurs.*

Lecteur	2	3	4	5	6	7	8	9	10	11	12
1	0,90	0,89	0,89	0,85	0,76	0,88	0,82	0,81	0,86	0,74	0,81
2		0,88	0,86	0,81	0,74	0,89	0,77	0,80	0,92	0,70	0,77
3			0,88	0,82	0,75	0,84	0,82	0,77	0,85	0,70	0,80
4				0,79	0,76	0,84	0,83	0,82	0,82	0,65	0,81
5					0,72	0,79	0,72	0,75	0,82	0,66	0,74
6						0,73	0,74	0,68	0,71	0,68	0,67
7							0,79	0,85	0,88	0,74	0,79
8								0,81	0,73	0,68	0,81
9									0,80	0,70	0,75
10										0,70	0,73
11											0,67

(a) Corrélation inter-lecteur

Lecteur	senior	junior
senior	0.77	0.79
junior		0.78
moyenne	0.78	

(b) Corrélation inter-expert suivant l'expérience

Table *9.5* – *Etude de la corrélation inter-experts mesurée sur la lecture avec CADx (a) entre chaque lecteur (b) en moyenne entre les lecteurs juniors et seniors et sur l'ensemble des lecteurs.*

9.3.2 Influence du CADx sur la confiance des lecteurs

On étudie dans cette partie l'évolution de la confiance du lecteur dans l'interprétation des images, c'est-à-dire sa propension à coder "0" et "4" avec et sans l'avis du CADx. On cherche ainsi à estimer si les lecteurs se prononcent de façon plus tranchée sur la classe des cibles, en attribuant moins le score "2". On aimerait également savoir si les scores attribués aux vrais positifs (VP) sont plus proches de "4" et si les scores attribués aux vrais négatifs (VN) sont plus proches de "0".

Le tableau 9.6 répertorie les changements réalisés, en moyenne sur toutes les cibles, dans l'attribution des scores entre la lecture 2 et la lecture avec CADx. On observe une forte variabilité intra-observateur dans l'interprétation des images et une influence notable du CADx dans la prise de décision : le score "2" est sensiblement moins attribué après avis du CADx (-6% en moyenne), le score "4" l'est un peu plus (+ 2% en moyenne) et le score "0" l'est beaucoup plus (+12% en moyenne).

Lecteur	1	2	3	4	5	6	7	8	9	10	11	12
score "0"	+11%	+14%	+9%	+7%	+11%	+2%	+14%	+22%	+18%	+17%	+6%	+11%
score "1"	0	0	-3%	+5%	-2%	-1%	+7%	-14%	-15%	-2%	+1%	-13%
score "2"	-11%	-9%	-7%	+6%	-11%	0	-10%	-9%	-7%	-11%	-1%	-1%
score "3"	-1%	-5%	+2%	-16%	-3%	-6%	-9%	+1%	0	-10%	-2%	-1%
score "4"	+1%	+5%	-1%	-1%	+6%	+5%	-1%	0	+3%	+7%	-3%	+3%

Table *9.6* *– Evolution de la répartition des scores avant et après prise en compte de l'"avis' du CAD. Pour chaque score $i \in [0, 1, 2, 3, 4]$, on calcule :*

$$\sum_{\text{lecture avec CAD}} Ind(score{=}i) - \sum_{\text{lecture 2}} Ind(score{=}i),$$

normalisé par le nombre total de cas.

Approche statistique

Un modèle logistique hiérarchique à effet mixte a été proposé pour étudier l'évolution de la propension des lecteurs à attribuer un score de "0" pour les cas sains et un score de "4" pour les cas pathologiques, entre les lectures 1, 2 et avec CADx. L'objectif est donc d'étudier la capacité de prise de décision (i.e. la confiance dans le diagnostic) en fonction de la lecture, seul ou avec l'avis du CADx. En reformulant, les deux objectifs de cette section sont :

a) de comparer entre deux lectures la propension à coder {0} (défini comme le "succès") versus {1, 2, 3, 4} (défini comme "l'échec"), chez les cibles saines ;

b) de comparer entre deux lectures la propension à coder {4} (le "succès") versus {0, 1, 2, 3,} ("l'échec"), chez les cibles pathologiques.

Les méthodologies relatives à ces deux objectifs sont strictement identiques. Ainsi, seule la méthodologie permettant d'étudier la propension à coder "0" sur les cibles saines est décrite ici.

Dans le modèle logistique hiérarchique utilisé, la probabilité du "succès", i.e. la propension à coder "0" sur une cible saine, a été modélisée en fonction de la lecture (1, 2 ou avec CADx) et du lecteur (n°$1, \ldots, 12$). Le modèle utilisé permet d'évaluer l'impact de la lecture (1, 2 ou avec CADx) et de l'expérience des lecteurs (lecteurs juniors versus lecteurs expérimentés) sur la propension à coder "0". De plus, la hiérarchie des données a été prise en compte, en considérant dans ce modèle différents niveaux :

- le niveau patient ;
- le niveau cible (à un patient peut correspondre plusieurs cibles) ;
- le niveau "mesure" qui correspond au résultat obtenu d'une lecture par un lecteur sur une cible ;
- le niveau lecteur : chaque lecteur a évalué toutes les cibles lors des trois lectures.

Le caractère hiérarchique du modèle permet de prendre en compte les différentes corrélations entre scores (inter-cible, inter-lecteur...). Par exemple, il est assez fréquent qu'une cible donnée soit assez facilement identifiable comme une cible non-cancéreuse, tandis que le diagnostic pour une deuxième soit plus délicat à poser. Les lecteurs seront globalement tous plus enclins à coder "0" pour la première mais ils auront moins tendance à coder "0" pour la deuxième. De manière similaire, la propension à coder "0" pour un lecteur bénéficiant d'une solide expérience sera plus élevée que pour un lecteur débutant ou avec peu de d'expérience.
Dans ce modèle, les propensions à coder "0" entre deux lectures sont comparées en termes de rapport de cotes (*odds ratio* en anglais). La cote d'une probabilité $\mathbb{P}$ est définie par $\frac{\mathbb{P}}{1-\mathbb{P}}$ et peut être vue intuitivement comme le nombre de "succès" divisé par le nombre d'"échecs". Le rapport de cotes de la probabilité de "succès" $\mathbb{P}_A$ dans une catégorie A par rapport à la probabilité de "succès" $\mathbb{P}_B$ dans une catégorie de référence B est le rapport :

$$\frac{\dfrac{\mathbb{P}_A}{1-\mathbb{P}_A}}{\dfrac{\mathbb{P}_B}{1-\mathbb{P}_B}} \tag{9.1}$$

Un *odds ratio* valant k indique que le nombre de "succès" dans la catégorie A est k fois supérieur à celui dans la catégorie B, ramené à un même nombre fixé d'"échecs" dans chacune des catégories A et B. Un *odds ratio* de 1 correspond au cas où les probabilités $\mathbb{P}_A$ et $\mathbb{P}$ sont égales.

L'annexe B section B.3 comprend une introduction aux modèles de régression linéaire généralisée pour un critère binaire, une description détaillée du modèle utilisé dans cette partie et de l'interprétation des paramètres choisis, ainsi que l'ensemble des résultats obtenus.
L'analyse statistique a été réalisée avec le logiciel R® (the R foundation, Vienne, Autriche). Dans la suite, les p-valeurs ≤ 0.05 sont considérées comme indicatrices d'une différence significative d'un point de vue "statistique".

Résultats : évolution de la propension à coder 0 chez les cibles bénignes

D'après le modèle statistique précédemment introduit, on obtient que :

- la propension à coder 0 entre les lectures 1 et 2 est similaire avec une très légère tendance à coder moins de fois 0 lors de la lecture 1 (odd-ratio : 0.86) ;
- la propension à coder 0 est bien plus importante pour la lecture avec CADx par rapport à la lecture 2 (odd-ratio : 7.02). Cet odd-ratio est significativement différent de 1 au seuil de 5% (p-valeur < 0.0001) ;
- la propension à coder 0 est inférieure pour les lecteurs juniors (odd-ratio : 0.25). En termes d'ampleur d'effet, les différences de propension à coder 0 sont importantes entre lecteurs juniors et seniors.

Ces résultats sont détaillés section B.3.2.

A partir des estimations des paramètres du modèle statistique, les propensions globales à coder "0" pour les cibles saines ont été prédites par lecture et expérience des lecteurs. Elles sont données dans la tableau 9.7.

	Tous	**Junior**	**Senior**
Lecture 1	4.5 [1.4 ;13]	2.6 [0.6 ;8.7]	9.7 [2.6 ;29.5]
Lecture 2	5.3 [1.7 ;14]	3 [0.8 ;9.6]	11.1 [3 ;31.6]
Lecture avec CADx	28.4 [11.5 ;54.8]	17.9 [5.1 ;43.8]	46.8 [16.5 ;78.8]

Table *9.7* – *Estimations (IC à 95%) des propensions globales à coder "0" pour des cibles saines par lecture et expérience des lecteurs, exprimées en pourcentage.*

Résultats : évolution de la propension à coder 4 chez les cibles malignes

D'après le modèle statistique précédemment introduit, on obtient que :

- la propension à coder 4 était plus faible lors de la lecture 1 par rapport à la lecture 2 (odd-ratio : 0.58). Cet odd-ratio n'est pas significativement différent de 1 au seuil de 5% (p-valeur de 0.08) ;
- la propension à coder 4 est plus importante pour la lecture avec CADx par rapport à la lecture 2 (odd-ratio : 1.61). Cet odd-ratio est significativement, d'un point de vue "statistique", différent de 1 au seuil de 5% (p-valeur de 0.036) ;
- la propension à coder 4 était similaire chez les lecteurs juniors et seniors (odd-ratio : 1.19).

Ces résultats sont détaillés section B.3.3.

A partir des estimations des paramètres du modèle statistique, les propensions globales à coder "4" pour les cibles pathologiques ont été prédites par lecture et expérience des lecteurs. Elles sont données dans le tableau 9.8.

	Tous	**Junior**	**Senior**
Lecture 1	16 [4 ;42.3]	17 [3.5 ;44.1]	14.7 [3.4 ;43.9]
Lecture 2	24.8 [6.3 ;55.3]	26.2 [6.2 ;58.6]	23 [5.6 ;60.2]
Lecture avec CADx	34.6 [10.4 ;65.4]	36.3 [10.4 ;68.4]	32.4 [9.1 ;69]

Table *9.8* – *Estimations [IC à 95%] des propensions globales à coder "4" pour des cibles pathologiques par lecture et expérience des lecteurs, exprimées en pourcentage.*

9.3.3 Impact du CADx sur les performances diagnostiques

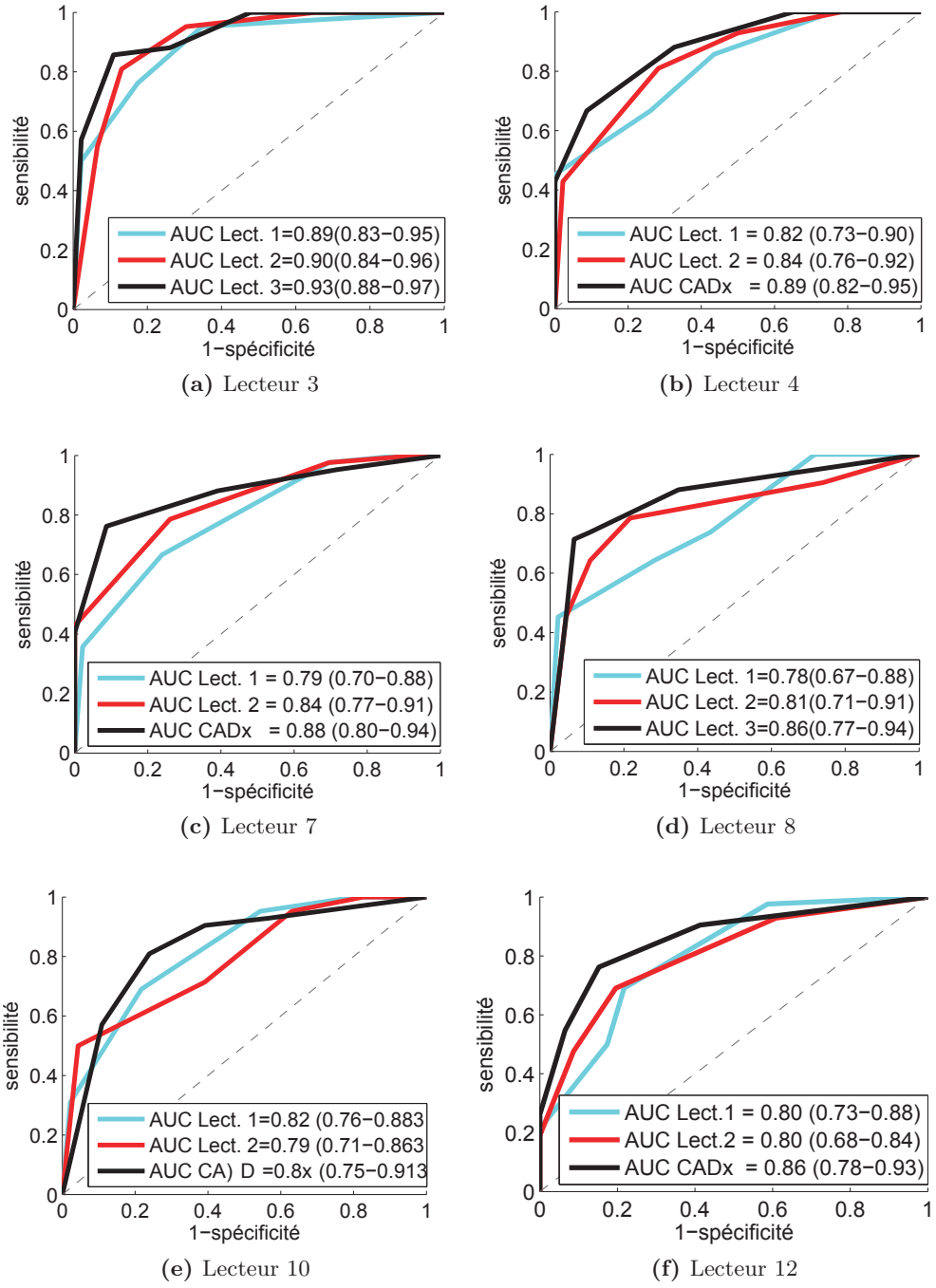
Courbes ROC expérimentales

La figure 9.1 montre l'évolution des courbes ROC mesurées au cours des trois lectures, pour six lecteurs : 3 juniors (lecteurs 7, 8, 10) et 3 seniors (lecteurs 3, 4, 12).
Nous comparons les performances diagnostiques, mesurées en termes d'aire sous la courbe ROC (AUC), réalisées par les différents lecteurs au cours des différentes lectures. Le tableau 9.9 liste les valeurs d'AUC ainsi que les intervalles de confiance à 5% correspondants. Ces valeurs sont calculées à l'aide d'une méthode d'estimation non-paramétrique présentée dans l'annexe B.1 et reposant sur l'article de Rutter [106], prenant en compte le fait que les données sont *clusterisées* (plusieurs ROIs par patient). Le graphique 9.2 permet de visualiser l'évolution des valeurs d'AUC au cours des trois lectures en distinguant les lecteurs juniors des lecteurs seniors.

	AUC			**Différence d'AUC**
	Lecture 1	Lecture 2	Lecture CAD	Lecture CAD- Lecture 2
Lecteur 1	82.4 [74.8 ;89.6]	85.8 [78.5 ;92.8]	88.1 [81.5 ;94]	2.3 [-2.9 ; 7.2]
Lecteur 2	84.7 [78.3 ;90.6]	86.1 [79.2 ;92.1]	87.2 [80 ;93.9]	1.2 [-5.8 ; 8.3]
Lecteur 3	88.9 [82.6 ;94.6]	90.5 [84 ;96.2]	92.7 [87.5 ;96.9]	2.2 [-2.5 ; 7.1]
Lecteur 4	81.9 [73.3 ;89.7]	84.5 [75.8 ;91.9]	88.8 [82 ;94.5]	4.4 [-0.7 ; 9.9]
Lecteur 5	80.2 [71.3 ;88.1]	80.3 [71.2 ;88.6]	84.8 [76.6 ;92]	4.5 [-0.4 ; 10]
Lecteur 6	79.0 [70.8 ;86.9]	76.6 [66.5 ;85.9]	81.2 [72 ;89.4]	4.6 [-3.5 ; 12.2]
Lecteur 7	79.2 [69.9 ;87.5]	84.3 [76.7 ;91.1]	87.8 [79.7 ;94.4]	3.4 [-2.2 ; 9.4]
Lecteur 8	78.2 [66.8 ;88.3]	81.4 [71.2 ;90.8]	86.2 [77 ;94.1]	4.8 [-2.9 ; 13.7]
Lecteur 9	79.2 [71.8 ;86.5]	77.6 [68.3 ;85.9]	82.7 [74.1 ;90.6]	5.1 [-0.9 ; 11.2]
Lecteur 10	82.0 [75.7 ;87.8]	78.8 [71.1 ;86.4]	83.1 [74.9 ;90.7]	4.3 [-4.6 ; 13.5]
Lecteur 11	64.2 [55.1 ;72.9]	75.8 [67.6 ;83.9]	79.5 [71.7 ;86.6]	3.7 [-0.8 ; 8.5]
Lecteur 12	80.6 [73.1 ;87.6]	80.6 [72.6 ;88.4]	86.1 [78 ;93.1]	5.5 [1.6 ; 9.7]

Table *9.9* – *AUC estimées, avec intervalles de confiance (IC), pour chaque lecteur lors des 3 lectures*

Ces résultats illustrent la diversité des performances diagnostiques selon le lecteur. Hormis le cas particulier du lecteur 11 présentant un gros écart entre lectures 1 et 2, aucune tendance nette de supériorité des performances d'une des deux lectures 1 ou 2 ne se dégage, tandis que les performances des lecteurs lors de la lecture avec CADx sont toutes supérieures aux performances mesurées lors des lectures 1 et 2. Les gains de performance entre lectures 2 et avec CADx sont plus ou moins importants (cf. table 9.9 ci-dessus), selon



(a) Lecteur 3 **(b)** Lecteur 4

(c) Lecteur 7 **(d)** Lecteur 8

(e) Lecteur 10 **(f)** Lecteur 12

Figure *9.1* – *Evolution des courbes ROC au cours des trois lectures pour six lecteurs : 3 juniors (lecteurs 7, 8 10) et 3 seniors (lecteurs 3, 4, 12).*

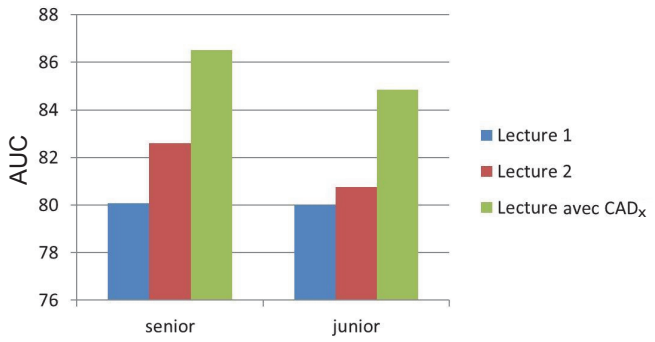


Figure *9.2* – *Evolution des valeurs d'AUC seniors/juniors au cours des trois lectures.*

les lecteurs et varient de 1.2 à 5.5 points.

Modélisation des courbes ROC

Afin de comparer, d'un point de vue "statistique", les performances diagnostiques selon la lecture, les courbes ROC ont été modélisées en utilisant un modèle de régression introduit par Alonzo et Pepe [3]. En annexe B section B.2 sont disponibles une description de la méthode proposée par Alonzo et Pepe, le détail du modèle retenu pour la modélisation des courbes ROC et les résultats complets obtenus.

Les courbes ROC ont été modélisées par des courbes binormales admettant la forme paramétrique suivante :

$$\mathrm{ROC}_{(l,k)}(t) = \Phi\Big(a_{l,k} + b_{l,k}\Phi^{-1}(t)\Big), \tag{9.2}$$

pour t une fraction de faux positifs (correspondant à une valeur de 1-spécificité) comprise entre 0 et 1, et où :

- Φ désigne la fonction de répartition d'une variable aléatoire Z de loi normale de moyenne 0 et de variance 1 ;
- l désigne un lecteur et k une lecture.

Le paramètre $a_{l,k}$ est le paramètre de "localisation" tandis que $b_{l,k}$ représente la "pente".

Une fois obtenues les estimations de ces paramètres, les AUC des courbes ROC sont estimées par :

$$\Phi\Big(\frac{\hat{a}_{l,k}}{\sqrt{1+\hat{b}_{l,k}^2}}\Big). \tag{9.3}$$

Résultats : comparaison des performances diagnostiques entre lectures

Les résultats présentés ici sont issus du modèle détaillé en annexe B partie B.2.3.
Les performances diagnostiques des lectures 1, 2 et avec CADx sont comparées sur l'ensemble des lecteurs. L'unique covariable considérée est donc le type de lecture (les lecteurs

sont pris en compte comme des effets aléatoires dans le modèle).

Les performances des lectures 1 et 2 sont similaires, tandis que la lecture avec le CADx se démarque des deux autres lectures par des performances supérieures. Cependant, notre étude n'a pas permis de mettre en évidence des différences statistiquement significatives (ce qui peut être simplement dû à un manque de puissance s'expliquant par le nombre limité de cibles considérées...).

Les AUC des courbes ROC ont été estimées à 83 [77.9 ;88] pour la lecture 1, à 84.1 [78.1 ;88.7] pour la lecture 2 et à 87.2 [81 ;92] pour la lecture 2 avec le CADx. L'AUC de la courbe ROC pour la lecture 2 est ainsi supérieure de 1.1 [-5.8 ;8.4] à l'AUC de la lecture 1 et inférieure de 3 [0 ;6.6] à l'AUC de la lecture avec CADx. L'AUC de la lecture 2 avec CADx est supérieure de 4.2 de celle de la lecture 1. Ces résultats sont résumés dans les tableaux 9.10 et 9.11.

	Lecture 1	**Lecture 2**	**Lecture 2 avec CADx**
Tous lecteurs confondus	83 [77.9 ;88]	84.1 [78.1 ;88.7]	87.2 [81 ;92]

Table *9.10* *– AUC estimées, avec IC, à partir des courbes ROC modélisées, pour les lectures* 1*,* 2*, et avec CADx (globalement pour l'ensemble des lecteurs)*

Lectures	**différence d'AUC**	**Erreur Standard**	**p-value**
Lecture 2 - Lecture 1	1.1 [-5.8 ;8.4]	3.7	0.763
Lecture CADx - Lecture 1	4.2 [-3.4 ;11.5]	3.8	0.2736
Lecture CADx - Lecture 2	3.0 [0 ;6.6]	1.7	0.0769

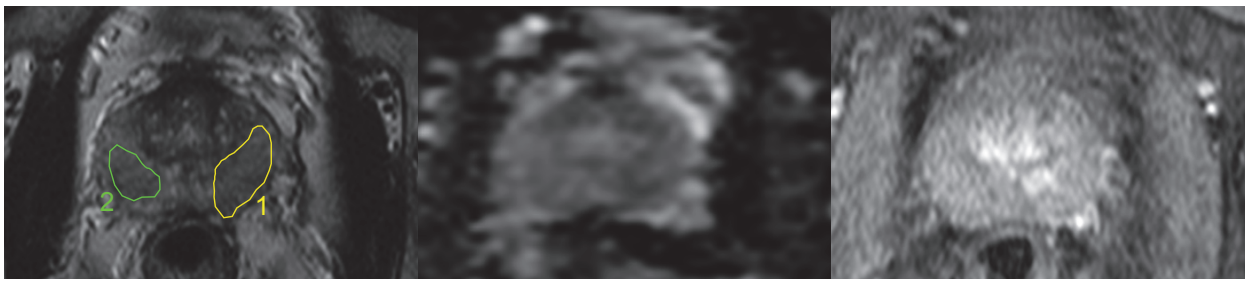
Table *9.11* *– Différences estimées d'AUC, avec IC, à partir des courbes ROC modélisées, entre les lectures* 1 *vs* 2*,* 1 *vs CAD, et* 2 *vs CAD (globalement pour l'ensemble des lecteurs)*

9.4 Discussion

Les résultats obtenus montrent, de manière systématique, une amélioration des performances de discrimination, mesurées par l'aire sous la courbe ROC (AUC) et cela même chez les radiologues les plus expérimentés. Les résultats de la modélisation des courbes ROC ont montré un gain d'AUC substantiel de 3 points globalement sur tous les lecteurs. On observe également une augmentation systématique de la confiance dans le diagnostic, que nous avons mesuré via l'étude des propensions à coder "0", (respectivement "4") sur les cas sains (respectivement malins). Le CADx permet également de diminuer la variabilité inter-experts, indépendamment de leur niveau d'expertise. On remarque également que tous les lecteurs, quel que soit leur niveau d'expertise, ont pris en compte l'avis du CADx, manifestant ainsi leur confiance dans un tel système. Dans l'ensemble, les résultats sont donc positifs et prometteurs quant à l'avenir d'un tel système.

9.5 Exemples de cas étudiés

Nous présentons ici quelques exemples de cas inclus dans cette étude. Bien entendu, dans les conditions de l'étude, les images sont en 3-dimensions, la séquence dynamique (DCE) est 3-D + temps (temps total de 3 minutes) et les contours des cibles à analyser, ici affichés sur la séquence T2-w, peuvent être supprimés/re-affichés à la guise du lecteur pour plus de lisibilité.



(a) T2-w **(b)** ADC **(c)** DCE

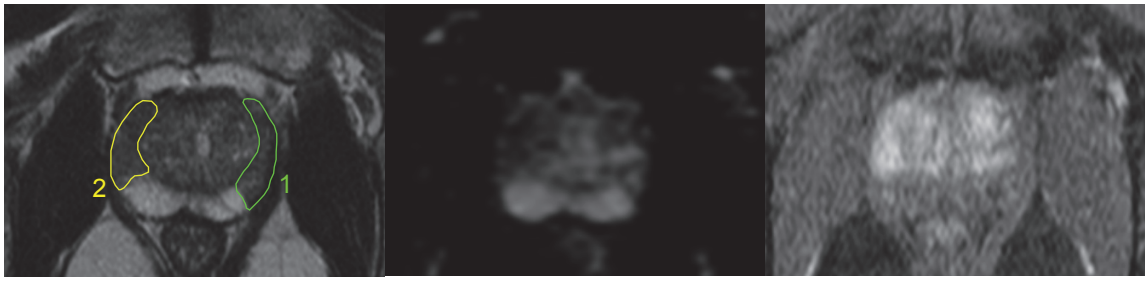
Figure *9.3* – *Exemple 1 : cas clinique présentant deux cibles malignes 1 et 2 pour lesquelles le CADx renvoie respectivement un score de 0.93(/1) et 0.98(/1).*

	Lecteur n°	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	
Lecture 1	*Score cible 1*	*4*	*4*	*4*	*4*	*3*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	
	Score cible 2	*1*	*2*	*0*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*2*	*1*	
Lecture 2	*Score cible 1*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	
	Score cible 2	*2*	*2*	*1*	*1*	*2*	*1*	*2*	*0*	*1*	*2*	*2*	*1*	***CADx***
Lecture avec CADx	*Score cible 1*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	***0.93***
	Score cible 2	*2*	*3*	*3*	*3*	*3*	*1*	*3*	*2*	*2*	*3*	*2*	*2*	***0.98***

(a) T2-w **(b)** ADC **(c)** DCE

Figure *9.4* – *Exemple 2 : cas clinique présentant une cible bénigne pour laquelle le CADx renvoie un score de 0.1(/1).*

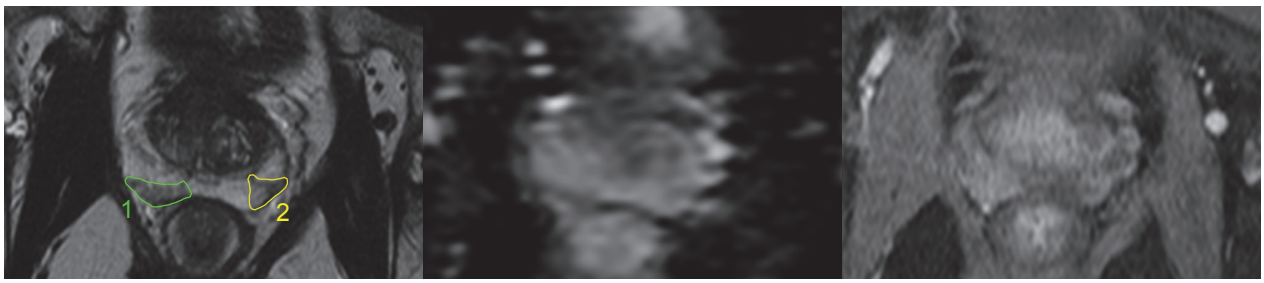
	Lecteur n°	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	
Lecture 1	*Score cible 1*	*1*	*2*	*1*	*1*	*3*	*3*	*1*	*3*	*2*	*3*	*2*	*1*	
Lecture 2	*Score cible 1*	*1*	*3*	*1*	*3*	*2*	*3*	*1*	*3*	*2*	*3*	*2*	*1*	***CADx***
Lecture avec CADx	*Score cible 1*	*0*	*2*	*1*	*1*	*2*	*3*	*0*	*1*	*2*	*3*	*2*	*0*	***0.1***



(a) T2-w **(b)** ADC **(c)** DCE

Figure *9.5* – *Exemple 3 : cas clinique présentant deux cibles, l'une maligne, l'autre bénigne, pour lesquelles le CADx renvoie respectivement un score de 1(/1) et 0.2(/1).*

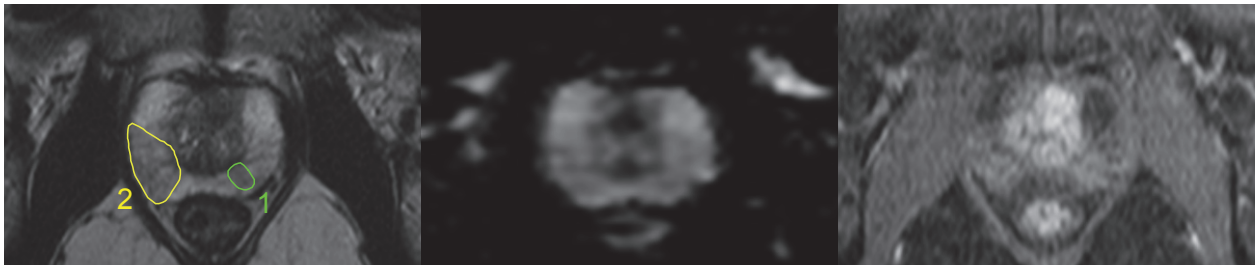
Lecteur n°		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	
Lecture 1	*Score cible 1*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*3*	*4*	*4*	*3*	
	Score cible 2	*1*	*1*	*1*	*1*	*1*	*1*	*2*	*3*	*1*	*2*	*4*	*1*	
Lecture 2	*Score cible 1*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*3*	
	Score cible 2	*1*	*2*	*1*	*2*	*2*	*2*	*2*	*1*	*1*	*2*	*2*	*1*	***CADx***
Lecture avec CADx	*Score cible 1*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	***1***
	Score cible 2	*0*	*1*	*1*	*2*	*1*	*2*	*2*	*1*	*1*	*2*	*2*	*0*	***0.2***



(a) T2-w **(b)** ADC **(c)** DCE

Figure *9.6* – *Exemple 4 : cas clinique présentant deux cibles malignes pour lesquelles le CADx renvoie respectivement un score de 0.87(/1) et 0.99(/1).*

Lecteur n°		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	
Lecture 1	*Score cible 1*	*2*	*2*	*3*	*3*	*3*	*2*	*3*	*2*	*1*	*2*	*2*	*1*	
	Score cible 2	*4*	*4*	*4*	*4*	*2*	*4*	*4*	*4*	*3*	*4*	*4*	*3*	
Lecture 2	*Score cible 1*	*2*	*3*	*3*	*3*	*2*	*3*	*4*	*3*	*1*	*2*	*3*	*1*	
	Score cible 2	*3*	*4*	*4*	*4*	*3*	*4*	*3*	*4*	*4*	*4*	*4*	*3*	***CADx***
Lecture avec CADx	*Score cible 1*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*4*	*3*	*3*	*3*	*2*	***0.87***
	Score cible 2	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*3*	***0.99***



(a) T2-w **(b)** ADC **(c)** DCE

Figure *9.7* – *Exemple 5 : cas clinique présentant deux cibles bénignes pour lesquelles le CADx renvoie un score de 0(/1).*

	Lecteur n°	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	
Lecture 1	*Score cible 1*	*3*	*2*	*1*	*3*	*2*	*3*	*2*	*3*	*3*	*2*	*3*	*3*	
	Score cible 2	*2*	*2*	*0*	*3*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*0*	
Lecture 2	*Score cible 1*	*2*	*2*	*0*	*2*	*2*	*2*	*3*	*2*	*3*	*3*	*3*	*2*	
	Score cible 2	*1*	*2*	*0*	*2*	*2*	*2*	*2*	*1*	*2*	*3*	*2*	*1*	***CADx***
Lecture avec CADx	*Score cible 1*	*1*	*1*	*0*	*1*	*0*	*1*	*1*	*0*	*3*	*0*	*3*	*1*	***0***
	Score cible 2	*1*	*1*	*0*	*1*	*0*	*1*	*1*	*0*	*1*	*0*	*2*	*0*	***0***

(a) T2-w **(b)** ADC **(c)** DCE

Figure *9.8* – *Exemple 6 : cas clinique présentant deux cibles malignes pour lesquelles le CADx renvoie respectivement un score de 0.99(/1) et 1(/1).*

	Lecteur n°	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	
Lecture 1	*Score cible 1*	*3*	*4*	*4*	*4*	*2*	*3*	*3*	*4*	*4*	*3*	*3*	*3*	
	Score cible 2	*3*	*4*	*3*	*3*	*3*	*4*	*4*	*4*	*3*	*3*	*3*	*3*	
Lecture 2	*Score cible 1*	*3*	*4*	*4*	*4*	*3*	*3*	*4*	*4*	*4*	*4*	*4*	*3*	
	Score cible 2	*3*	*3*	*4*	*4*	*2*	*3*	*4*	*4*	*3*	*3*	*4*	*2*	***CADx***
Lecture avec CADx	*Score cible 1*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	***0.99***
	Score cible 2	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*3*	***1***

III SVM étendu au cas des étiquettes incertaines

Chapitre 10

Introduction

10.1 Contexte et motivation

Corrélation anatomo-radiologique : une tâche complexe

L'analyse des limites du système expert présenté dans la partie II (chapitre 8) et des schémas classiques de systèmes supervisés en général souligne l'influence majeure de la base d'apprentissage sur les performances. Celle-ci doit en effet être à la fois riche (i.e. contenir un grand nombre d'exemples) et fiable (i.e. reposer sur une vérité terrain indiscutable, la référence histologique dans notre problématique).

Pour tous les patients constituant la base de données sur laquelle s'appuie l'étude précédemment présentée (lire chapitre 6), la présence de foyer(s) malin(s) a été révélée par biopsie. La localisation et l'extension de ce(s) foyer(s) a ensuite été vérifiée *a posteriori* par analyse de la pièce de prostatectomie, permettant d'établir une cartographie précise du cancer sur la glande. Finalement, les coupes histologiques et les images IRM-mp ont été analysées conjointement et les régions malignes repérées lors de l'analyse histologique reportées, après consensus entre radiologues et anatomo-pathologistes, sur les images IRM-mp (lire section 6.5).

L'établissement d'une telle base de données, où les lésions malignes sont annotées sur les images IRM-mp de manière précise et exhaustive, est une tâche difficile et fastidieuse. Elle nécessite 1) d'avoir accès à la pièce histologique et 2) la mobilisation de radiologues et d'anatomo-pathologistes pour permettre une corrélation anatomo-radiologique.

Comme nous l'avons vu dans la section 4.7 dédiée à l'analyse de l'état de l'art, de nombreuses études CAD doivent donc se limiter à l'utilisation de jeux de données annotées de taille parfois très restreinte. Par exemple, les travaux de Madabushi *et coll.* [67, 68] s'appuient sur les données IRM histologiques de seulement 5 patients (dont seules 33 coupes au total sont étudiées) ; l'étude présentée par Viswanath *et coll.* [131] repose quant à elle sur les données de 6 patients (dont seulement 18 coupes sont étudiées).

Enfin, beaucoup d'études (voir en particulier les études [18,75,93,124] présentées section 4.7) ne peuvent pas s'appuyer sur une telle corrélation anatomo-radiologique et la vérité terrain utilisée est souvent le résultat :

- soit d'une analyse histologique partielle utilisant les résultats des biopsies *randomisées* (qui ne représente qu'un échantillonnage de la glande...), à partir de laquelle le radiologue positionne, seul, les foyers sur les images IRM,
- soit d'une analyse *en aveugle* (et donc subjective...) des images IRM-mp, réalisée par un radiologue expert.

Toutes deux sont entachées d'incertitude quant au véritable statut carcinologique des cibles (ROIs) contourées.
Néanmoins, ces études qui ne reposent que sur une "vérité expert" (par opposition à la "vérité histologique"), ont l'avantage de pouvoir présenter des bases de données plus riches en nombre de cas, avantage non négligeable en classification supervisée. Par exemple, l'étude de Puech *et coll.* [93] utilisent les données de 84 patients, dont 47 sont annotées par un radiologue expert.

La figure 10.1 illustre les différentes étapes de la constitution d'une base de données d'apprentissage clinique pour le cancer de la prostate.

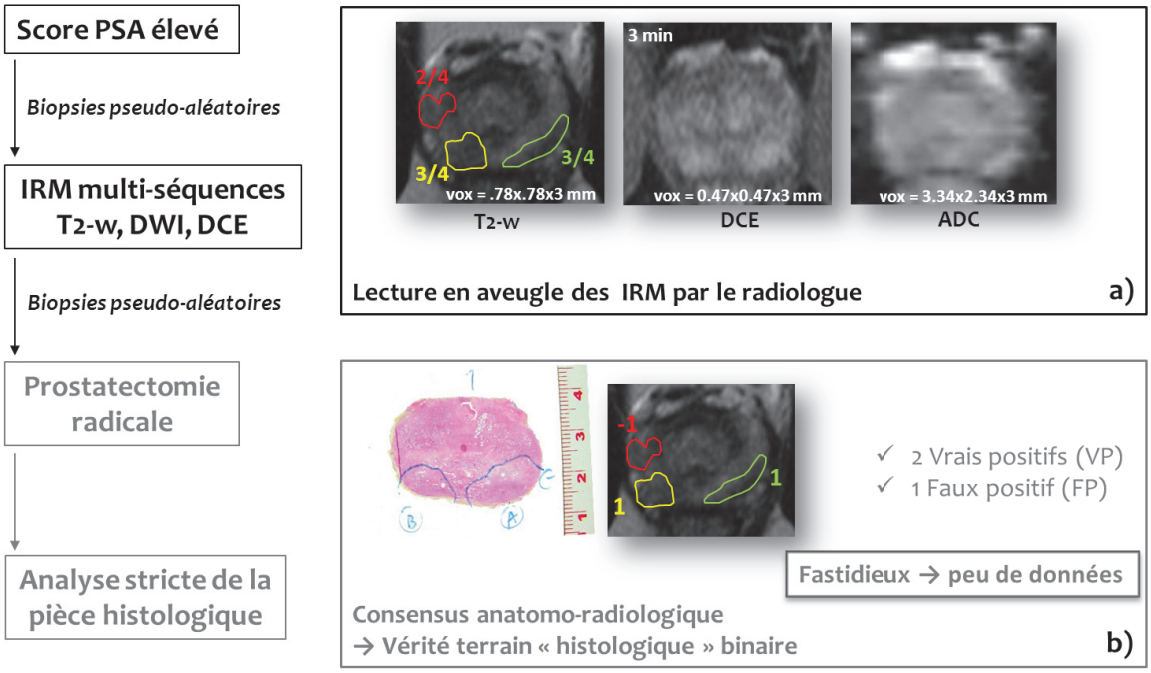


Figure *10.1* – *Construction d'une base de données cliniques.*
On rappelle qu'une mesure élevée du taux de PSA dans le sang peut être un marqueur de la présence de lésions cancéreuses dans la prostate. C'est donc un résultat élevé au test du PSA, qui engage le processus de recherche de présence de lésions cancéreuses dans la glande.
La référence actuelle pour le diagnostic du cancer de la prostate reste l'analyse histologique réalisée par le biais de biopsies "pseudo-aléatoires" qui seules attestent de la présence ou de l'absence de lésions cancéreuses sur la glande.
L'acquisition et l'analyse IRM-mp sont réalisées soit :
(1) avant les biopsies, pour les orienter dans les régions présentant un signal IRM suspect,
(2) en complément des biopsies, pour qualifier l'extension de la tumeur détectée à la biopsie et orienter le choix du traitement.
Le radiologue analyse ces images et repère les zones suspectes auxquelles il affecte un score de suspicion de malignité variant entre 0 et 4 (encart ***a****).*
Une fois la présence du cancer avérée, la chirurgie peut être envisagée. Si tel est le cas, une analyse de la pièce de prostatectomie peut être réalisée afin de localiser précisément les foyers malins et établir la "vérité terrain" (encart ***b****). Cette cartographie du cancer de la prostate peut alors être reportée sur les images IRM-mp après consensus anatomo-radiologique. Les parties grisées représentent les étapes qui ne sont pas réalisées de manière systématique : certaines études CAD ne reposent pas sur une analyse des pièces de prostatectomie et une corrélation anatomo-radiologique pour constituer la base d'apprentissage.*

Utilisation des scores experts

Lors de l'acquisition des images IRM-mp (voir illustration 10.1, a) le radiologue analyse conjointement les signaux des différentes séquences, et repère les ROIs qui lui paraissent suspectes. La pratique standard, recommandée par les directives européennes [6, 27], consiste alors à leur affecter un score subjectif à 5 points (échelle de Likert) reflétant le degré de suspicion de malignité. Ce score est échelonné de la manière suivante : 0 = bénignité certaine, 1 = bénignité probable, 2 = score intermédiaire d'incertitude, 3 = malignité probable et 4 = malignité certaine.

Aucune des études CAD précédemment présentées section 4.7 n'exploite actuellement ce score. Afin de pouvoir s'appuyer sur des algorithmes de classification "classiques" (lire section 4.4) c'est en effet une version binaire (sain/pathologique) de l'analyse radiologique experte réalisée *en aveugle* qui est utilisée. Celle-ci peut être vue comme un seuillage du score de suspicion.

Or, ce score est porteur d'informations sur le degré de certitude du médecin et la difficulté d'analyse de la ROI, il s'apparente à une probabilité d'appartenance à l'une ou l'autre des deux classes de tissus et mérite donc d'être exploité en l'absence de référence histologique fiable.

Objectif général

La plupart des études se limitent en nombre de cas étudiés car la construction d'une vérité terrain dans laquelle tous les tissus sains/pathologiques sont discriminés de manière fiable, en se basant sur la référence histologique, est très difficile et contraignante.

Or, c'est la richesse d'une base de données d'apprentissage qui permet de créer un classifieur robuste avec une bonne capacité de généralisation.

Au lieu de rejeter les données IRM-mp pour lesquelles la vérité histologique n'est pas accessible ou incertaine, nous proposons au contraire de les inclure dans l'apprentissage en exploitant le score de suspicion affecté par le radiologue lors de son analyse "en aveugle".

10.2 Proposition

L'objectif de ce travail est de pouvoir se rapprocher de la problématique clinique dans laquelle on essaie de construire un système de classification de données en deux classes (sain/ pathologique), en se basant sur une base de données annotées par l'homme et donc possiblement entachée d'incertitude. Ce degré d'incertitude peut être directement estimé par l'expert en charge de l'annotation qui l'exprime par le biais d'une probabilité d'appartenance à une classe.

Dans la partie II précédente, nous avons présenté les résultats prometteurs d'un système CADx reposant sur un algorithme de type séparateur à vaste marge (SVM). Or, si les SVM

sont des algorithmes de discrimination efficaces, classiquement utilisés pour discriminer deux classes de données étiquetées de manière binaire (sain/pathologique), ils ne peuvent cependant pas être appliqués directement dans le cas où les données contiennent à la fois des étiquettes de classe certaines (sain/pathologique) et des étiquettes de classe incertaines (telles que des probabilités d'appartenance à une classe).

Dans la seconde partie de ce travail de thèse, nous proposons une nouvelle formulation des SVM permettant d'intégrer l'incertitude de l'expert sur la classe de certains exemples. Cette incertitude peut être exprimée par le score de suspicion de malignité (échelle de Likert). L'idée est d'apprendre une fonction discriminante qui, à la fois, maximise les performances de classification sur les étiquettes certaines et prédit au mieux les probabilités sur les étiquettes incertaines. Il s'agit donc de moduler classification et régression.

A ce propos, il nous a été suggéré que dans ces travaux, nous nous attaquons à un problème ouvert proposé dans le livre *Learning with kernels* par B. Scholkopf and A. Smola [109]. Il est en effet proche du problème 7.11 de la page 223 de ce livre. Nous proposons une manière de résoudre ce problème en prenant en compte une information probabiliste.

La suite de ce chapitre est organisée comme suit. Nous rappelons dans un premier temps la formulation classique des SVM, puis décrivons l'extension proposée. L'évaluation de ce nouvel algorithme, que l'on nommera P-SVM (pour *Probabilistic* SVM) afin de faciliter les discussions, est d'abord réalisée à l'aide d'exemples jouets puis sur une série de données cliniques. Nous montrons que le P-SVM permet de pondérer l'influence des cas d'interprétation difficile et d'obtenir de meilleures performances que celles réalisées en utilisant une *vérité expert* binaire.

Nous tenons à souligner que le travail que nous proposons dans cette partie est le fruit d'une collaboration forte avec Stéphane Canu et Rémi Flamary du laboratoire LITIS de Rouen (France). Nous les en remercions sincèrement.

Chapitre 11

Le séparateur à vaste marge (SVM) : approche classique

11.1 Introduction

Ce premier chapitre est dédié à la description et la résolution de la formulation classique des SVM telle qu'elle a été proposée par Vapnik et Cortes en 1995 [23] et sur lesquelles nous nous appuierons chapitre 12 pour présenter notre développement. Nous nous intéressons uniquement à la classification binaire c'est-à-dire à la recherche d'une fonction de décision permettant de distinguer entre deux classes.

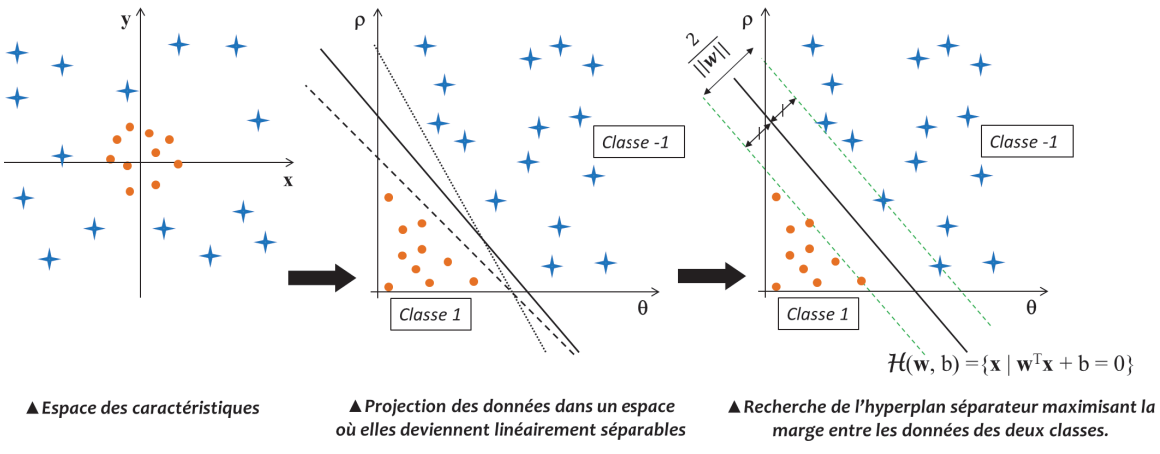


Figure *11.1* – *Schéma de principe des SVM*

Le séparateur à vaste marge (SVM) est un algorithme de classification supervisée qui, à partir d'un ensemble de points d'apprentissage étiquetés ± 1, cherche à construire un séparateur linéaire (hyperplan) définissant la frontière entre les classes "+1" et "−1". Notons que lorsqu'il existe une séparatrice linéaire entre des points d'apprentissage, il en existe en général une infinité ; néanmoins, l'objectif du SVM n'est pas seulement de **bien classifier** les exemples d'apprentissage mais aussi de **bien généraliser** à de nouveaux exemples. Vapnik *et coll.* [128] ont ainsi proposé de traduire la robustesse de la fonction de décision en considérant la marge séparant les exemples de la classe "+1" des exemples de la classe "-1". L'hyperplan séparateur optimal $\mathcal{H}$, qui sépare "au mieux" (au sens de Vapnik) les deux nuages de points, est construit de manière à maximiser la marge entre les exemples des deux classes. Dans le cas où les données d'apprentissage ne sont pas linéairement séparables dans leur espace de définition, l'idée est de se ramener à un espace de plus grande dimension, par l'intermédiaire d'une fonction de projection ϕ, dans lequel elles deviennent séparables. L'illustration 11.1 rappelle le principe des SVM.

Dans la section 11.3, nous rappelons la formulation des SVM linéaires introduite section 7.4.3, partie II (page 114). La section 11.4 décrit le passage à la formulation duale à partir de laquelle les formulations introduites dans le cas d'un problème linéairement séparable pourront être généralisées au cas non-linéaire en introduisant les fonctions noyaux, section 11.5. Finalement, la section 11.6 présente une méthode d'estimation des probabilités *a posteriori* d'appartenance à une classe, à partir des paramètres obtenus par SVM.

11.2 Notations

On rappelle que notre problème d'apprentissage supervisé par SVM repose sur un ensemble d'apprentissage $\mathcal{A}$ contenant un certain nombre n d'individus $\mathbf{x} \in \mathcal{X}$ où $\mathcal{X}$ définit l'espace des caractéristiques. Dans nos travaux, nous nous limitons à des espaces euclidiens de la forme $\mathcal{X} = \mathbb{R}^d$. A chaque observation $\mathbf{x}$ est associée une étiquette, ou classe, $y \in \mathcal{Y}$ que nous voulons arriver à prédire. Le problème de classification nous intéressant étant binaire, nous avons $\mathcal{Y} = \{-1, 1\}$. Finalement, l'ensemble d'apprentissage est composé de n couples d'observation/étiquette $(\mathbf{x}_i, y_i)$ tel que $\mathbf{x}_i = [x_1, x_2 \dots x_d] \in \mathbb{R}^d$ et $y_i \in \{-1, 1\}$, pour $i = 1 \dots n$.

11.3 Le problème des SVM linéaires (dans le primal)

Un séparateur à vaste marge est un discriminateur reposant sur la construction d'un hyperplan séparateur optimal de l'espace des données, au sens de la maximisation de la marge entre les exemples de classes différentes. La figure 11.2a illustre son principe.

Soit $\mathcal{H}$ un hyperplan de l'espace des données. Il est donné par une équation de la forme :

$$\mathcal{H}(\mathbf{w}, b) \;=\; \{\mathbf{x} \in \mathbb{R}^d |\; h(\mathbf{x}) = 0\},\ \text{où}$$

$$h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = \mathbf{w}^\top \mathbf{x} + b = \sum_{j=0}^{d} w_j x_j + b,$$

avec $\mathbf{w} \in \mathbb{R}^d$ un vecteur orthogonal à $\mathcal{H}$, et $b \in \mathbb{R}$ le biais.

La distance $d(\mathbf{x},\mathcal{H})$ d'un point $\mathbf{x}$ à l'hyperplan $\mathcal{H}(\mathbf{w}, b)$ est égale à $\frac{|\langle \mathbf{w},\mathbf{x} \rangle|}{||\mathbf{w}||}$ puisque $\mathbf{w}$ est orthogonal à $\mathcal{H}$.

On cherche la séparatrice optimale parmi toutes les séparatrices possibles comme étant celle qui sépare "au mieux" (au sens de Vapnik) les deux nuages de points d'étiquettes "-1" et "+1", c'est-à-dire l'hyperplan dont la distance minimale aux exemples d'apprentissage est la plus grande possible (pour maximiser le pouvoir de généralisation).

Cet hyperplan optimal est donné par :

$$\arg\max_{\mathbf{w},b} \ \min\{\|\mathbf{x} - \mathbf{x}_i\| \mid \mathbf{x} \in \mathbb{R}^d, h(\mathbf{x}) = 0 \text{ et } \mathbf{x}_i \in \mathcal{A}, i = 1 \dots n\}.$$

On normalise $\mathbf{w}$ et b de façon à ce que la demi-marge (i.e. la distance au(x) point(s) le(s) plus proche(s) de l'hyperplan, voir figure 11.2a) vaille $\frac{1}{\|\mathbf{w}\|}$ (lire section 7.4.3 page 114).
Maximiser la marge entre les deux classes revient donc à minimiser $\|\mathbf{w}\|$ sous les contraintes de bonne classification (hyperplan séparateur) :

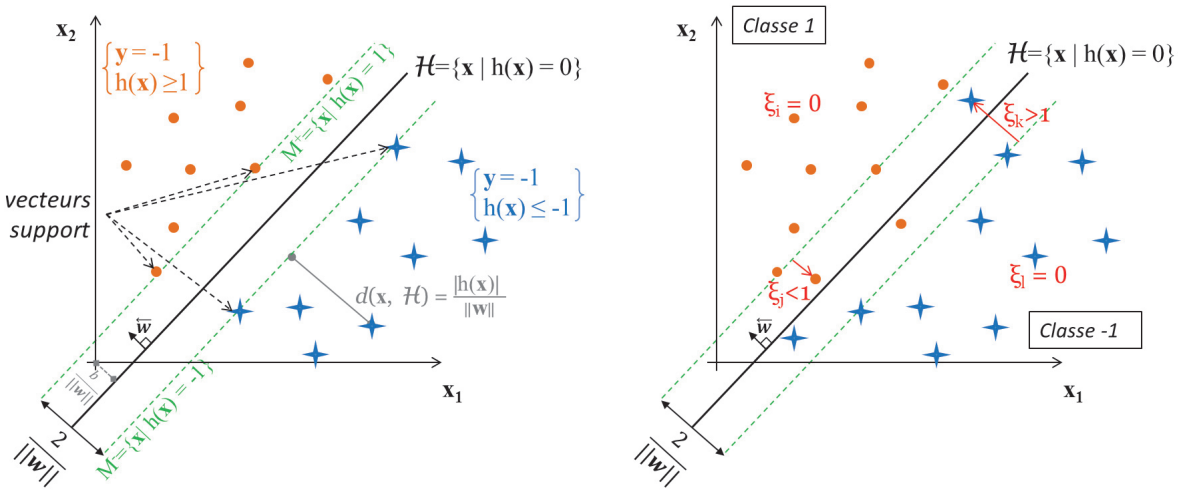
$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \forall i \in [1 \dots n].$$

En général, il n'est pas possible de trouver une séparatrice linéaire. On introduit alors le concept de "marge souple" qui permet de relâcher la contrainte de localisation par rapport à la marge en tolérant les mauvais classements. Ceci se traduit par le problème d'optimisation suivant :

$$\begin{cases} \min\limits_{\mathbf{w},\xi} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n} \xi_i & \\ \text{sous les contraintes} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, & i = 1, \dots, n \\ \text{et} & \xi_i \geq 0, & i = 1, \dots, n \end{cases} \tag{11.1}$$

où ξ_i est appelée variable "ressort". Comme l'illustre la figure 11.2b, elle représente la distance à la marge de l'exemple $\mathbf{x}_i$. Le paramètre C représente le coût de mauvaise classification, qui permet de contrôler le compromis entre le nombre d'erreurs et la largeur de la marge (valeur à fixer selon les données).

Une fois les paramètres optimaux ($\mathbf{w}$, b) calculés, on utilise la fonction de décision $\text{sign}(h(\mathbf{x}))$ pour prédire la classe ± 1 de l'entrée $\mathbf{x}$.



(a) Exemples linéairement séparables **(b) Exemples non séparables**

Figure *11.2* – *Schéma de principe des SVM linéaires dans les cas (a) séparable et (b) non séparable (introduction de variables "ressort" ξ).*

11.4 Formulation duale des SVM linéaires

D'après le théorème de Kuhn-Tucker, un problème d'optimisation possède une forme duale. Dans le cas où la fonction objectif (ici : $(\mathbf{w}, \xi) \mapsto \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$) et les contraintes (ici : $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$ et $\xi_i \geq 0$, $i = 1, \ldots n$) sont strictement convexes, résoudre l'expression duale revient à trouver la solution du problème original.
Notons que ces critères de convexité sont réalisés dans le problème (15).
De manière générale, la forme duale peut ou non être plus simple que le problème original (primal). Dans le cas des SVM, les contraintes sont plus simples mais la raison principale justifiant l'utilisation de la forme duale est de mettre le problème sous une forme permettant l'utilisation du *kernel Trick*, que nous expliciterons section 11.5.

Notons $\mathcal{L}$ le Lagrangien associé au problème. $\mathcal{L}$ est défini comme étant la somme de la fonction objectif et d'une combinaison linéaire des contraintes dont les coefficients sont appelés les multiplicateurs de Lagrange (ou encore les variables duales) :

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - (1 - \xi_i)) - \sum_{i=1}^{n}\beta_i\xi_i \qquad (11.2)$$

avec α_i, β_i (pour $i = 1...n$) les multiplicateurs de Lagrange.
Dans les conditions satisfaites par notre problème de minimisation sous contraintes (15), le théorème de Kuhn-Tucker établit qu'il y a équivalence entre trouver $(\mathbf{w}, \xi)$ satisfaisant le problème d'optimisation (15) et l'existence de vecteurs $\alpha = [\alpha_1, \ldots, \alpha_n]$ et $\beta = [\beta_1, \ldots, \beta_n]$ tels que :

– $(\mathbf{w}, b, \xi, \alpha, \beta)$ est un point selle du Lagrangien,

– $\alpha, \beta \geq 0$,
– pour $i = 1...n$, $\alpha_i(y_i(\mathbf{w}^T x_i + b) - (1 - \xi_i)) = 0$ et $\beta_i \xi_i = 0$.

On cherche donc à résoudre

$$\begin{cases} \max\limits_{\alpha,\beta} \min\limits_{\mathbf{w},b,\xi} \mathcal{L}(\mathbf{w}, b, \xi, \alpha, \beta), \\ \\ \text{avec} \quad \alpha \geq 0 \text{ et } \beta \geq 0. \end{cases}$$

Au point selle, les dérivées du Lagrangien par rapport à chaque variable primaire $\mathbf{w}, b$ et ξ doivent s'annuler. On a :

- $\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \xi, \alpha, \beta) = \mathbf{w} - \sum\limits_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$
- $\frac{\partial \mathcal{L}}{\partial b}(\mathbf{w}, b, \xi, \alpha, \beta) = -\sum\limits_{i=1}^{n} \alpha_i y_i$
- $\nabla_{\xi} \mathcal{L}(\mathbf{w}, b, \xi, \alpha, \beta) = Ce - \alpha - \beta$, avec $e = [\underbrace{1 \ldots 1}_{\text{n fois}}]^\top$, le vecteur unité.

Ainsi les conditions d'optimalité sont obtenues pour :

$$\begin{cases} \mathbf{w} &= \sum\limits_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \sum\limits_{i=1}^{n} \alpha_i y_i &= 0 \\ Ce &= \alpha + \beta \end{cases} \tag{11.3}$$

Sous ces conditions d'optimalité, on peut simplifier l'expression 11.2 du Lagrangien de la façon suivante :

$$\begin{aligned} &\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \beta) = \\ &= \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i y_i \mathbf{w}^\top x_i - \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{=0} b + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i \xi_i - \sum_{i=1}^{n} \beta_i \xi_i \\ &= \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j - \sum_{i=1}^{n} \alpha_i y_i \sum_{j=1}^{n} \alpha_j y_j x_j^\top x_i + \sum_{i=1}^{n} \alpha_i + \underbrace{C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \xi_i - \sum_{i=1}^{n} \beta_i \xi_i}_{=0} \\ &= -\tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j + \sum_{i=1}^{n} \alpha_i \end{aligned}$$

On élimine ainsi les variables primaires. Sachant que $\beta \geq 0$, la condition :

$$Ce = \alpha + \beta$$

exprimée dans 11.3, devient simplement :

$$0 \leq \alpha_i \leq C,\ i = 1 \dots n$$

.

On obtient finalement la forme duale du problème d'optimisation, dans laquelle on cherche les multiplicateurs de Lagrange tels que :

$$\begin{cases} \max\limits_{\alpha} & -\frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum\limits_{i=1}^{n}\alpha_i \\ \text{avec} & 0 \leq \alpha_i \leq C \qquad\qquad i = 1, \dots, n \\ \text{et} & \sum\limits_{i=1}^{n}\alpha_i y_i = 0 \end{cases} \tag{11.4}$$

L'hyperplan solution peut alors s'écrire :

$$h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i^\top \mathbf{x} + b \tag{11.5}$$

où les α_i sont solutions de 11.4 et b peut être obtenu en substituant n'importe quel exemple support $(\mathbf{x}_s,\, y_s)$ dans l'équation $b = y_s - \mathbf{w}^\top \mathbf{x}_s$.

Finalement, en écrivant $\mathbf{w}^\top \mathbf{w} = \alpha^\top G \alpha$ avec la matrice G de terme général : $G_{ij} = y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$ (matrice des influences), on reformule le problème dual sous la forme d'un problème matriciel :

$$\begin{cases} \min\limits_{\alpha} & \frac{1}{2}\alpha^\top G \alpha - e^\top \alpha \\ \text{avec} & 0 \leq \alpha_i \leq C, \qquad i = 1 \dots n \\ \text{et} & y^\top \alpha = 0. \end{cases} \tag{11.6}$$

La dernière condition du théorème de Kuhn-Tucker montre que seuls les points $\mathbf{x}_s$ qui sont sur les hyperplans frontière $M^{\pm}$ tels que $\mathbf{w}^\top \mathbf{x}_s + b = \pm 1$ ou à l'intérieur de la marge $-1 < \mathbf{w}^\top \mathbf{x}_s + b < 1$ jouent un rôle dans la résolution du système. Ces points pour lesquels les multiplicateurs de Lagrange α_s sont non nuls sont appelés vecteurs de support. Ce sont eux qui déterminent l'hyperplan optimal, tandis que les autres exemples ne jouent pas de rôle dans cette analyse (méthode parcimonieuse).

11.5 Formulation généralisée : les fonctions noyaux

L'analyse précédente suppose que les classes d'exemples sont linéairement séparables dans l'espace d'entrée (ou presque). Ce cas est rare dans la pratique. Il s'agit donc d'identifier une fonction de projection ϕ qui permette de passer de l'espace d'entrée $\mathcal{X}$ dans un espace image $\mathcal{H}$ (appelé espace de redescription) de grande dimension où une séparatrice linéaire (de la forme $h(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b = 0$) peut être trouvée (voir illustration

11.1). En utilisant le même cheminement que dans le cas sans transformation, le problème d'optimisation se transcrit alors :

$$\begin{cases} \max\limits_{\alpha} & -\frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle\phi(\mathbf{x}_i),\phi(\mathbf{x})\rangle_{\mathcal{H}} + \sum\limits_{i=1}^{n}\alpha_i \\ \text{avec} & 0 \leq \alpha_i \leq C \qquad\qquad i = 1,\ldots,n \\ \text{et} & \sum\limits_{i=1}^{n}\alpha_i y_i = 0. \end{cases} \tag{11.7}$$

Et l'équation de l'hyperplan séparateur dans l'espace des projections devient :

$$h(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + b = \sum_{i=1}^{n}\alpha_i y_i \langle\phi(\mathbf{x}_i),\phi(\mathbf{x})\rangle_{\mathcal{H}} + b. \tag{11.8}$$

Nous pouvons remarquer que l'hyperplan solution ne requiert que le calcul des produits scalaires $\langle\phi(\mathbf{x}_i),\phi(\mathbf{x})\rangle_{\mathcal{H}}$ entre des vecteurs de l'espace de redescription. L'espace de redescription étant a priori de dimension élevée, ces produits scalaires peuvent se révéler coûteux en termes de calculs. Néanmoins, le fait que seuls des produits scalaires doivent être calculés est un avantage. Cela va permettre d'utiliser l'astuce des noyaux (*Kernel Trick*). En effet, plutôt que de déterminer et d'utiliser de manière explicite une transformation non-linéaire $\phi : \mathcal{X} \mapsto \mathcal{H}$, on utilise une fonction $k : \mathcal{X}\times\mathcal{X} \mapsto \mathbb{R}$, appelé fonction noyau. Une fonction noyau est une forme bilinéaire symétrique positive qui se comporte comme un produit scalaire dans l'espace des projections $\mathcal{H}$ (théorème de Mercer). Elle traduit la répartition des exemples dans cet espace. Le problème d'optimisation associé devient :

$$\begin{cases} \max\limits_{\alpha} & -\frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) + \sum\limits_{i=1}^{n}\alpha_i \\ \text{avec} & 0 \leq \alpha_i \leq C \qquad\qquad i = 1,\ldots,n \\ \text{et} & \sum\limits_{i=1}^{n}\alpha_i y_i = 0 \end{cases} \tag{11.9}$$

ou, sous forme matricielle :

$$\begin{cases} \min\limits_{\alpha} & \frac{1}{2}\alpha^\top G\alpha - e^\top\alpha \\ \text{avec} & 0 \leq \alpha_i \leq C, \qquad i = 1\ldots n \\ \text{et} & y^\top\alpha = 0, \qquad i = 1\ldots n \end{cases} \tag{11.10}$$

avec $G_{ij} = y_i y_j k(\mathbf{x}_i,\mathbf{x}_j)$.

L'expression de l'hyperplan séparateur h devient, en fonction de la fonction noyau :

$$h(\mathbf{x}) = \sum_{i=1}^{n}\alpha_i y_i k(\mathbf{x}_i,\mathbf{x}) + b. \tag{11.11}$$

L'intérêt de la fonction noyau est double. D'une part le calcul se fait dans l'espace

d'origine $\mathcal{X}$, ce qui est beaucoup moins coûteux qu'un produit scalaire en grande dimension. D'autre part, la transformation ϕ n'a pas besoin d'être connue explicitement, seule la fonction noyau intervient dans les calculs. On peut donc envisager des transformations complexes, et même des espaces de redescription de dimension infinie.

En pratique, on ne connait pas la transformation ϕ et on construit directement une fonction noyau. Les noyaux usuels employés avec les SVM sont par exemple :

- le noyau linéaire : $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$ (noyau trivial, qui revient au cas du classifieur linéaire décrit section 11.3),
- le noyau polynomial : $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + 1)^d$,
- le noyau gaussien : $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2})$.

11.6 Estimation de la probabilité *a posteriori*

Nous avons montré dans les sections précédentes, comment, à partir d'un ensemble de points d'apprentissage étiquetés $(\mathbf{x}_i, y_i)_{i=1...n}$, nous pouvons construire une séparatrice optimale (au sens de Vapnik) dans $\mathcal{X}$, permettant de discriminer les points suivant leur classe d'appartenance ($\mathcal{Y} = \{-1, +1\}$).
La classe d'un nouvel exemple test observé dans $\mathcal{X}$ peut alors être prédite suivant sa position par rapport à la séparatrice, c'est à dire en utilisant le signe de la fonction de prédiction h :

$$class(\mathbf{x}) = sign(h(\mathbf{x})).$$

Cependant, plutôt qu'une décision binaire brute, il peut être intéressant d'estimer une prédiction probabiliste de la classe d'un point test, permettant d'avoir une information sur l'incertitude de cette prédiction.
Soit $\mathbf{x}$ un point observé dans $\mathcal{X}$. On cherche à estimer sa probabilité d'appartenance à la classe "+1".

Un classifieur probabiliste est une fonction $h_{prob} : \mathcal{X} \longrightarrow [0, 1]$ qui retourne une estimation de la probabilité conditionnelle de la classe "+1", ie.

$$h_{prob}(\mathbf{x}) \approx P(Y = 1|\mathbf{x})$$

Une approche standard de la classification probabiliste consiste à calibrer un classifieur numérique. Considérons la fonction h associée à notre classifieur SVM, on cherche une fonction de calibration $\varphi : \mathbb{R} \longrightarrow [0, 1]$ telle que :

$$\varphi(h(\mathbf{x})) \approx P(Y = 1|\mathbf{x}).$$

L'estimation de cette probabilité d'appartenance à partir du score $h(\mathbf{x})$ est une tâche complexe à réaliser et plusieurs méthodes ont été proposées pour répondre à ce pro-

blème [92,114,134,142]. Néanmoins, la majorité des publications s'intéressant à l'estimation de ces probabilités *a posteriori* utilise l'algorithme de Platt [92] [1]. Les études comparatives réalisées par Niculescu-Mizil & Elkan [76] et Rüping [104] montrent d'ailleurs que l'approche proposée par Platt est l'une des plus efficaces. Elle consiste à transformer le score brut $h(\mathbf{x})$ (distance à l'hyperplan séparateur) en utilisant une fonction logistique de la forme :

$$\mathbb{P}(y = 1 \mid \mathbf{x}) = \varphi(h(\mathbf{x})) = \frac{1}{1 + \exp(-A.h(\mathbf{x}) + B)} \tag{11.12}$$

où les paramètres A et B ($\in \mathbb{R}$) sont appris en maximisant la log-vraisemblance sur l'ensemble d'apprentissage (descente de gradient). On reconnaît dans cette forme une version unidimensionnelle de la régression logistique. La méthode de Platt consiste donc à effectuer une régression logistique sur le score retourné par la fonction de décision SVM.

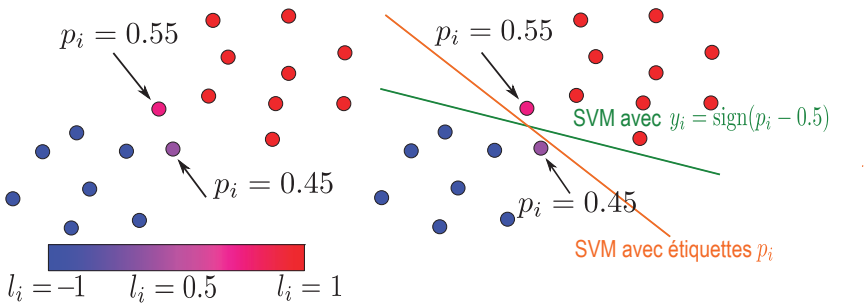
La constante B a été introduite par Platt pour permettre un biais pour le seuil de Bayes optimal de telle manière qu'une probabilité de sortie 0.5 puisse correspondre à une valeur de score $h(\mathbf{x})$ non nulle du SVM. En pratique, certaines études [83, 105] utilisent une version simplifiée de cette formule, avec $A = 1$ et $B = 0$.

1. corrigé par Lin [62] en 2007

Chapitre 12

Extension des SVM classiques au cas des étiquettes incertaines

12.1 Introduction



(a) Exemples d'entraînement **(b) Frontières de discrimination**

Figure *12.1* – *Enjeux de l'étiquetage probabiliste.* ***(a)*** *Soit deux nuages de points avec une étiquette "±1" pour les exemples "certains" et avec pour étiquette une probabilité d'appartenance à la classe "+1" pour les autres.* ***(b)*** *On représente la séparatrice obtenue avec un SVM classique (en vert) et celle que l'on souhaite obtenir avec un P-SVM (en orange).*

La figure 12.1 illustre les limites d'utilisation des SVM classiques lorsque les étiquettes de certains points d'apprentissage sont des probabilités d'appartenance à une classe. Elle illustre l'enjeu d'un algorithme capable de prendre en compte, dans l'apprentissage, à la

fois des étiquettes binaires et des probabilités. Dans cette section, nous proposons un nouvel algorithme basé sur le principe des SVM, mais capable d'apprendre sur un étiquetage mixte (binaire et probabiliste) en effectuant simultanément une classification et une régression. Pour faciliter les discussions, ce dernier algorithme sera noté P-SVM dans la suite de cette thèse.

L'illustration 12.1a représente un ensemble d'exemples d'entraînement répartis selon deux nuages de points bleus et rouges étiquetés "-1" et "+1" pour les exemples "certains" et de la probabilité p d'appartenance à la classe "1" pour les autres. Les variations de ton sont liées à celles de la probabilité p. On s'intéresse particulièrement aux deux exemples représentés en rose ($p = 0.55$) et mauve ($p = 0.45$) pour lesquels la probabilité d'appartenance à l'une ou l'autre des deux classes est proche du hasard ($p \approx 0.5$).
L'illustration 12.1b propose une comparaison des frontières de séparation que l'on obtiendrait avec les approches SVM (en vert) et P-SVM (en orange). La séparatrice construite avec les SVM utilise uniquement les étiquettes binaires. Tous les points sont "binarisés" (i.e. en prenant un seuil arbitraire de 0.5 : le point rose est classifié "1" et le point mauve est classifié "-1") et traités de manière identique, indépendamment de la valeur p. La frontière séparatrice construite pour maximiser la marge va faire des points mauve et rose des points "supports", biaisant ainsi la position de la marge.
La séparatrice que l'on souhaite construire avec un algorithme de type P-SVM prend, elle, en compte la probabilité d'appartenance des points "incertains". L'objectif est différent. L'algorithme P-SVM effectue une classification sur les point "sûrs" et une régression sur les points de classe "incertaine" (i.e. les prédictions de probabilités réalisées par les P-SVM sur les points d'apprentissage incertains sont contraints de rester proches de la probabilité *a priori* p_i apprise). En prenant en compte le degré de certitude sur la classe, on souhaite arriver à construire la frontière orange, séparant de manière plus naturelle les deux nuages de points.

L'objectif des P-SVM est triple. Il s'agit de :

bien classifier les exemples pour lesquels l'étiquetage est "sûr" ($y_i = \pm 1$) ;

bien estimer les probabilités d'appartenance des exemples pour lesquels l'étiquetage est une probabilité *a priori* ($p_i \in [0, 1]$) ;

bien généraliser (en classification et estimation de probabilité) à de nouveaux exemples n'appartenant pas à la base d'apprentissage ayant servi à construire la frontière de décision.

12.2 Formulation du problème

Dans tout ce chapitre, on considère un ensemble de données d'apprentissage composé de m couples d'observation/étiquette $(\mathbf{x}_i, l_i)_{i=1\ldots m}$ tels que $\mathbf{x}_i \in \mathcal{X}(= \mathbb{R}^d)$ et :

- $l_i = y_i \in \{-1, 1\}$ pour $i = 1 \ldots n$, et
- $l_i = p_i \in [0, 1]$ pour $i = n + 1 \ldots m$

L'étiquette p_i, associée au point $\mathbf{x}_i$, représente sa probabilité d'appartenance à la classe "+1" :

$$p_i = P(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i).$$

Elle permet d'exprimer le degré de certitude quant à la classe de la variable $\mathbf{x}_i$.

Notre objectif est de construire une fonction de prédiction $h : \mathcal{X} \longrightarrow \mathcal{Y}$ qui permette à la fois de bonnes performances en classification (ainsi qu'en estimation de probabilités) en étant apprise, à profit, sur un ensemble de données d'étiquettes mixtes ($\mathcal{Y} = \{-1, 1\} \cup [0, 1]$).
L'idée est donc de construire cette fonction h par apprentissage sur l'ensemble étiqueté $(\mathbf{x}_i, l_i)_{i=1\ldots m}$ de façon à minimiser l'erreur de classification sur les n points $\mathbf{x}_i$ d'étiquettes $y_i \in \{-1, 1\}$ et à minimiser l'erreur d'estimation de probabilité sur les $m - n$ points $\mathbf{x}_i$ d'étiquettes $p_i \in [0, 1]$.

Pour construire la fonction h à partir des étiquettes probabilistes, l'approche que nous proposons consiste à effectuer le schéma inverse de l'estimation des probabilités *a posteriori* détaillée section 11.6. Dans cette section, nous avions en effet explicité comment, à partir d'une fonction de classification h construite par l'approche SVM "classique", il était possible d'estimer les probabilités *a posteriori* en introduisant une fonction de calibration φ dont les paramètres étaient ajustés sur les données de façon à ce que $\varphi(h(\mathbf{x}))$ soit un bon estimateur de $P(Y = 1|\mathbf{x})$.
Dans notre nouvelle problématique, on cherche au contraire à construire h à partir des p_i (et toujours des y_i ...).

12.3 Le problème des P-SVM linéaires (dans le primal)

Dans cette première section, on se ramène au cas général où les données sont linéairement séparables. Ce cas a été présenté, dans le cas du SVM classique, à la section 11.3. Soit $\mathcal{H}$ l'hyperplan séparateur que l'on cherche à construire. On peut écrire $\mathcal{H}$ de la manière suivante : $\mathcal{H} = \{\mathbf{x} \in \mathcal{X} \mid \mathbf{w}^\top \mathbf{x} + b = 0\}$ avec $\mathbf{w} \in \mathbb{R}^d$ un vecteur orthogonal à l'hyperplan séparateur $\mathcal{H}$, et $b \in \mathbb{R}$ le biais.
Une fois l'hyperplan optimal défini, une observation $\mathbf{x}$ de $\mathcal{X}$ pourra alors être classifiée comme appartenant à l'une ou l'autre des deux classes en fonction de sa position par rapport à l'hyperplan, donnée par $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$.

Suivant la méthode introduite section 11.6, on cherche à estimer les probabilités des exemples $(\mathbf{x}_i)_{i=n+1\ldots m}$ sous la forme :

$$p_i = \varphi(h(\mathbf{x}_i)) \tag{12.1}$$

où nous choisissons φ comme étant la fonction logistique de la forme :

$$\begin{aligned} \varphi : \mathbb{R} &\to [0,1] \\ u &\mapsto \frac{1}{1+\exp(-A.u)} \end{aligned} \tag{12.2}$$

avec $A \in \mathbb{R}$ est un paramètre d'échelle à déterminer.
Autrement dit, on cherche à construire la fonction de prédiction h de façon à ce que $\varphi(h(\mathbf{x}))$ soit un bon estimateur de la probabilité étiquetée p.

En pratique, nous cherchons plutôt à contraindre la prédiction de probabilité faite pour le point $\mathbf{x}_i$ à être au plus à une certaine distance η de p_i, i.e. dans un tube fin tel que :

$$\mid \varphi(h(\mathbf{x}_i)) - p_i \mid < \eta \text{ , pour } i{=}n+1\ldots m \tag{12.3}$$

où η indique l'écart toléré entre la probabilité *a priori* p_i de l'exemple, et la probabilité $\varphi(h(\mathbf{x}_i))$ prédite [1].
Pour éviter d'introduire des valeurs non définies dans la suite de notre développement (lors de l'inversion de φ), on contraint les valeurs d'étiquette p_i à appartenir à l'intervalle [η, 1-η], quitte à ré-étiqueter les données de la façon suivante :

$$l_i = \begin{cases} \text{-1 si } p_i \text{ - } \eta \leq 0 \\ \text{+1 si } p_i + \eta \geq 1 \\ p_i \text{ sinon.} \end{cases}$$

1. Le coefficient η dépend de la précision de l'étiquetage et de la confiance en l'étiquetage. La précision, notée ε, représente le pas d'échantillonnage (dans notre cas d'étude, les scores affectés par les radiologues étant $\{0,1,2,3,4\} \Leftrightarrow \{0, 0.25, 0.5, 0.75, 1\}$, on choisit $\varepsilon = 0.25$). La confiance δ représente l'écart maximum espéré entre la probabilité vraie et la valeur affectée (on choisit $\delta = \varepsilon/2 = 0.125$ dans notre cas pratique).

La condition 12.3 peut se réécrire de manière équivalente :

$$\begin{array}{rrcl} & p_i - \eta \leq & \varphi(h(\mathbf{x}_i)) & \leq p_i + \eta, \\ \iff & p_i - \eta \leq & \frac{1}{1+\exp(-A.h(\mathbf{x}_i))} & \leq p_i + \eta, \\ \iff & -\frac{1}{A}.\ln(\frac{1}{p_i-\eta}-1) \leq & h(\mathbf{x}_i) & \leq -\frac{1}{A}.\ln(\frac{1}{p_i+\eta}-1), \\ \iff & a.z_i^- \leq & h(\mathbf{x}_i) & \leq a.z_i^+, \end{array} \tag{12.4}$$

avec :

$$z_i^{\pm} = \frac{1}{a}\varphi^{-1}(p_i \pm \eta) = \ln(\frac{1}{p_i \pm \eta} - 1) \text{ et } a = -\frac{1}{A}.$$

La figure 12.2 donne le tracé des fonctions z^+ et z^- :

$$\begin{array}{c} z^{\pm} : [0,1] \to \mathbb{R}^{\pm} \\ p \mapsto z^{\pm} \end{array} \tag{12.5}$$

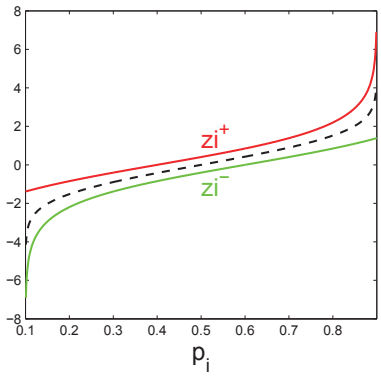


Figure *12.2* – *Rôle de z_i^+ et z_i^-. Représentation de la contrainte de localisation des images par h des points d'étiquette probabiliste en fonction de la valeur de la probabilité p_i associée. La contrainte de localisation définie par les bornes z_i^+ et z_i^- vise à maintenir la prédiction faite sur le point $\mathbf{x}_i$ dans des limites définies. Plus la valeur étiquetée p_i est proche d'une valeur extrême 0 ou 1 (à la distance minimum η), moins la contrainte de localisation sur $h(\mathbf{x}_i)$ est forte ($\to \infty$).*

Une hypothèse raisonnable d'après l'inégalité 12.3 est que les bords du tube définissant la contrainte sur la prédiction de probabilité correspondent aux cas extrêmes $h(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b = \pm 1$. On a alors : $\varphi(h(\mathbf{x}_i)) = \eta$ pour les $\mathbf{x}_i$ tels que $\mathbf{w}^\top \mathbf{x}_i + b = -1$ et $\varphi(h(\mathbf{x}_i)) = 1 - \eta$ pour les $\mathbf{x}_i$ tels que $\mathbf{w}^\top \mathbf{x}_i + b = +1$, ce qui, dans les deux cas, nous donne :

$$A = \ln(\frac{1}{\eta} - 1). \tag{12.6}$$

Cette hypothèse permet de fixer le paramètre A mais peut être jugée arbitraire. Dans le problème général qui suit, on considère donc A (ou de manière équivalente a) comme une variable d'échelle à apprendre. On aura ainsi le choix de fixer *a priori* (avec (12.6)) ou d'apprendre ce paramètre.

Finalement, on cherche à estimer notre fonction de prédiction h à partir des exemples $(\mathbf{x}_i, l_i)_{i=1...m}$ de façon à :

- ✓ **bien prédire la classe** (lire section 11.3) ± 1 des exemples $(\mathbf{x}_i)_{i=1...n}$ via la fonction de décision $\mathbf{x} \mapsto h(\mathbf{x})$,
  Condition : $y_i(h(\mathbf{x}_i)) \geq 1$

- ✓ **bien estimer les probabilités** p_i des exemples $(\mathbf{x}_i)_{i=n+1...m}$, via la fonction d'estimation $\mathbf{x} \mapsto \varphi(h(\mathbf{x}))$,
  Condition : $\varphi^{-1}(p_i - \eta) \leq h(\mathbf{x}_i) \leq \varphi^{-1}(p_i + \eta)$

- ✓ **bien généraliser** en maximisant la marge entre les exemples de classes différentes.
  Condition : $\max_{\mathbf{w}} \frac{1}{\|\mathbf{w}\|}$

Le problème d'optimisation peut finalement s'écrire de la façon suivante :

$$\begin{cases} \min\limits_{\mathbf{w},b,a} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{avec} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \qquad i = 1...n \\ & a.z_i^- \leq \mathbf{w}^\top \mathbf{x}_i + b \leq a.z_i^+, \quad i = n+1...m. \end{cases} \tag{12.7}$$

Évidemment, si $n = m$, on est ramené au problème SVM classique décrit dans la section 11.3.

On introduit, comme en 11.3, le concept de "marge souple", qui tolère les mauvais classements. Les variables "ressort" (*slack variables*) ξ_i représentent le coût de mauvaise classification (à minimiser), pondéré par la constante C. De même, on introduit des variables "ressort" ν_i^- et ν_i^+, qui sont le pendant des variables ξ_i pour les points à étiquette probabiliste, et $\tilde{C}$ le coût associé. Elles permettent de relâcher les contraintes sur la prédiction de probabilité, définies en (12.3).

$$\begin{cases} \min\limits_{\mathbf{w},b,a,\xi,\nu^-,\nu^+} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n}\xi_i + \tilde{C}\sum\limits_{i=n+1}^{m}(\nu_i^- + \nu_i^+) & \\ \text{avec} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, & i = 1...n \\ & a.z_i^- - \nu_i^- \leq \mathbf{w}^\top \mathbf{x}_i + b \leq a.z_i^+ + \nu_i^+, & i = n+1...m \\ & 0 \leq \xi_i, & i = 1...n \\ & 0 \leq \nu_i^-, & i = n+1...m \\ & 0 \leq \nu_i^+, & i = n+1...m \end{cases} \tag{12.8}$$

12.4 Formulation duale des P-SVM linéaires

Suivant la méthodologie introduite section 11.4, on ré-exprime le problème sous sa forme duale. On cherche un point selle du lagrangien :

$$\max_{\alpha,\beta,\mu^+,\mu^-,\gamma^+,\gamma^-} \min_{\mathbf{w},b,a,\xi,\nu^+,\nu^-} \mathcal{L}(\mathbf{w},b,a,\xi,\alpha,\beta,\nu^-,\nu^+,\mu^-,\mu^+,\gamma^-,\gamma^+)$$

où $\alpha,\beta,\mu^+,\mu^-,\gamma^+,\gamma^- \geq 0$ sont les multiplicateurs de Lagrange et $\mathcal{L}$ le lagrangien défini ci-dessous :

$$\begin{aligned}\mathcal{L}(\mathbf{w},b,a,\xi,\alpha,\beta,\nu^-,\nu^+,\mu^-,\mu^+,\gamma^-,\gamma^+) = \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \tilde{C}\sum_{i=n+1}^{m}(\nu_i^- + \nu_i^+) \\ & - \sum_{i=1}^{n}\alpha_i(y_i(\mathbf{w}^\top\mathbf{x}_i + b) - (1-\xi_i)) - \sum_{i=1}^{n}\beta_i\xi_i \\ & - \sum_{i=n+1}^{m}\mu_i^-((\mathbf{w}^\top\mathbf{x}_i + b) - (az_i^- - \nu_i^-)) - \sum_{i=n+1}^{m}\gamma_i^-\nu_i^- \\ & - \sum_{i=n+1}^{m}\mu_i^+((az_i^+ + \nu_i^+) - (\mathbf{w}^\top\mathbf{x}_i + b)) - \sum_{i=n+1}^{m}\gamma_i^+\nu_i^+\end{aligned}$$

On calcule les dérivées relativement à chacune des variables primales $\mathbf{w}, b, \xi, \nu^-, \nu^+$ et on cherche les valeurs en lesquelles ces dérivées s'annulent afin d'obtenir un point selle du Lagrangien $\mathcal{L}$ (cf. Théorème de Kuhn-Tucker 11.4) :

- $\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i - \sum_{i=n+1}^{m}\mu_i^-\mathbf{x}_i - \sum_{i=n+1}^{m}(-\mu_i^+\mathbf{x}_i)$
  $= \mathbf{w} + \sum_{i=n+1}^{m}(\mu_i^+ - \mu_i^-)\mathbf{x}_i - \sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i$
- $\frac{\partial\mathcal{L}}{\partial b} = -\sum_{i=1}^{n}\alpha_i y_i - \sum_{i=n+1}^{m}\mu_i^- - \sum_{i=n+1}^{m} -\mu_i^+ = \sum_{i=n+1}^{m}(\mu_i^+ - \mu_i^-) - \sum_{i=1}^{n}\alpha_i y_i$
- $\frac{\partial\mathcal{L}}{\partial a} = \sum_{i=n+1}^{m}\mu_i^- z_i^- - \sum_{i=n+1}^{m}\mu_i^+ z_i^+$
- $\nabla_\xi\mathcal{L} = Ce_1 - \beta - \alpha$
- $\nabla_{\nu^-}\mathcal{L} = \tilde{C}e_2 - \mu^- - \gamma^-$
- $\nabla_{\nu^+}\mathcal{L} = \tilde{C}e_2 - \mu^+ - \gamma^+$

avec e_1 et e_2 définis par : $e_1 = [\underbrace{1\ldots1}_{\text{n fois}}\ \underbrace{0\ldots0}_{\text{(m-n) fois}}\ ]^\top$ et $e_2 = [\underbrace{0\ldots0}_{\text{n fois}}\ \underbrace{1\ldots1}_{\text{(m-n) fois}}\ ]^\top$.

Les conditions d'optimalité sont obtenues pour :

$$\left\{\begin{array}{lcl} \mathbf{w} & = & \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i - \sum_{i=n+1}^{m}(\mu_i^+ - \mu_i^-)\mathbf{x}_i \\ \sum_{i=1}^{n}\alpha_i y_i & = & \sum_{i=n+1}^{m}(\mu_i^+ - \mu_i^-) \\ \sum_{i=n+1}^{m}\mu_i^- z_i^- & = & \sum_{i=n+1}^{m}\mu_i^+ z_i^+ \\ Ce_1 & = & \alpha + \beta \\ \tilde{C}e_2 & = & \mu^- + \gamma^- = \mu^+ + \gamma^+. \end{array}\right. \tag{12.9}$$

Ainsi, l'expression du lagrangien au point selle peut se simplifier en injectant les conditions d'annulation précédentes :
$\mathcal{L}(\mathbf{w}, b, a, \xi, \alpha, \beta, \nu^-, \nu^+, \mu^-, \mu^+, \gamma^-, \gamma^+)$

$$\begin{array}{rl} = & \frac{1}{2}\mathbf{w}^\top\mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{w}^\top\mathbf{x}_i - \sum_{i=n+1}^{m}(\mu_i^- - \mu_i^+)\mathbf{w}^\top\mathbf{x}_i + \sum_{i=1}^{n}\alpha_i + \sum_{i=n+1}^{m}\mu_i^- z_i^- - \sum_{i=n+1}^{m}\mu_i^+ z_i^+ \\ & \underbrace{+C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\beta_i\xi_i}_{=0} + \underbrace{\sum_{i=n+1}^{m}\mu_i^+ b - \sum_{i=n+1}^{m}\mu_i^- b - \sum_{i=1}^{n}\alpha_i y_i b}_{=0} \\ & \underbrace{+\tilde{C}\sum_{i=n+1}^{m}\nu_i^- - \sum_{i=n+1}^{m}\mu_i^-\nu_i^- - \sum_{i=n+1}^{m}\gamma_i^-\nu_i^-}_{=0} + \underbrace{\tilde{C}\sum_{i=n+1}^{m}\nu_i^+ - \sum_{i=n+1}^{m}\mu_i^+\nu_i^+ - \sum_{i=n+1}^{m}\gamma_i^+\nu_i^+}_{=0} \\ = & -\frac{1}{2}\mathbf{w}^\top\mathbf{w} + \sum_{i=1}^{n}\alpha_i + \sum_{i=n+1}^{m}\mu_i^- z_i^- - \sum_{i=n+1}^{m}\mu_i^+ z_i^+. \end{array}$$

De plus, sachant que $\beta \geq 0$, $\gamma^+ \geq 0$, $\gamma^- \geq 0$, les conditions 12.9 :

$$\left\{\begin{array}{lcl} Ce_1 & = & \alpha + \beta, \\ \tilde{C}e_2 & = & \mu^- + \gamma^- = \mu^+ + \gamma^+, \end{array}\right.$$

deviennent simplement :

$$\left\{\begin{array}{lll} 0 & \leq \alpha_i \leq C, & i = 1...n \\ 0 & \leq \mu_i^+ \leq \tilde{C}, & i = n+1...m \\ 0 & \leq \mu_i^- \leq \tilde{C}, & i = n+1...m. \end{array}\right.$$

On obtient finalement la forme duale du problème d'optimisation :

$$\begin{cases} \max\limits_{\alpha} & -\frac{1}{2}\mathbf{w}^\top\mathbf{w} + \sum\limits_{i=1}^{n}\alpha_i + \sum\limits_{i=n+1}^{m}\mu_i^- z_i^- - \sum\limits_{i=n+1}^{m}\mu_i^+ z_i^+ & \\ \text{avec} & 0 \leq \alpha_i \leq C, & i = 1...n \\ & 0 \leq \mu_i^+ \leq \tilde{C}, & i = n+1...m \\ & 0 \leq \mu_i^- \leq \tilde{C}, & i = n+1...m \\ \text{et} & y^\top\alpha = \sum\limits_{i=n+1}^{m}(\mu_i^+ - \mu_i^-) & \end{cases} \tag{12.10}$$

L'expression de $\mathbf{w}^\top\mathbf{w}$ peut se ré-écrire :

$$\begin{aligned} \mathbf{w}^\top\mathbf{w} &= (\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i - \sum_{i=n+1}^{m}(\mu_i^+ - \mu_i^-)\mathbf{x}_i)^\top(\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i - \sum_{i=n+1}^{m}(\mu_i^+ - \mu_i^-)\mathbf{x}_i) \\ &= \sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top\mathbf{x}_j + \sum_{i,j=n+1}^{m}(\mu_i^+ - \mu_i^-)(\mu_j^+ - \mu_j^-)\mathbf{x}_i^\top\mathbf{x}_j \\ &\quad -\sum_{i=1}^{n}\sum_{j=n+1}^{m}(\mu_j^+ - \mu_j^-)\alpha_i y_i \mathbf{x}_i^\top\mathbf{x}_j - \sum_{i=n+1}^{m}\sum_{j=1}^{n}(\mu_i^+ - \mu_i^-)\alpha_j y_j \mathbf{x}_i^\top\mathbf{x}_j \\ &= \sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top\mathbf{x}_j - \sum_{i,j=n+1}^{m}(\mu_i^+\mu_j^+ + \mu_i^-\mu_j^- - \mu_i^-\mu_j^+ - \mu_i^+\mu_j^-)\mathbf{x}_i^\top\mathbf{x}_j \\ &\quad -\sum_{i=1}^{n}\sum_{j=n+1}^{m}(\alpha_i\mu_j^+ y_i \mathbf{x}_i^\top\mathbf{x}_j - \alpha_i\mu_j^- y_i \mathbf{x}_i^\top\mathbf{x}_j) - \sum_{i=n+1}^{m}\sum_{j=1}^{n}(\alpha_j\mu_i^+ y_j \mathbf{x}_i^\top\mathbf{x}_j - \alpha_j\mu_i^- y_j \mathbf{x}_i^\top\mathbf{x}_j). \end{aligned}$$

Soit $\Gamma = [\alpha_1 \dots \alpha_n \ \mu_{n+1}^+ \dots \mu_m^+ \ \mu_{n+1}^- \dots \mu_m^-]^\top$ un vecteur de dimension $2m - n$. On a :

$$\mathbf{w}^\top\mathbf{w} = \Gamma^\top \mathrm{G}\Gamma$$

avec :

$$\mathrm{G} = \begin{pmatrix} K_1 & - & K_2 & K_2 \\ - & K_2^\top & K_3 & - & K_3 \\ & K_2^\top & - & K_3 & K_3 \end{pmatrix}$$

et :

$$\begin{cases} K_1 = (y_i y_j \mathbf{x}_i^\top\mathbf{x}_j)_{i,j=1...n}, \\ K_2 = (\mathbf{x}_i^\top\mathbf{x}_j y_i)_{i=1...n, j=n+1...m}, \\ K_3 = (\mathbf{x}_i^\top\mathbf{x}_j)_{i,j=n+1...m}, \end{cases} \tag{12.11}$$

On a ainsi éliminé les variables primales et on peut finalement reformuler le problème

dual sous forme matricielle :

$$\left\{\begin{array}{ll} \min\limits_{\Gamma} & \frac{1}{2}\Gamma^\top \mathrm{G}\Gamma - \tilde{e}^\top\Gamma, \\ \text{avec} & \tilde{e} = [\underbrace{1\dots1}_{nfois}\underbrace{-z_{n+1}^+\cdots-z_m^+}_{n-mfois}\underbrace{z_{n+1}^-\dots z_m^-}_{n-mfois}] \\ & f^\top\Gamma = 0 \\ & g^\top\Gamma = 0 \\ \text{avec} & f^\top = [y^\top, \underbrace{-1\cdots-1}_{n-mfois}, \underbrace{1\dots1}_{n-mfois}] \\ \text{et} & 0 \leq \Gamma \leq [\underbrace{C\dots C}_{nfois}\underbrace{\tilde{C}\dots\tilde{C}}_{n-mfois}\underbrace{\tilde{C}\dots\tilde{C}}_{n-mfois}]^\top \\ & g^\top = [\underbrace{0\dots0}_{nfois}, \underbrace{-z_{n+1}^+\cdots-z_m^+}_{n-mfois}\underbrace{z_{n+1}^-\dots z_m^-}_{n-mfois}]. \end{array}\right. \tag{12.12}$$

On obtient un système d'optimisation sous forme d'un problème d'optimisation quadratique (QP) standard. Celui-ci peut être résolu en s'appuyant sur *les solveurs* utilisés dans le cadre des SVM classiques.

12.5 Formulation généralisée (noyaux)

De la même façon que dans la section 11.5, on peut généraliser la formulation du problème présenté précédemment dans le cadre des séparatrices linéaires au cas des exemples non linéairement séparables en introduisant les fonctions noyaux.

On obtient alors la même formulation matricielle du problème qu'en 12.12, avec :

$$\begin{aligned} K_1 &= (y_i y_j k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1\dots n}, \\ K_2 &= (k(\mathbf{x}_i, \mathbf{x}_j) y_i)_{i=1\dots n, j=n+1\dots m}, \\ K_3 &= (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=n+1\dots m}, \end{aligned}$$

12.6 Conclusion

Nous avons proposé dans ce chapitre un nouveau problème d'apprentissage dans lequel nous entraînons un classifieur à partir d'un étiquetage mixte comprenant des étiquettes de classe "sûre" et des probabilités d'appartenance à une classe.

Le code que nous avons développé conjointement avec R. Flamary et S. Canu pour tester cet algorithme est en libre accès. Il peut être téléchargé sur le site dédié au partage de programmes de *machine learning* mloss.org. Notre code s'appuie sur la toolbox SVM-KM (*SVM and Kernel Methods Toolbox*), développée par S. Canu *et coll.* [15].

Les chapitres 13 et 14 suivants seront dédiés à l'évaluation, sur des exemples jouets et sur des données cliniques, de cet algorithme dont nous comparerons les performances avec celles obtenues avec une approche SVM classique.

Chapitre 13

Exemples sur données simulées

13.1 Introduction

Dans ce chapitre, nous proposons de tester les propriétés des P-SVM sur un ensemble de cas simulés, afin d'en appréhender les performances en termes de précision d'étiquetage (classification), d'estimation de probabilité et de robustesse au bruit d'étiquetage.

13.2 Mesures de performance

Afin d'évaluer et comparer les performances de prédiction réalisées par un classifieur SVM classique et notre algorithme P-SVM, nous utilisons différentes mesures définies ci-dessous.

- Pour évaluer les performances de *classification*, nous utilisons :
  - **La précision** (Acc), qui correspond au pourcentage de bonnes classifications (± 1) des données du jeu de test, définie section 4.5.1,
  - **L'aire sous la courbe ROC** (AUC), définie section 4.5.2.
- Pour évaluer la précision des *estimations de probabilités*, nous introduisons différentes métriques, avec P et Q deux distributions de probabilités :
  - **Divergence de Kullback-Leibler** (D_{KL}) (ou entropie relative),
$$D_{KL}(P||Q) = \sum_{i=1}^{n} \mathrm{P}(y_i = 1|\mathbf{x}_i)\log(\frac{P(y_i = 1|\mathbf{x}_i)}{Q(y_i = 1|\mathbf{x}_i)})$$
  - **Erreur d'alignement** (Err_{Align}),
$$Err_{Align} = 1 - \frac{\sum_{i=1}^{n} P(y_i = 1|\mathbf{x}_i)Q(y_i = 1|\mathbf{x}_i)}{\sqrt{\sum_i P(y_i = 1|\mathbf{x}_i)^2}\sqrt{\sum_i Q(y_i = 1|\mathbf{x}_i)^2}}$$

– **Distance** L_1 (en valeur absolue) **moyenne**,
$$D_{L1} = \frac{1}{n}\sum_{i=1}^{n} |P(y_i = 1|\mathbf{x}_i) - Q(y_i = 1|\mathbf{x}_i)|$$
– **Mean Cross-Entropy** (MCE),
$$MCE = -\frac{1}{n}\sum_{i=1}^{n}(P(y_i = 1|\mathbf{x}_i) * \log(Q(y_i = 1|\mathbf{x}_i)) + (1 - P(y_i = 1|\mathbf{x}_i))\log(1 - Q(y_i = 1|\mathbf{x}_i)))$$
– **Mean Squared Error** (MSE),
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P(y_i = 1|\mathbf{x}_i) - Q(y_i = 1|\mathbf{x}_i))^2$$

Toutes ces métriques expriment une mesure de dissimilarité entre deux distributions de probabilités P et Q : plus leurs valeurs augmentent, plus les distributions sont éloignées. Elles ont été choisies de par leur utilisation dans diverses publications impliquant des comparaisons d'estimées de probabilités [16, 104]. Dans notre évaluation, P est la probabilité "vraie" calculée et Q la probabilité estimée par la méthode de Platt (pour le SVM classique) ou par le P-SVM (directement).

13.3 Comportement en présence d'un *outlier*

13.3.1 Description du test

On simule deux nuages de points (2-D) étiquetés "+1" et "-1", tirés selon deux lois gaussiennes, $\mathcal{N}(\mu_{-1}, \sigma)$ et $\mathcal{N}(\mu_1, \sigma)$, de mêmes variances et de moyennes différentes, qui ne se chevauchent pas.
On introduit arbitrairement un point singulier (*outlier*) $\mathbf{x}$ situé à égale distance des centres des deux nuages de points, tel que $P(Y = 1 \mid X = \mathbf{x}) = 0.51$.
On étudie la perturbation provoquée par ce point sur la construction de la frontière de décision en fonction de la valeur de l'étiquette qui lui est affectée :

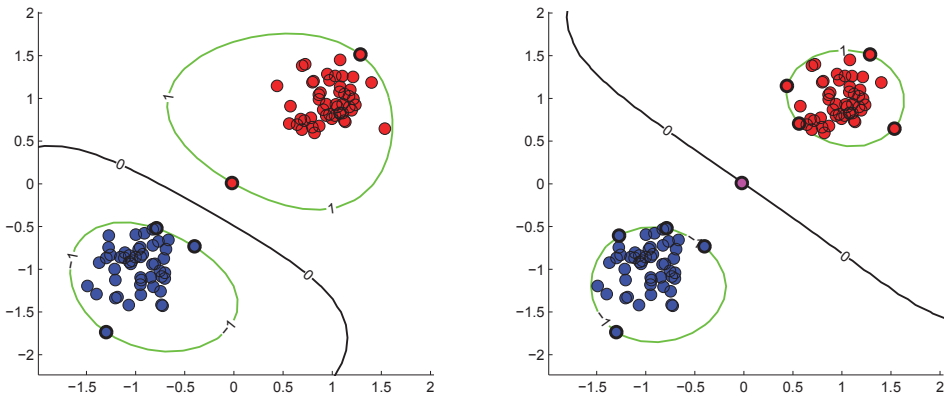
– étiquette "sûre" : $y = 1$, puisque $P(Y = 1 \mid X = \mathbf{x}) > 0.5$,
versus
– étiquette "probabiliste" : $p = 0.51$.

On "apprend" ensuite notre classifieur P-SVM sur les deux jeux de données ainsi construits. Dans le cas où la séparatrice est apprise uniquement sur des étiquettes binaires, on est ramené à un SVM classique.

13.3.2 Résultat du test

La figure 13.1 illustre ce test avec $n_{app} = 100$ points d'apprentissage (50 points par distribution) tirés aléatoirement selon les deux distributions $\mathcal{N}(\mu_{-1}, \sigma)$ et $\mathcal{N}(\mu_1, \sigma)$ telles que : μ_{-1}=(-1,-1), μ_{+1}=(1,1) et σ=0.3. On choisit η=0.1 (A=2.2 d'après l'équation (12.6)), C = 100 et $\hat{C}$= 100 ; on utilise un noyau gaussien de paramètre $\sigma = 1$.
Dans le cas où l'étiquette du point $\mathbf{x}$ vaut 1 (jeu de données étiquetées de manière binaire), la frontière est construite de façon à minimiser l'erreur de classification et maximiser la

marge : l'*outlier* étant le point de la classe "1" le plus proche des points de la classe "-1", il devient un point support et conditionne la position de la séparatrice (figure 13.1a). Celle-ci est largement déviée vers le nuage de points de la classe "-1". On perd donc en généralisation.
Au contraire, dans le cas où l'étiquette du point $\mathbf{x}$ vaut p, l'approche P-SVM tire profit de l'information probabiliste apprise. La classe de ce point étant très incertaine ($p \simeq 0.5$), il se retrouve positionné sur la frontière de décision (figure 13.1b) tandis que le reste des points d'étiquettes binaires sont séparés de manière optimale (au sens de la maximisation de la marge).



(a) L'*outlier* est étiqueté y=+1 **(b) L'*outlier* est étiqueté p=0.51**

Figure *13.1* – *Exemple jouet : présence d'un outlier. On simule deux nuages de points ($n_{app} = 100$) étiquetés "+1" (rouges) et "-1" (bleus). On observe l'évolution de la position de la séparatrice (tracée en noir) en fonction de l'étiquette affectée à l'outlier : (a) l'outlier est affecté à la classe +1, (b) l'outlier est étiqueté de sa probabilité $P(Y = 1 \mid X = \mathbf{x}) = 0.51$.*

13.4 Estimation des probabilités

13.4.1 Description du test

On simule deux ensembles de points (1-D) tirés selon deux lois gaussiennes, $\mathcal{N}(\mu_{-1}, \sigma)$ et $\mathcal{N}(\mu_1, \sigma)$, représentant deux classes de données "-1" et "+1", de moyennes différentes et de même variance. Soit $(\mathbf{x}_{app_i})_{i=1\ldots n_{app}}$ les données d'apprentissage ainsi créées.
On calcule les probabilités de chacun des points d'appartenir à la classe "+1".

Pour $i = 1 \ldots n_{app}$, on a

$$P(y_{app_i} = 1 | \mathbf{x}_{app_i}) = \frac{P(\mathbf{x}_{app_i} | y_{app_i} = 1)}{P(\mathbf{x}_{app_i} | y_{app_i} = 1) + P(\mathbf{x}_{app_i} | y_{app_i} = -1)} \tag{13.1}$$

$$\text{avec} \quad P(\mathbf{x}_{app_i} | y_{app_i} = 1) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{\mathbf{x}_{app_i} - \mu_1}{\sigma})^2)$$

$$\text{et} \quad P(\mathbf{x}_{app_i} | y_{app_i} = -1) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{\mathbf{x}_{app_i} - \mu_{-1}}{\sigma})^2).$$

On construit un premier jeu de données $(\mathbf{x}_{app_i}, y_{app_i})_{i=1\dots n_{app}}$, pour lequel les données sont étiquetées comme suit :

$$\begin{array}{llllll} \text{si} & P(y_{app_i}=1|\mathbf{x}_{app_i}) & > & 0.5, & \text{alors} & y_{app_i}=+1; \\ \text{si} & P(y_{app_i}=1|\mathbf{x}_{app_i}) & \leq & 0.5, & \text{alors} & y_{app_i}=-1, \end{array} \tag{13.2}$$

pour $i = 1 \dots n_{app}$.
Les données $(\mathbf{x}_{app_i}, y_{app_i})_{i=1\dots n_{app}}$ ainsi étiquetées de manière binaire sont utilisées comme données d'apprentissage par le classifieur SVM classique.

On définit un autre jeu de données $(\mathbf{x}_{app_i}, \hat{y}_{app_i})_{i=1\dots n_{app}}$ tel que, pour $i = 1 \dots n_{app}$:

$$\begin{array}{lll} \text{si} \quad P(y_{app_i}=1|\mathbf{x}_{app_i}) > 1-\eta, & \text{alors} & \hat{y}_{app_i}=1; \\ \text{si} \quad P(y_{app_i}=1|\mathbf{x}_{app_i}) < \eta, & \text{alors} & \hat{y}_{app_i}=-1; \\ \hat{y}_{app_i} = P(y_{app_i}=1|\mathbf{x}_{app_i}) & \text{sinon.} & \end{array} \tag{13.3}$$

où η, introduit page 172, représente la précision de l'étiquetage.
Les données $(\mathbf{x}_{app_i}, \hat{y}_{app_i})_{i=1\dots n_{app}}$, comprenant à la fois des données d'étiquettes binaires et probabilistes sont, quant à elles, utilisées comme base d'apprentissage par un classifieur P-SVM.

On tire de manière aléatoire un nouveau jeu de points test $(\mathbf{x}_{test_i})_{i=1\dots n_{test}}$ afin de tester et comparer les prédictions de nos classifieurs.

Dans le cas du classifieur SVM classique, les probabilités d'appartenir à la classe d'étiquette "1" sont estimées en utilisant l'algorithme de Platt (décrit section 11.6, page 166). Dans le cas du classifieur P-SVM, les probabilités prédites sur les points $\mathbf{x}_{test}$ sont directement données par $\varphi(h(\mathbf{x}_{test}))$, où φ est la fonction logistique définie page 172 et h la fonction de prédiction.

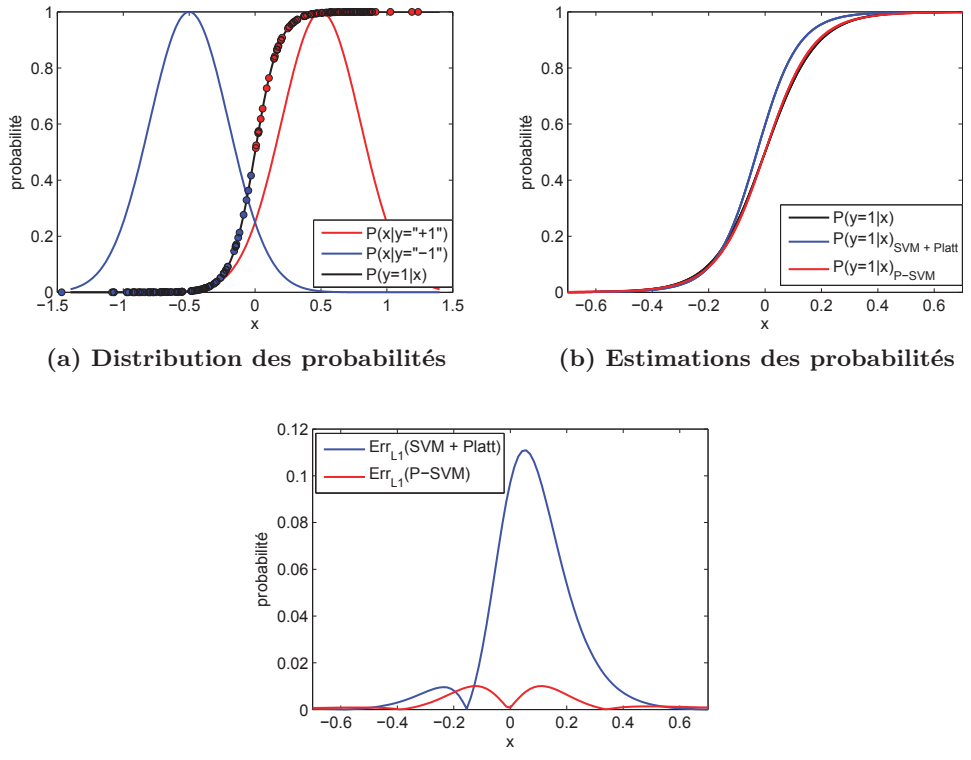
On estime l'erreur entre les probabilités vraies $(P(y_{test_i}=1|\mathbf{x}_{test_i}))_{i=1\dots n_{test}}$ (calculée comme en 13.1) et la probabilité estimée par les deux algorithmes sur le jeu de données test.

13.4.2 Résultat du test

La figure 13.2 illustre les résultats obtenus en utilisant $n_{app} = 100$ points d'apprentissage tirés aléatoirement selon les deux distributions $\mathcal{N}(\mu_{-1}, \sigma)$ et $\mathcal{N}(\mu_1, \sigma)$ telles que : μ_{-1}=-0.5, μ_{+1}=0.5 et σ=0.3. On choisit η=0.1 (A=2.2), C = 100 et $\hat{C}$= 100 et on utilise un noyau gaussien de paramètre $\sigma = 1$.
Elle permet une évaluation visuelle de l'amélioration des performances en termes de prédiction de probabilités apportée par les P-SVM en comparaison des performances de la combinaison SVM classique + Platt. Le tableau 13.1 permet une évaluation quantitative sur $n_{test} = 1000$ points de test tirés aléatoirement, en utilisant les métriques définies section 13.2.

	AUC	Acc	D_{KL}	Err_{Align}	D_{L1}	MCE	MSE
SVM + Platt	1	0.99	11	0.01	6.10^{-4}	0.13	8.10^{-4}
P-SVM	1	1	0.4	2.10^{-3}	1.10^{-5}	0.12	1.10^{-5}

Table *13.1* – *Mesures des performances de prédiction de probabilités réalisées par un classifieur P-SVM et par un SVM classique couplé à l'algorithme de Platt testés sur $n_{test} =$ 1000 points de test tirés aléatoirement. Le test est illustré sur la figure 13.2.*



(a) Distribution des probabilités

(b) Estimations des probabilités

(c) Erreur d'estimations

Figure *13.2* – *Comparaison de l'estimation des probabilités obtenues par l'algorithme de Platt appliqué à la sortie d'un SVM classique avec celles directement estimées par les P-SVM.*
(a) Représentation des distributions de probabilités des données d'apprentissage ($n_{app} = 100$), les points d'apprentissage sont représentés par des étoiles ;
(b) Représentation des probabilités prédites par le SVM classique couplé à l'algorithme de Platt d'une part (en bleu) et directement par P-SVM d'autre part (en rouge), superposées aux probabilités vraies (en noir) ;
(c) Représentation de l'erreur d'estimation : Err_{L1} représente l'écart moyen, en valeur absolue, entre les probabilités prédites et les probabilités "vraies".

13.5 Robustesse au bruit d'étiquetage

13.5.1 Description du test

On génère deux nuages de points (2-D), correspondant respectivement aux classes "-1" et "+1". De la même manière que précédemment (cf. formules (13.1)), on calcule les probabilités $P(y = 1|X = \mathbf{x})$ d'appartenance à la classe "+1". On simule alors des erreurs d'étiquetage en ajoutant artificiellement un bruit uniforme, noté δ, aux probabilités ainsi estimées. Pour $i = 1 \dots n$:

$$\hat{P}(y_i = 1|\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i) + \delta_i$$

On étiquette ensuite les données à partir de ces nouvelles probabilités $\hat{P}$ suivant la même méthode de seuillage que celle détaillée au paragraphe précédent (cf. formules (13.2) et (13.3)). On construit ainsi deux jeux d'apprentissage : l'un, d'étiquettes binaires, utilisé pour construire un SVM classique et l'autre, d'étiquettes mixtes, pour construire un P-SVM.

On génère aléatoirement un jeu de données test et on compare les performances des deux classifieurs en termes de précision d'étiquetage et d'estimation de probabilités sur ce jeu de test.

13.5.2 Résultats du test

On tire aléatoirement deux nuages de points d'apprentissage ($n_{app} = 100$) (50 points par distribution) distribuées selon des lois gaussiennes $\mathcal{N}(\mu_{-1}, \sigma)$ et $\mathcal{N}(\mu_1, \sigma)$ telles que : μ_{-1}=(0.3,0.5), μ_{+1}=(-0.3,-0.5) et σ=0.7. On choisit η=0.125 (A=1.9), C = 100 et $\hat{C}$= 100 ; on utilise un noyau gaussien de paramètre $\sigma = 1$.

Illustration sur des données linéairement séparables

La figure 13.3 illustre cette simulation lorsqu'un bruit uniforme δ d'amplitude 0.1 est appliqué sur les étiquettes des nuages d'apprentissage.

La figure 13.3a représente la vérité terrain. Les points d'apprentissage d'étiquettes ("$\pm$1") bruitées y sont représentés par des cercles bleus (classe "-1") et rouges (classe "+1"). Les probabilités "vraies" calculées suivant les formules explicitées en (13.1) sont représentées par l'échelle de couleurs.
La figure 13.3b représente le résultat de l'apprentissage sur les n_{app} points d'un classifieur SVM classique et l'estimation des probabilités réalisée par l'algorithme de Platt. La frontière de décision construite est tracée en noir tandis que les probabilités estimées sur l'ensemble de la grille sont représentées par l'échelle de couleurs.
Enfin, la figure 13.3c représente le résultat de l'apprentissage sur les n_{app} points d'un classifieur P-SVM. Là encore, la frontière de décision construite est tracée en noir et les

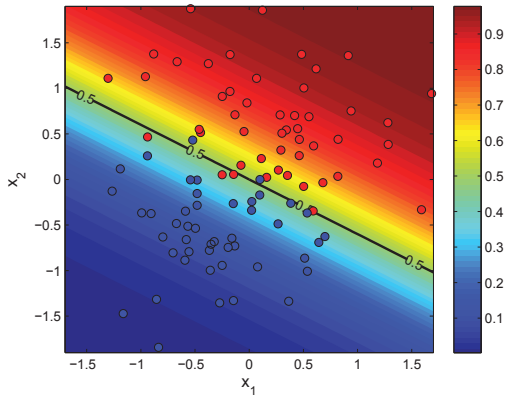
probabilités sont représentées par l'échelle de couleurs.

On remarque visuellement l'impact fort qu'ont les données bruitées sur la construction de la marge dans le cas d'un SVM classique. Celles-ci deviennent effectivement des points support intervenant directement sur la frontière de décision qui, pour satisfaire les contraintes de bonne classification et de généralisation, va autant que possible être localisée entre ces points.
Au contraire, la frontière de décision construite avec le P-SVM est quasiment superposable avec la séparatrice représentée sur la figure 13.3a.
On observe également que les probabilités prédites représentent mieux la distribution de probabilités "vraies" dans le cas du P-SVM que lorsqu'il est estimé avec l'algorithme de Platt à partir du score SVM classique.

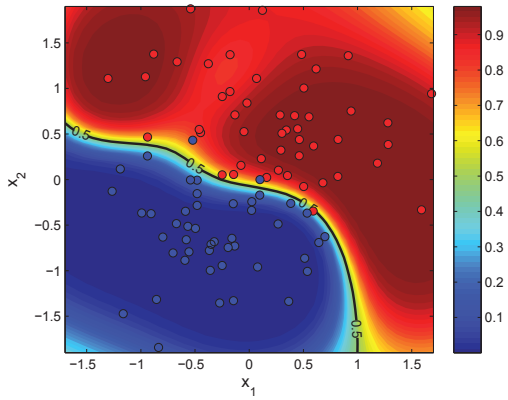
Le tableau 13.2 permet une évaluation quantitative des performances de prédiction. Elles sont mesurées sur $n_{test} = 1000$ points de tests tirés aléatoirement.

	AUC	Acc	D_{KL}	Err_{Align}	D_{L1}	MCE	MSE
SVM + Platt	0.99	0.95	175	0.04	0.16	0.80	0.04
P-SVM	1	0.99	13	5.10^{-3}	0.04	0.45	4.10^{-3}

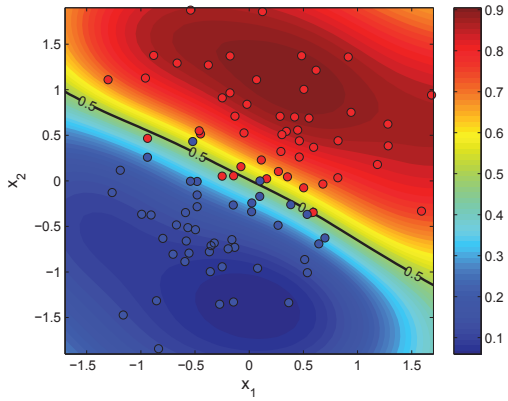
Table *13.2* – *Mesures des performances de prédictions réalisées par un classifieur SVM classique et un classifieur P-SVM exercés sur données bruitées (δ d'amplitude 0.1) et testés sur $n_{test} = 1000$ points de test tirés aléatoirement. Le test est illustré sur la figure 13.3.*



(a) Distribution des probabilités



(b) Estimations des probabilités en sortie du SVM + Platt



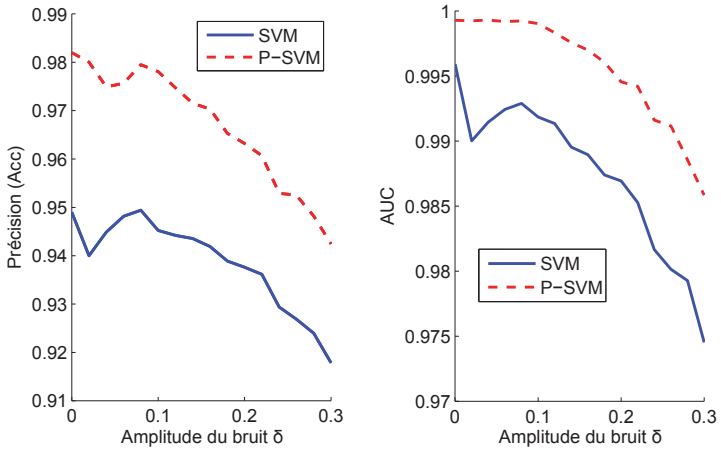
(c) Estimations des probabilités en sortie du P-SVM

Figure *13.3* – *Estimation des probabilités à partir d'un apprentissage sur données bruitées. Les points d'apprentissage d'étiquettes bruitées sont représentés par des cercles bleus (classe "-1") et rouges (classe "+1"). La frontière de décision construite est tracée en noir. Les probabilités estimées sur l'ensemble de la grille sont représentées par l'échelle de couleurs.*

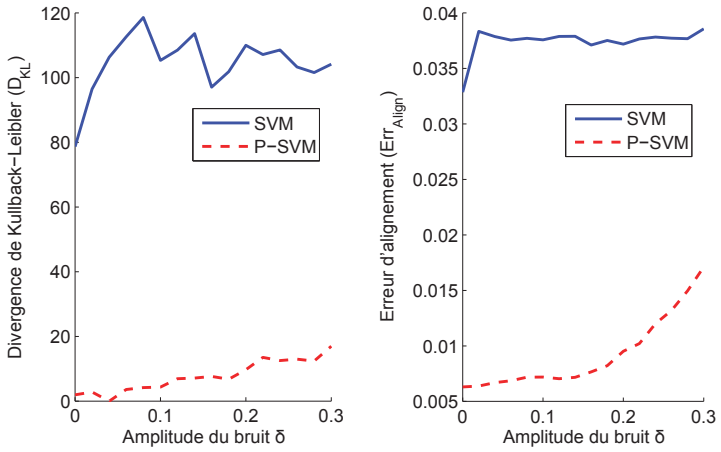
Impact de l'amplitude du bruit d'étiquetage

On tire aléatoirement $n_{app} = 100$ points d'apprentissage et $n_{test} = 1000$ points de test. On fait varier l'amplitude δ du bruit (uniforme) d'étiquetage des données d'apprentissage. Pour chaque valeur de δ dans [0, 0.3], on bruite aléatoirement les étiquettes des n_{app} données d'apprentissage (le processus est répété 100 fois, afin de moyenner les valeurs des indicateurs mesurés).

La figure 13.4 montre l'impact de l'amplitude du bruit d'étiquetage sur les performances des classifieurs. Comme attendu, on observe une décroissance des performances de classification (AUC et Acc) et de prédiction de probabilités (D_{KL} et Err_{Align}) avec l'augmentation du bruit d'étiquetage, pour le SVM classique (+ Platt) et le P-SVM. La comparaison des performances réalisées par ces deux algorithmes reste toujours favorable au P-SVM.



(a) Performances de classification



(b) Performances en estimation de probabilités

Figure *13.4* – *Robustesse au bruit en fonction de son amplitude. Evolution des performances de classification et d'estimation de probabilités en fonction de l'amplitude δ du bruit des SVM et P-SVM ($n_{test} = 1000$, tirage aléatoire répété 100 fois).*

Impact du nombre d'étiquettes probabilistes

On étudie l'évolution des performances du classifieur P-SVM en fonction de la proportion d'étiquettes probabilistes du jeu de données d'apprentissage.

On tire aléatoirement $n_{app} = 100$ points d'apprentissage et $n_{test} = 1000$ données test (afin d'évaluer les performances de discrimination).

Les étiquettes du jeu d'apprentissage sont bruitées avec un bruit uniforme δ d'amplitude 0.2. En suivant le schéma d'étiquetage présenté en (13.3), on obtient 35% d'étiquettes probabilistes et 65% de points d'étiquettes binaires dans le jeu de données $(\mathbf{x}_{app_i}, \hat{y}_{app_i})_{i=1...n_{app}}$ servant à l'apprentissage du P-SVM.

On fait ensuite arbitrairement varier le pourcentage t d'étiquettes probabilistes du jeu d'apprentissage de 5% à 35%.

Pour ce faire, on tire aléatoirement t échantillons parmi les 35 points d'étiquettes probabilistes. Les probabilités des 35-t données restantes sont alors *binarisées* en suivant le schéma d'étiquetage présenté en (13.2) (avec un seuil de 0.5). On obtient ainsi un jeu d'apprentissage avec une proportion d'étiquettes probabilistes égale à t.

Les résultats de ce test sont présentés sur la figure 13.5. Comme attendu, on observe que plus la proportion de données d'apprentissage étiquetées avec les probabilités est importante, plus la précision augmente et meilleures sont les estimations de probabilités. Remarquons que les performances du SVM classique (+ Platt), mesurées sur les données test (pointillées), sont constantes puisque seule varie la proportion d'étiquettes probabilistes présentes dans le jeu d'apprentissage (qui ne sont pas exploitées par le SVM, qui en utilise la version *binarisée*).

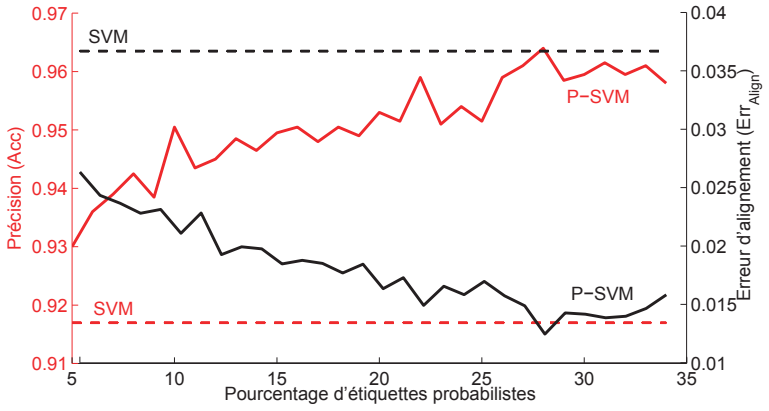


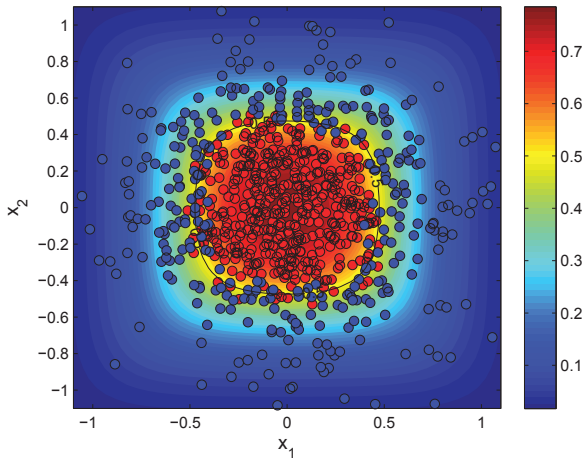
Figure *13.5* – *Evolution des performances de classification et d'estimation de probabilités du P-SVM en fonction de la proportion d'étiquettes probabilistes introduites à l'apprentissage. Comparaison avec les performances du SVM classique (+Platt).*

Illustration sur des données non linéairement séparables

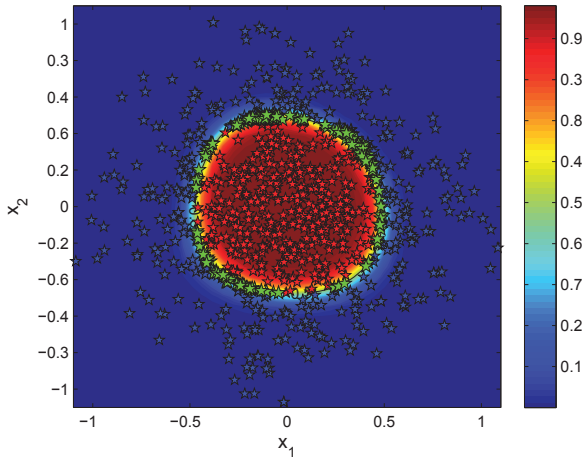
On propose d'illustrer par la figure 13.6 l'expérience de robustesse au bruit sur un cas non-linéairement séparable, exploitant et illustrant ainsi la puissance du SVM couplé à un noyau gaussien. L'image 13.6a présente la vérité terrain. Les $n_{app} = 800$ points d'apprentissage bruités sont représentés par des cercles bleus (classe "-1") et rouges (classe "1"). Ces points sont utilisés pour apprendre deux classifieurs de type SVM classique (couplé à l'algorithme de Platt, figure 13.6b) et P-SVM (figure 13.6c).
Quelques points parmi les $n_{test} = 8000$ points de test utilisés pour l'évaluation des performances sont représentés par des étoiles sur les figures 13.6b et 13.6a. Les étoiles bleues et rouges correspondent aux données correctement classifiées (classe "-1" et "+1"). Les étoiles vertes correspondent aux mauvaises classifications. Les performances obtenues sur les n_{test} points de test sont données dans le tableau 13.3.

	AUC	Acc	D_{KL}	Err_{Align}	D_{L1}	MCE	MSE
SVM + Platt	1	97%	2936	0.06	0.25	1.49	0.07
P-SVM	1	99%	59	4.10^{-4}	0.01	0.59	3.10^{-4}

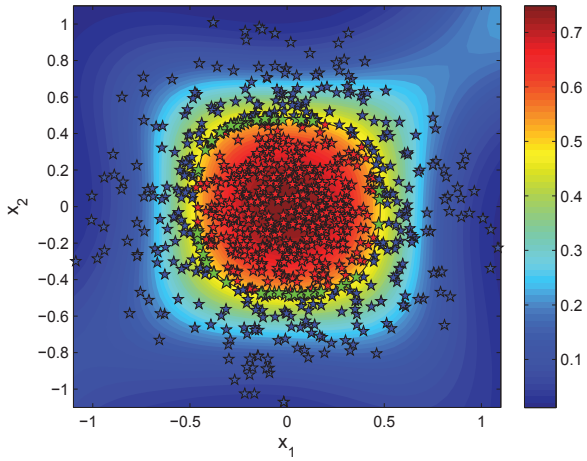
Table *13.3* – *Mesures des performances de prédiction réalisées par un classifieur SVM classique et un P-SVM exercés sur données bruitées (δ d'amplitude 0.1) et testés sur $n_{test} = 8000$ points de test tirés aléatoirement. Le test est illustré sur la figure 13.6.*



(a) Distribution des probabilités



(b) Distribution des probabilités



(c) Distribution des probabilités

Figure *13.6* – *Estimation des probabilités sur une grille de valeurs à partir d'un apprentissage sur données bruitées (bruit uniforme d'amplitude 0.1). Illustration du cas non linéairement séparable.*

13.6 Conclusion

Dans ce chapitre nous avons construit un ensemble d'exemples jouets afin de tester et comparer le SVM classique (+ Platt) et le schéma de classification P-SVM que nous avons proposé. L'objectif était d'évaluer à la fois les performances en classification et en estimation de probabilités dans différentes configurations et notamment en présence de données d'étiquettes bruitées (i.e. données d'apprentissage mal étiquetées). L'ensemble des comparaisons réalisées apparaît favorable à l'utilisation du P-SVM. En combinant classification et régression, le P-SVM s'avère efficace dans les deux tâches et relativement robuste au bruit d'étiquetage.

Chapitre 14

Tests sur données cliniques

14.1 Introduction

Dans ce chapitre, nous proposons d'appliquer les P-SVM aux données cliniques IRM-mp de prostate présentées partie II, en exploitant cette fois les scores affectés en *aveugle* par un radiologue expert (15 ans d'expérience). On se place ainsi dans le cas où la vérité histologique n'est pas disponible. On utilisera néanmoins la connaissance de la vérité terrain histologique comme référence pour l'estimation des performances.

14.2 Matériel et méthode

14.2.1 Base de données cliniques

La base de données cliniques est composée des images IRM-mp de 49 patients, acquises selon le protocole décrit section 6.3 page 90, pour lesquelles deux sessions d'annotations ont été réalisées : une analyse *a priori* en aveugle, effectuée par un radiologue expert lors de l'acquisition des images IRM-mp et une annotation *a posteriori* des images IRM-mp, faite après l'analyse des données histologiques (obtenues par prostatectomie) constituant la vérité terrain. Ces données correspondent au jeu de 30 patients de la base d'apprentissage utilisée dans la partie II, complété de 19 autres cas qui ont été ajoutés au cours de l'avancement de cette thèse. On notera que l'annotation des images de ces 19 nouveaux cas a été plus difficile pour diverses raisons : présence de petits foyers, cas où l'anomalie de signal est beaucoup plus importante/plus petite que la surface réelle de tissu malin mesurée sur les lames histologiques, cas où des foyers malins et suspects sont proches et difficile-

ment distinguables, etc. Elle a requis deux séances de relecture anatomo-radiologique *a posteriori* pour établir la vérité terrain sur les images.

Annotations en aveugle. L'analyse en aveugle est réalisée par un radiologue expert (15 ans d'expérience) qui contoure les régions d'intérêt (ROIs) suspectes et les annote selon un degré de suspicion échelonné de 0 à 4 de la manière suivante :

- *score* = 0 : bénignité certaine,
- *score* = 1 : bénignité probable,
- *score* = 2 : incertitude sur la ROI,
- *score* = 3 : malignité probable,
- *score* = 4 : malignité certaine.

Dans notre étude, nous ramenons ce score dans l'intervalle [0,1].

Annotations guidées par la vérité terrain histologique. La pièce de prostatectomie est analysée et toutes les zones de cancer sont délimitées précisément sur les lames histologiques. Chaque image IRM-mp est ensuite annotée après consensus entre radiologues et anatomo-pathologistes de manière à reporter la vérité histologique sur les images. On obtient alors un ensemble de ROIs de cancer, étiquetées "+1"(vrais positifs ou faux négatifs du radiologue) et des ROIs bénignes, étiquetées "-1". Celles-ci comportent des régions bénignes mais d'apparence suspecte sur les images IRM-mp (faux positifs des médecins ou régions fortement suspectées lors de l'analyse en aveugle) ainsi que des ROIs saines présentant un signal "normal", extraites aléatoirement sur la glande.

Finalement, cette étude repose sur un ensemble de 66 ROIs malignes {M}, 78 ROIs normales mais suspectes {NS} et 206 ROIs saines {N}.

14.2.2 Notations

Soit $(\mathbf{x}_i, l_i)_{i=1...n}$ un jeu de données où chacun des n vecteurs $\mathbf{x}_i$ représente une donnée d'apprentissage (i.e. un vecteur de caractéristiques descriptif d'une ROI) et où les l_i sont les étiquettes associées.

Dans la suite de ce chapitre, on distinguera trois types d'étiquettes :

- $l = c_{histo} \in \{-1, +1\}$ où c_{histo} est l'étiquette de classe correspondant à la vérité terrain histologique. On est ramené au cas idéal étudié dans la partie II de ce manuscrit où les étiquettes de classes l_i sont obtenues par la mise en correspondance des données histologiques et IRM ;
- $l = c_{expert} \in \{-1, +1\}$ où c_{expert} est l'étiquette de classe issue de l'analyse en aveugle des images IRM par un radiologue expert. On est ramené au cas exposé dans la section 10.1 où l'analyse de la référence histologique n'est pas accessible et où la vérité terrain est construite à partir de l'expertise du médecin. Les étiquettes c_{expert} sont construites par seuillage des scores de la manière suivante :

$$c_{expert} = \begin{cases} \text{-1 si } score < 0.5 \\ +1 \text{ si } score \geq 0.5\,; \end{cases}$$

- $l = s_{expert} \in \{-1, +1\} \cup [0, 1]$ où s_{expert} correspond au score expert. On a :

$$s_{expert} = \begin{cases} \text{-1 si } score = 0 \\ +1 \text{ si } score = 1 \\ score \text{ sinon} \end{cases}$$

Les étiquettes binaires c_{histo} et c_{expert} peuvent être utilisées comme données d'apprentissage par un classifieur SVM classique.
Les étiquettes mixtes (binaires et probabilistes) s_{expert} peuvent, elles, être utilisées comme données d'apprentissage par un classifieur P-SVM.

14.2.3 Distinction des problèmes de classification

On rappelle que notre base de données est structurée selon trois types de ROIs : malignes {M}, saines {N} et bénignes mais suspectes {NS}. Suivant l'approche employée dans la partie II de cette thèse, on distinguera deux problèmes de classification :

- PB1 : H0 = {N, NS} versus H1 = {M},
- PB2 : H0 = {NS} versus H1 = {M},

le deuxième problème de discrimination étant a priori plus difficile.

14.2.4 Procédure d'évaluation

L'évaluation des performances présentée ci-après section 14.4 est réalisée suivant la procédure de validation croisée type *leave-one-patient-out* (LOPO) décrite section 7.5.3 page 121.

14.3 Description du test

On teste et compare les résultats obtenus par les deux classifieurs SVM classique et P-SVM suivant deux types de bases de données servant à l'apprentissage.

14.3.1 Apprentissage sur données annotées suivant la référence histologique

Dans ce test, la vérité terrain utilisée pour l'apprentissage du classifieur a été établie à partir de la référence histologique (comme dans la partie II de cette thèse). Les étiquettes, que nous avons noté c_{histo}, sont de la forme {-1,+1}.

14.3.2 Apprentissage sur données annotées par l'expert

Dans ce test, on se met dans la situation où la vérité terrain histologique n'est pas accessible pour l'apprentissage du classifieur. L'apprentissage se base donc uniquement

sur les annotations, effectuées en "aveugle" par le radiologue expert. Les étiquettes, que nous avons notées s_{expert}, sont de la forme {-1, 0.25, 0.5, 0.75, +1}.
Dans le cas du P-SVM, l'intérêt ici est de considérer, dans l'apprentissage, l'incertitude des médecins en ne seuillant pas le score subjectif à 0.5 comme on l'aurait fait pour pouvoir réaliser un apprentissage par SVM classique, qui requiert des données binaires.
On compare les performances obtenues suivant que la référence utilisée pour l'évaluation des résultats tests est :

- **la base de données annotées par l'expert :** les prédictions réalisées sont-elles proches des scores annotés par l'expert ? L'objectif ici n'est alors pas fondamentalement d'améliorer les performances de l'expert mais d'être capable de mimer ses décisions. Dans ce cas, le système fourni pourra servir de "second avis" à un médecin peu expérimenté, par exemple.
- **la vérité terrain histologique :** les prédictions réalisées sur un système ayant appris sur une base de données annotée en "aveugle" par un radiologue expert peuvent-elles être aussi fiables que celles d'un système reposant sur une vérité terrain histologique objective ? Bien entendu, les conclusions sont corrélées aux capacités d'analyse de l'expert.

14.4 Résultats

Les tableaux 14.1, 14.2 et 14.3 suivants présentent les performances de *discrimination* et d'*estimation de probabilité* réalisées par les classifieurs (SVM/P-SVM) suivant le type d'annotations (*score expert*/*vérité histologique*) utilisées pour l'apprentissage et la référence (*score expert*/*vérité histologique*) prise en compte pour l'évaluation des prédictions effectuées lors du test. Les paramètres C et σ (noyaux gaussiens) des SVM et P-SVM correspondent à ceux optimisés dans l'étude présentée partie II ; pour le P-SVM, on choisit $\tilde{C} = C$, η=0.125 (les scores étant échantillonnés avec un pas de 0.25) et A=1.9.

14.4.1 Apprentissage et test sur la vérité terrain

On évalue ici les performances de classification d'un SVM classique (couplé à l'algorithme de Platt pour l'estimation des probabilités) sur la base de données $(\mathbf{x}_i, c_{histo,i})_{i=1...n}$ étiquetées en utilisant la référence histologique. Les résultats sont présentés dans le tableau 14.1.

Quelques remarques :
– attention, les mesures $\mathrm{D}_{\mathbf{KL}}$, $\mathrm{Err}_{\mathbf{Align}}$, $\mathrm{D}_{\mathbf{L1}}$, MCE et MSE évaluant la qualité des estimées de probabilités sont calculées par rapport à une vérité terrain binaire. Elles sont données à titre indicatif mais il vaut mieux se référer au valeurs d'AUC et de précision (Acc) pour l'évaluation des performances de classification.

– les performances mesurées en termes d'AUC sont moins bonnes que celles obtenues dans la section 8 de la partie II sur la base constituée des données de 30 patients. Comme nous l'avons souligné, les données des 19 nouveaux patients qui ont pu être ajoutées dans cette dernière étude sont plus complexes et ont nécessité plusieurs relectures histo/IRM avant de pouvoir établir une vérité terrain sur les images IRM-mp. Les performances moindres peuvent donc trouver une explication dans la difficulté d'analyse inhérente aux données ou dans la comparaison à une vérité terrain moins fiable.

		Acc	**AUC**	$\mathbf{D_{KL}}$	$\mathbf{Err_{Align}}$	$\mathbf{D_{L1}}$	**MCE**	**MSE**
PB1	SVM	0.87	0.86	77	0.31	0.15	0.35	0.10
PB2	SVM	0.73	0.76	77	0.26	0.34	0.71	0.21

Table *14.1* – *Performances SVM de discrimination réalisées sur le jeu de données annotées suivant la référence histologique.*
L'apprentissage et le test se font sur les données annotées suivant la référence histologique {-1,+1}.

14.4.2 Apprentissage et test sur les scores expert

On évalue et compare les performances obtenues en exerçant un SVM classique (+ Platt) et un P-SVM sur la base de données d'apprentissage dont les étiquettes sont les scores de suspicion affectés par le radiologue expert. On se place ainsi dans l'hypothèse où la référence histologique est inconnue.
Les *scores expert* sont ici utilisés à la fois comme étiquetage d'apprentissage et de test. Pour les P-SVM, l'apprentissage de la fonction de décision est réalisé sur la base de données $(\mathbf{x}_i, s_{expert,i})_{i=1...n}$, c'est-à-dire en utilisant directement les scores (normalisés entre 0 et 1) comme étiquettes.
Pour les SVM classiques, l'apprentissage est réalisé sur la base de données $(\mathbf{x}_i, c_{expert,i})_{i=1...n}$, c'est-à-dire en transformant les scores experts en scores binaires (seuillage à 0.5).
Durant la phase de test, l'évaluation de la prédiction se fait en utilisant pour référence les annotations de l'expert : l'évaluation des performances en *classification* compare les classes prédites aux scores experts seuillés $(c_{expert,i})_{i=1...n}$, tandis que l'évaluation des performances en *estimation de probabilités* compare les probabilités prédites aux scores expert $(s_{expert,i})_{i=1...n}$.
Les résultats sont présentés dans le tableau 14.2.
Ils montrent que les P-SVM permettent à la fois de bien classifier les données relativement à la *vérité expert* (AUC_{P-SVM}= 0.89) et de bien prédire les étiquettes ($D_{KL_{P-SVM}}$= 43) en fonction du degré de certitude de la base d'apprentissage. On observe qu'ils surpassent les SVM classiques tant en termes de *classification* (AUC_{SVM}= 0.85) qu'en *estimation de probabilités* ($D_{KL_{P-SVM}}$= 76).

		Acc	**AUC**	$\mathbf{D_{KL}}$	$\mathbf{Err_{Align}}$	$\mathbf{D_{L1}}$	**MCE**	**MSE**
PB1	P-SVM sur scores	0.91	0.89	43	0.25	0.15	0.31	0.06
	SVM sur scores seuillés	0.91	0.85	76	0.31	0.14	0.38	0.07
PB2	P-SVM sur scores	0.79	0.80	43	0.21	0.26	0.64	0.12
	SVM sur scores seuillés	0.79	0.78	76	0.25	0.27	0.81	0.14

Table *14.2* *– Performances de discrimination et d'estimation de probabilités pour les classifieurs SVM et P-SVM sur le jeu de données annotées par l'expert.*

14.4.3 Apprentissage sur les scores expert et test sur la vérité terrain

Nous évaluons ici les performances réalisées en utilisant les *scores expert* comme étiquetage d'apprentissage et la vérité terrain histologique comme étiquetage de test.
L'apprentissage se fait sur les données annotées par l'expert : pour les P-SVM en utilisant directement les scores $(s_{expert,i})_{i=1...n}$, et pour les SVM classiques en utilisant les scores seuillés à 0.5 $(c_{expert,i})_{i=1...n}$.
Le test se fait en utilisant pour référence la vérité terrain histologique $(c_{histo,i})_{i=1...n}$.
Les résultats sont présentés dans le tableau 14.3.
Ils nous montrent qu'utiliser les P-SVM sur une base de données étiquetées avec les scores experts issus d'une annotation *en aveugle* permet de bien classifier les données au sens de la vérité histologique. Inclure l'incertitude expert dans l'étape d'apprentissage *via* les P-SVM permet de pondérer l'influence des données incertaines lors de l'apprentissage et ainsi de réaliser de meilleures performances par rapport à la vérité terrain histologique (AUC_{P-SVM}= 0.86) que celles obtenues avec les SVM classiques (AUC_{SVM}= 0.82).

		Acc	**AUC**	$\mathbf{D_{KL}}$	$\mathbf{Err_{Align}}$	$\mathbf{D_{L1}}$	**MCE**	**MSE**
PB1	P-SVM sur scores	0.86	0.86	75	0.32	0.15	0.33	0.10
	SVM sur scores seuillés	0.85	0.82	118	0.38	0.15	0.43	0.12
PB2	P-SVM sur scores	0.72	0.74	75	0.28	0.36	0.68	0.22
	SVM sur scores seuillés	0.69	0.72	118	0.34	0.36	0.95	0.26

Table *14.3* – *Performances de discrimination obtenues en entraînant les classifieurs SVM et P-SVM sur le jeu de données annotées par l'expert et en effectuant le test avec la vérité histologique pour référence.*

14.5 Discussion et perspectives

Dans les deux cas de comparaison des SVM classiques et des P-SVM, on observe que les P-SVM surpassent les SVM en termes de performances de *classification* et de *prédiction de probabilités*.

Il est intéressant de remarquer que les performances de *classification* obtenues en entraînant le SVM sur la vérité histologique (voir tableau 14.1) et celles obtenues en exerçant les P-SVM sur les scores expert (voir tableau 14.3) sont proches (AUC= 0.86). Ceci suggère que l'expertise du radiologue pourrait être suffisante pour permettre la construction d'un classifieur robuste.

Il sera bien sûr nécessaire de confirmer ces bons résultats par une validation à plus grande échelle.

Ainsi, à condition que les données soient analysées par un radiologue ayant une bonne expertise, utiliser une base de données expert pourrait permettre d'atteindre des perfor-

mances de classification semblables à celles obtenues en utilisant une base de données reposant sur une corrélation anatomo-radiologique. Etant donnée la complexité de construction d'une telle base de données, qui représente une tâche longue, coûteuse et fastidieuse, les perspectives ouvertes par ces résultats préliminaires devraient être considérées avec attention.

Remarquons que l'apprentissage par P-SVM a ici été effectué sur les données annotées par un seul radiologue expert. Il sera intéressant d'utiliser conjointement les scores attribués par deux ou plusieurs experts. Ceci permettrait en effet, en moyennant les scores attribués par les $N_{experts}$ experts, une meilleure estimation des probabilités d'appartenance :

$$p_i = \frac{1}{N_{experts}} \sum_{i=1}^{N_{experts}} s_{expert_i},$$

avec une échelle de valeurs plus fine.
Une autre perspective concerne l'utilisation du paramètre η (la contrainte sur la prédiction de probabilité, lire section 12.3). Il serait en effet intéressant d'utiliser non pas une valeur fixe de η comme c'est le cas actuellement, mais une valeur adaptative, η_i, par point d'apprentissage i, de façon à ce que :

- lorsque les scores de suspicion de malignité attribués par les experts à la cible i sont très concordants (variance faible), η_i soit petit, traduisant ainsi la confiance dans l'annotation de la cible ;
- plus les scores attribués par les différents experts à la cible i sont discordants (variance forte), plus η_i augmente (on relâche la contrainte sur la prédiction), traduisant ainsi une plus grande incertitude sur la classe de cette cible.

14.6 Conclusion

Dans cette partie, nous avons proposé un schéma de classification supervisée reposant sur un étiquetage mixte (valeurs binaires/probabilités) des données d'apprentissage. Notre approche permet d'apprendre une unique fonction qui réalise à la fois l'objectif de discrimination des données étiquetées de manière "sûre" (binaire) et d'estimation de probabilités sur les données "incertaines" (en réalisant une régression probabiliste). Contrairement à la problématique d'estimation de probabilités à partir d'une fonction de discrimination SVM qui a été largement étudiée dans la littérature (voir les études comparatives de Niculescu-Mizil et Rüping [76, 104]), notre problématique d'apprentissage d'un SVM à partir d'un étiquetage mixte n'a pas, à notre connaissance, déjà fait l'objet de recherches.

Les résultats obtenus avec le P-SVM, à la fois dans le cadre d'exemples jouets et sur les données cliniques, montrent que la formulation proposée répond à notre problématique.

Chapitre 15

Conclusions

Contributions

Dans cette thèse, nous nous sommes intéressés à la conception d'un système d'aide au diagnostic (CAD) pour le cancer de la prostate utilisant des données d'imagerie IRM-multiparamétrique (IRM-mp).

*
* *

La première partie de cette thèse était consacrée à la description du contexte clinique et scientifique de ce travail. Après une introduction détaillant les caractéristiques du cancer de la prostate, les enjeux cliniques d'une imagerie IRM-mp en plein essor pour une meilleure visualisation des tissus prostatiques et les limites actuelles de l'interprétation des images, nous avons présenté l'état de l'art des méthodes d'aide au diagnostic du cancer de la prostate proposées dans la littérature.

La construction d'un système d'aide au diagnostic (CAD) implique différents choix méthodologiques à différentes étapes de la classification : définition des caractéristiques (ou *features*), réduction de la dimension de l'espace des caractéristiques, choix du classifieur, de la méthode de validation et du critère d'évaluation. Dans la deuxième partie de cette thèse, nous avons évalué les performances de différentes stratégies de sélection de caractéristiques et de classification pour l'élaboration d'un système automatique d'aide au diagnostic. Cette analyse a été effectuée sur une même base de données, rendant la comparaison objective. De plus, la base clinique sur laquelle s'appuie notre étude est riche

par le nombre de cas inclus et par l'annotation des données, réalisée de manière fiable et exhaustive en utilisant les pièces histologiques pour référence. Nous avons pu mettre en évidence un schéma de classification optimal, couplant la méthode de sélection par test t à un classifieur SVM (pour *Séparateur à vaste Marge*). Afin de quantifier l'apport d'un tel système dans la pratique clinique, nous avons testé notre prototype CAD auprès de douze radiologues d'expérience variable. Nous avons pu montrer que le "second avis" offert par notre CAD permettait non seulement d'améliorer les performances des experts dans la discrimination des tissus sains et malins mais aussi d'augmenter la confiance dans leur diagnostic.

L'une des limitations majeures lors de l'élaboration d'un système CAD supervisé est la richesse et la fiabilité de la base de données servant à l'apprentissage, les deux étant malheureusement souvent incompatibles. En effet, la pièce histologique de référence n'est pas souvent accessible et quand cette information est disponible, la construction d'une base de corrélations anatomo-radiologiques reste une tâche complexe et fastidieuse qui nécessite la mobilisation simultanée d'anatomo-pathologistes et de radiologues. Beaucoup d'études se limitent donc soit à l'utilisation d'une base de données histologiques/IRM de taille très limitée, soit à l'utilisation des résultats de l'analyse en aveugle des images IRM par un radiologue expert, seul.

Les scores de suspicion de malignité usuellement affectés par les radiologues aux cibles suspectées s'apparentent à des probabilités de présence du cancer. Afin de pouvoir les exploiter dans un schéma de classification SVM classique, il est nécessaire de les rendre binaire (en les seuillant de manière arbitraire pour affecter les cibles à la classe "cancer" ou "tissu sain") et ainsi d'enlever la part d'incertitude présente chez le praticien. Dans la dernière partie de cette thèse, nous avons proposé un nouveau schéma de classification, basé sur le SVM classique, permettant d'utiliser un étiquetage mixte, binaire et probabiliste, dans la phase d'apprentissage. Nous avons testé cette nouvelle approche sur un ensemble d'exemples jouets, nous permettant de mettre en évidence ses bonnes performances en classification et estimation de probabilités, ainsi que sa robustesse au bruit d'étiquetage, comparé aux résultats obtenus avec un SVM classique. Nous avons finalement appliqué notre méthode à notre base de données cliniques. Les résultats montrent qu'en combinant astucieusement classification et régression dans un même problème d'optimisation, notre méthode permet d'exploiter efficacement le degré de certitude de l'expert et d'obtenir, avec une vérité terrain expert, des performances proches de celles qui seraient réalisées si on pouvait connaître a priori la vérité histologique.

*
* *

Nos contributions principales sont :

Comparaison et évaluation de différentes stratégies de classification supervisée pour la discrimination des tissus prostatiques sains et malins :

- base de données d'apprentissage riche et fiable,
- extraction d'un grand nombre de caractéristiques de différents types,
- test de quatre méthodes de sélection de caractéristiques,
- test de quatre algorithmes de classification,
- comparaison sur deux problèmes de discrimination de difficultés différentes,
- confirmation de l'intérêt d'une approche multi-séquences et multi-attributs,
- mise en évidence d'un schéma de classification optimal.

Evaluation en conditions cliniques du système d'aide au diagnostic optimal :
- système utilisé comme "second avis",
- approche multi-lecteurs,
- quantification de l'apport en termes de confiance dans le diagnostic et en termes de performances.

Proposition d'un nouveau schéma de classification, basé sur les SVM, permettant de considérer à la fois des données étiquetées de manière certaine et des données étiquetées de manière probabiliste :
- mise en évidence de l'absence de considération de l'incertitude sur l'étiquetage,
- algorithme permettant l'apprentissage sur des données d'étiquettes mixtes (binaires et probabilistes),
- test sur données jouets,
- quantification de l'apport sur un jeu de données cliniques.

Limites et perspectives

De nombreuses pistes peuvent être envisagées pour améliorer les performances de l'approche CADx que nous avons proposée partie II.
Dans cette étude, nous avons fait des choix à tous les niveaux du schéma de classification (méthodes de sélection de type filtre, classification supervisée...). Les performances obtenues pour la tâche de discrimination la plus complexe ({NS} versus {M}) sont bonnes puisque nous avons pu montrer que le système permet d'aider le radiologue dans sa tâche diagnostique. Néanmoins, nous pourrions envisager une étude à plus grande échelle. Outre l'inclusion d'autres caractéristiques (tels que des paramètres image obtenus par décomposition en ondelettes, paramètres de Gabor ou encore paramètres de localisation spatiale dans la glande) nous pourrions évaluer les performances d'autres méthodes de sélection de caractéristiques (telles que les approches "enveloppantes") ou d'autres classifieurs (forêts aléatoires ou réseaux de neurones notamment). Il serait également intéressant de tester si la combinaison des sorties des classifieurs pourrait apporter une amélioration significative des résultats par rapport aux performances individuelles.

Dans notre travail, nous nous sommes largement focalisés sur l'utilisation du classifieur SVM dont les performances étaient meilleures sur nos données que celles des autres stratégies de classification testées (ADL, k-PPV, classifieur de Bayes). Il serait particulièrement intéressant de poursuivre l'étude de ce type de classifieur en testant l'approche MKL (pour *multiple kernels learning*) récemment proposée par Lanckriet [58]. Celle-ci consiste à apprendre une fonction de décision qui s'écrit comme une combinaison linéaire (convexe) de fonctions de décision SVM associées à chaque couple fonction noyau/caractéristique (ou un ensemble de caractéristiques, groupées par exemple selon la séquence IRM d'origine ou leur type). Cette approche permet simultanément l'optimisation des paramètres de chacune des fonctions de décision SVM (et noyaux associés) et la sélection du jeu de caractéristiques les plus discriminantes par le biais des poids de la combinaison linéaire apprise. Le problème s'exprime de la façon suivante (dans le cas d'un noyau linéaire trivial) :

$$\begin{cases} \min\limits_{\beta_k, \mathbf{w}_k, \xi} & \frac{1}{2}\Big(\sum\limits_{k=1}^{d} \beta_k \|\mathbf{w}_k\|\Big)^2 + C \sum\limits_{i=1}^{n} \xi_i \\ \text{tel que} & y_i\Big(\sum\limits_{k=1}^{d} \beta_k \mathbf{w}_k^\top \mathbf{x}_i + b\Big) \geq 1 - \xi_i, \qquad \xi_i \geq 0, \quad i = 1, \ldots, n. \end{cases}$$

Dans notre étude, nous nous sommes limités à l'utilisation de paramètres IRM. Or, d'autres modalités d'imagerie à fort potentiel méritent d'être testées. Il s'agit en particulier de techniques d'imagerie échographique actuellement en plein développement, telles que l'échographie de contraste ou l'élastographie. De même, les premières évaluations qualitatives de l'élastographie IRM (ERM) dans la discrimination des tissus prostatiques malins sont prometteurs. Ces modalités font l'objet de protocoles d'acquisition dans le cadre du projet INCa Cartographix (2011-2013) dans lequel les laboratoires LabTAU et CREATIS sont associés. Si le potentiel de ces méthodes d'imagerie est confirmé, nous proposerons d'inclure les paramètres d'élasticité, de contraste échographique et ERM dans notre prototype CADx en les ajoutant à la liste des caractéristiques discriminantes extraites de l'IRM. Néanmoins, pour pouvoir être utilisées conjointement, il sera nécessaire de développer des méthodes de recalage inter-modalités.
Cette étude des paramètres échographiques et ERM sera en particulier l'objet de la thèse de Jérôme Lehaire, poursuivant ce travail.

D'autre part, au cours de cette thèse, un nouvel imageur IRM à 3 Tesla (T) a été acquis à l'hôpital Edouard Herriot de Lyon. La base de données est à présent assez conséquente pour pouvoir tester la robustesse de notre schéma de classification sur ces nouvelles données et confirmer le niveau de performance mesuré sur les données de l'IRM 1.5 T.
De plus, dans un but expérimental, trois nouvelles acquisitions ont été ajoutées au protocole clinique : la séquence de diffusion à b=2000 s/mm^2, une carte d'anisotropie et une cartographie T2. Il sera nécessaire de quantifier leur apport respectif afin de savoir si ces acquisitions méritent, au regard du coût et de la durée de l'examen, d'être réalisées en

routine.

Le travail effectué dans une approche d'aide à la décision (CADx) peut être transposé relativement facilement à une approche d'aide à la détection (CADe) avec un schéma de classification identique. Nous avons d'ores et déjà travaillé sur cette transposition, en ajoutant à la fin de la chaîne de classification une étape de post-traitement simple utilisant des opérateurs morphologiques pour éliminer les micro-foyers de bruit et présenter une cartographie en *clusters* (voir la figure 15.1). Néanmoins, dans la perspective de l'établissement d'une carte de probabilité de présence des foyers tumoraux sur la glande, des méthodes plus pertinentes devront être mises en œuvre. Il sera important d'introduire une contrainte de régularité spatiale dans le classifieur, permettant d'éviter l'obtention d'une cartographie trop bruitée avec un nombre important de faux positifs comme on peut l'observer dans la littérature des CADe (cf. section 4.7). En effet, les systèmes d'aide au diagnostic prennent actuellement très peu en compte la distribution spatiale et anatomique des caractéristiques. Or la notion de voisinage peut en particulier être introduite dans le schéma d'optimisation du SVM. Quelques travaux ont d'ores et déjà été faits dans ce sens par Dundar *et coll.* [30], Tuia *et coll.* [127] et Cuignet [24], qui proposent respectivement l'introduction d'une pénalité de contiguité de la fonction de décision au voisinage direct, l'apprentissage d'un filtre spatial et une régularisation globale par graphe d'adjacence. Ces travaux ont été appliqués en imagerie cérébrale [24] ou satellitaire [30, 127] mais encore en imagerie de la prostate. Des atlas anatomiques de prostate sont par ailleurs en cours de développement ([143]). Leur utilisation comme information *a priori* dans les SVM (ajustement du score prédit sur les voxels suivant leur localisation dans la glande) est une piste de recherche.

La difficulté du passage à une approche CADe réside également dans le recalage interséquences rendu nécessaire par les mouvements involontaires du patient, les contractions du rectum au cours de l'acquisition et la présence de gaz intestinaux qui peut induire des distorsions.

Afin d'enrichir le système actuel, il serait également pertinent de mesurer les corrélations entre le degré de suspicion estimé par le CAD et l'agressivité tumorale mesurée par le score de Gleason (Gleason = 6, 7, 8, 9) afin de savoir si le CAD est plus spécifique aux tumeurs de haut grade. Il serait de même intéressant d'apprendre au classifieur à discriminer les tissus singuliers tels que les kystes, la néoplasie intraépithéliale de la prostate, les nodules myomateux, etc (cela ne sera envisageable que lorsque qu'un nombre de cas plus important correspondant à chacune des classes aura été inclus dans la base de données). Dans ces deux cas, une approche multi-classe devra être développée.

*
* *

Le travail réalisé sur les P-SVM apporte une contribution intéressante dans le cas où nous n'avons pas de base de référence (histologique $\simeq$ binaire) et où la vérité terrain

utilisée est une vérité expert (score $\simeq$ probabilité). Ce développement méthodologique a été mis à disposition de la communauté scientifique (mloss.org [1]) et doit permettre de répondre à d'autres problématiques que ce soit dans le domaine médical ou ailleurs.

Un autre axe de recherche doit également être exploré. En effet, lorsque la vérité histologique est connue, des imprécisions de report de cette vérité sur les images IRM peuvent survenir et impacter les résultats en particulier lors du passage à une approche cartographique. C'est par exemple le cas lorsqu'on considère de petits foyers, ou lorsque l'anomalie de signal visible à l'IRM est beaucoup plus importante/plus petite que la surface réelle mesurée sur les lames histologiques, ou encore lorsque des foyers malins et des foyers sains mais suspects sont proches et non distinguables, etc. Dans notre étude, nous avons rejeté les données pour lesquelles le report s'avérait hasardeux (cela représente environ 1/4 des données). Cette incertitude dans la segmentation a d'autant plus d'impact lorsque l'apprentissage des fonctions de discrimination se fait à un niveau pixel et non plus à un niveau ROI (en synthétisant l'information sur l'ensemble des pixels d'une ROI, l'impact de l'incertitude sur quelques pixels est moindre). On pourra envisager d'adapter notre approche P-SVM à cette problématique sans toutefois omettre de considérer d'autres approches, notamment les travaux récents de Stempfel et Ralaivola [118] qui proposent d'intégrer, dans le schéma d'optimisation SVM classique, une probabilité d'inversion (*flip*) de l'étiquetage {-1, +1}.

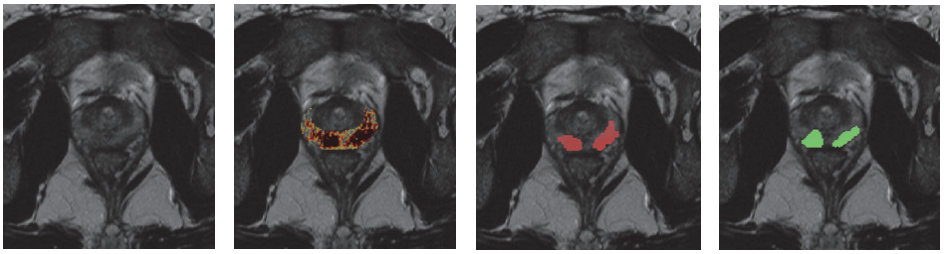


Figure *15.1* – *Exemple de transposition de la méthode CADx à la cartographie (CADe). On utilise la fonction discriminante optimale (test t + SVM) construite par apprentissage sur un ensemble de ROIs pour prédire la classe des pixels de l'image. On obtient ainsi une carte de scores. Une étape de post-traitement (morphologie mathématique) permet d'isoler des clusters et donc de définir des cibles suspectes.*

1. 366 téléchargements à l'heure où nous imprimons cette thèse.

Publications de l'auteur

Revues Internationales avec comité de lecture

- [publié] E. Niaf, O. Rouvière, F. Mège-Lechevallier, F. Bratan, and C. Lartizien, "Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI", Phys. Med. Biol., vol. 57, no. 12, pp. 3833-3851, 2012
- [soumis] F. Bratan, E. Niaf, C. Melodelima, A-L. Chesnais, R. Souchon, F. Mège-Lechevallier and O. Rouvière, *Factors influencing prostate cancer detection and localization at multiparametric MRI : a prospective study*, soumis à Radiology
- [soumis] A-L. Chesnais, E. Niaf, F. Bratan, F. Mège-Lechevallier, S Roche, M Rabilloud, M Colombel and O. Rouvière, *Differentiation of transitional zone prostate cancer from benign hyperplasia nodules : evaluation of discriminant criteria a multiparametric MRI*, soumis à European Radiology
- [en rédaction] E. Niaf, R. Flamary, O. Rouvière, C. Lartizien, S. Canu, *Kernel-based learning from both qualitative and quantitative labels : application to cancer diagnosis*, en rédaction pour IEEE Transactions on Signal Processing
- [en rédaction] E. Niaf, C. Lartizien, F. Bratan, N. Girouin, M. Papillard, G. Pagnoux, T. Vitry, L. Chamard, G. Renosi, T. Sanzalone, F. Mège-Lechevallier, O. Rouvière, *Clinical evaluation of a computer-aided decision system for prostate cancer screening in the peripheral zone : a multi-reader ROC study*

Congrès internationaux

- E. Niaf, R. Flamary, S. Canu, O. Rouvière, and C. Lartizien, *Handling learning samples uncertainties in SVM classification : application to MRI-based prostate cancer computer-aided diagnosis*, IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, Spain, abstract, 05/2012
- C. Lartizien, M. Rogez, A. Susset, F. Giammarile, E. Niaf, and F. Ricard, *Computer aided staging of lymphoma patients with FDG PET/CT imaging based on textural information*, IEEE International Symposium on Biomedical Imaging (ISBI), pp. 118-121, Barcelona, Spain, 05/2012
- E. Niaf, O. Rouvière, and C. Lartizien, *Computer-aided diagnosis for prostate cancer detection in the peripheral zone via multisequence MRI*, SPIE Medical Imaging, vol. 7963, Orlando, Florida, 02/2011. **Honorable poster award**.
- E. Niaf, R. Flamary, C. Lartizien, and S. Canu, *Handling uncertainties in SVM classification*, IEEE Workshop on Statistical Signal Processing (SSP), pp. 757-760, Nice, France, 06/2011.
- E. Niaf, O. Rouvière, and C. Lartizien, *Computer-Aided Diagnosis system for assisting radiologists in prostate cancer detection with Multiparametric MRI*, Focal Therapy and Imaging in Prostate and Kidney Cancer, Noordwijk, Netherlands, 05/2011
- F. Bratan, E. Niaf, A. - L. Chesnais, R. Souchon, F. Mège-Lechevallier, and O. Rou-

vière, *Detection and Localization of Prostate Cancers Using Multiparametric Magnetic Resonance Imaging*, Radiological Society of North America (RSNA) annual meeting, Chicago, Illinois, 11/2011

Congrès nationaux

- E. Niaf, C. Lartizien, F. Bratan, N. Girouin, M. Papillard, G. Pagnoux, T. Vitry, L. Chamard, G. Renosi, T. Sanzalone, F. Mège-Lechevallier, O. Rouvière, *Apport d'un système CADx pour la caractérisation des zones suspectes en IRM multiparamétrique de prostate*, Journées Françaises de Radiologie (JFR) , Paris, 10/2012
- E. Niaf, O. Rouvière, and C. Lartizien, *Aide au diagnostic du cancer de la prostate par IRM multi-sequences : une approche par classification supervisée*, Recherche en Imagerie et Technologies pour la Santé (RITS), Rennes, 04/2011
- E. Niaf, O. Rouvière, F. Bratan, A. - L. Chesnais, F. Mège-Lechevallier, and C. Lartizien, *Diagnostic assisté par ordinateur pour la détection du cancer de la prostate par IRM multi-paramétrique*, Journées Françaises de Radiologie (JFR) , Paris, 10/2011. **Prix poster mention honorable**.
- T. Vitry, E. Niaf, N. Girouin, R. Boutier, M. Papillard, and O. Rouvière, *L'IRM multiparamétrique peut-elle améliorer la détection des récidives post-HIFU du cancer prostatique ?*, Journées Françaises de Radiologie (JFR), Paris, 10/2011
- A. - L. Chesnais, F. Mège-Lechevallier, F. Bratan, E. Niaf, and O. Rouvière, *Nodules de la zone de transition prostatique : critères IRM de malignité*, Journées Françaises de Radiologie (JFR) , Paris, 10/2011

Annexes

Annexe A

Résultats complémentaires sur la comparaison des schémas CADx

A.1 Liste des caractéristiques sélectionnées

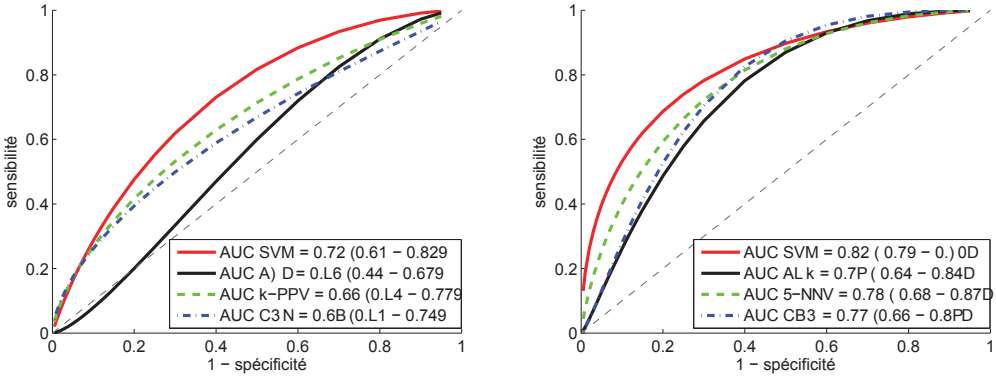
Le tableau A.1 présente la liste des 25 premières caractéristiques sélectionnées par chacune des 4 méthodes de sélection (test t, information mutuelle, mRMRd et mRMRq).

A.2 Courbes ROC modélisées

A titre d'exemple, le lecteur trouvera sur la figure A.1 le résultat de la modélisation des courbes ROC obtenu en utilisant un modèle bi-normal (ROCKIT®, version 1.1-beta, CE Metz, Université de Chicago). Ces courbes correspondent à la modélisation des résultats présenté figure 8.3, page 130.

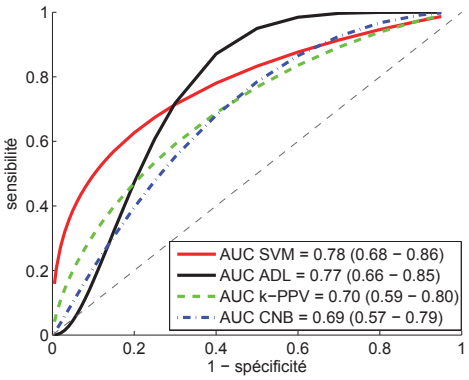
test t	IM	mRMRd	mRMRq
ADC Sobel_y MED	T2-w gradient δ_y MOY	ADC Sobel_y MED	ADC Sobel_y MED
T2-w gradient δ_y MOY	T2-w Sobel_y MOY	DCE T_{max} 75%	DCE T_{max} 75%
T2-w Sobel_y MOY	DCE WI MED	T2-w var MOY	T2-w ver MOY
ADC MED	T2-w var MOY	DCE T_{max} MOY	DCE WI MED
DCE MOY	ADC MED	ADC homo MED	T2-w gradient δ_y MOY
DCE MED	DCE T_{max} MOY	T2-w gradient δ_y MOY	ADC MED
T2-w Sobel_{xy} MOY	DCE MOY	DCE ve MED	DCE T_{max} MOY
ADC MOY	DCE MED	ADC MED	ADC homo MED
DCE AUGC 25%	T2-w var MED	ADC energ MED	T2-w inf1 MED
ADC 25%	ADC 25%	DCE T_{pic} MED	ADC Sobel_{xy} MED
T2-w Sobel_{xy} MED	T2-w inf1 MOY	ADC Sobel_{xy} MED	DCE ve 25%
ADC var MED	DCE AUGC 25%	T2-w entro MOY	DCE MOY
DCE T_{max} MOY	ADC Sobel_{xy} MED	DCE AUGC 25%	T2-w Sobel_y MOY
T2-w dissi MOY	ADC MOY	T2-w Sobel_{xy} MED	T2-w entro MOY
DCE 75%	DCE ve 25%	ADC cshad MOY	ADC 25%
T2-w dissi MED	T2-w Sobel_{norm} MOY	DCE WO 25%	T2-w Sobel_{xy} MED
T2-w Sobel_{xy} MED	ADC Sobel_{xy} MOY	T2-w 25%	T2-w 25 %
T2-w dent MED	T2-w Kirsh_{max} MED	ADC corrm MOY	ADC Sobel_{xy} MOY
T2-w Sobel_{norm} MOY	DCE T_{pic} MOY	DCE WI MED	DCE AUGC 25%
DCE WI 25%	T2-w Sobel_{xy} MOY	T2 corrm MED	DCE T_{pic} MED
T2-w sent MED	T2-w entro MOY	ADC gradient δ_x MED	T2-w var MED
ADC Sobel_{xy} MOY	T2-w homo MED	ADC maxpr MOY	T2-w Sobel_{xy} MOY
T2-w std MOY	T2-w sent MOY	DCE T_{onset} MOY	ADC cshad MOY
T2-w dent MOY	T2-w inf1 MED	ADC Sobel_{xy} MOY	DCE T_{pic} MOY
T2-w Kirsh_{max} MOY	T2-w inf2 MED	ADC 25% MOY	T2-w homo MED

Table *A.1* – *Liste des caractéristiques descriptives les plus discriminantes sélectionnées parmi les 25 premières pour chacune des quatre méthodes de sélection. Les abréviations de nom utilisées pour référencer chacune des caractéristiques sont données section 7.2. Les abbréviation MOY, MED, 25% et 75% correspondent respectivement à la valeur moyenne, médiane, 1er et 3^{e} quartile calculés pour résumer les valeurs sur une ROI.*

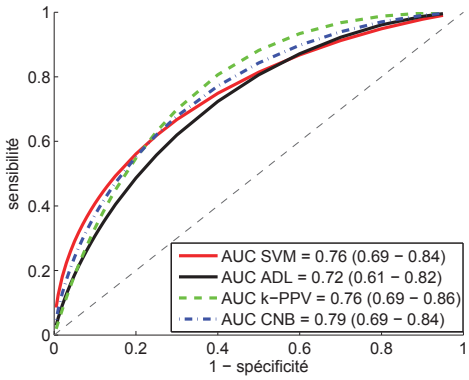


(a) sans selection

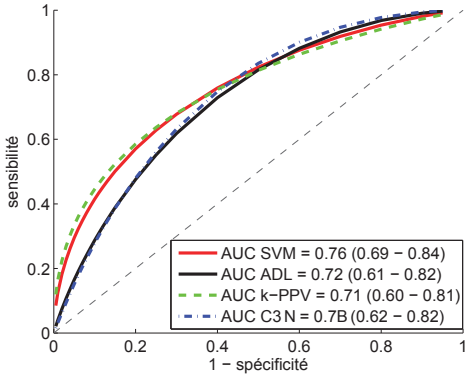
(b) test t



(c) IM



(d) mRMRd



(e) mRMRq

Figure *A.1* – *Performances des SVM, ADL, k-PPV et CNB en fonction du jeu de caractéristiques utilisé : (a) en utilisant toutes les caractéristiques extraites (sans sélection) ; (b) avec une sélection basée sur un test t. Ces courbes ROC modélisées (modèle paramétrique bi-normal) correspondent à la tâche de discrimination PB2 : H0={NS} versus H1={M}.*

Annexe B

Méthodes statistiques développées pour l'étude du CADx en condition clinique

B.1 Estimation non-paramétrique de l'aire sous la courbe (AUC) des courbes ROC

On cherche à évaluer les performances diagnostiques de chaque lecteur lors des trois lectures (première, deuxième et avec le CADx) par le biais des scores (entre 0 et 4) attribués. L'aire sous la courbe (AUC) est utilisée ici comme critère pour évaluer ces performances.

Les données de 30 patients sont utilisées pour l'évaluation, en condition clinique, du système CADx proposé. Ces patients sont repérés par un identifiant de 1 à 30.

L'unité diagnostique est ici une zone "cible" ou ROI, contourée sur la prostate ; un patient peut présenter plusieurs cibles suspectes. Le jeu de données comprend les résultats des lectures 1, 2 et avec le CADx pour l'ensemble des cibles ({NS} et {M}) analysées sur les 30 patients. Les données ont ici une structure hiérarchique. Deux niveaux sont présents : le niveau 1 "patient", et le niveau 2 "cible". Les résultats des cibles d'un même patient i peuvent être liés, à cause des caractéristiques de ce patient qui sont partagées pour toutes ses cibles. On parle dans ce cas de données en "*cluster*" (*clustered data*), le *cluster* étant ici le patient. Ceci doit être pris en compte dans les analyses : sinon, les résultats peuvent être erronés, notamment au niveau du calcul des intervalles de confiance des estimations de l'AUC.

Rutter (Academic Radiology, 2000) [106] a proposé une méthode non-paramétrique, reposant sur des estimations par "*bootstrap*" pour estimer l'AUC d'un score diagnostic dans le cas de données en *cluster*. La méthode "*bootstrap*" est une méthode générale d'estimation. Elle est souvent utilisée pour les calculs de variance et d'intervalles de confiance, lorsqu'il n'existe pas de formule simple pour les calculer. Pour un paramètre d'intérêt μ (ici ce sera l'AUC), l'idée globale est de simuler N jeux de données à partir des données réelles. Pour chaque jeu de données n simulé, on estime le paramètre d'intérêt sur le jeu

de données simulé. N "pseudo-estimations" sont ainsi obtenues. La théorie nous dit que la moyenne empirique de ces N "pseudo-estimations" estime correctement le paramètre μ et des intervalles de confiance peuvent être obtenus. Pour plus de détails sur le *bootstrap*, on pourra se référer aux livres d'Effron et al [32] et de Manly [70].

L'idée de Rutter pour prendre en compte les données en *cluster* lors des simulations est d'effectuer un tirage aléatoire sur les données au niveau du *cluster* patient (et non pas au niveau de la cible) : ici, $N = 2000$ simulations ont été réalisées (2000 jeux de données ont été simulés). Le jeu de données simulé pour chaque simulation n est obtenu de manière suivante :

- 30 tirages avec remise sont effectués parmi les 30 patients. Chaque jeu de données simulé contient donc les données de 30 pseudo-patients (avec possiblement des doublons) ;
- les données de l'ensemble des cibles des 30 patients tirés aléatoirement avec remise constituent le jeu de données simulé noté n.

L'AUC est ensuite estimée sur le jeu de données simulé n de manière non-paramétrique par :

$$\hat{A} = \frac{1}{n_M . n_{\hat{M}}} \sum_{\substack{j \in \{\text{unité } \hat{M}\} \\ j \in \{\text{unité } M\}}} \psi(t_i, t_j), \tag{B.1}$$

où t_i indique la valeur du score pour une cible i cancéreuse, t_j indique la valeur du score pour une cible j non-cancéreuse, n_M indique le nombre de cibles réellement cancéreuses malades, $n_{\hat{M}}$ le nombre d'unités réellement non cancéreuses, et

$$\psi(t_i, t_j) = \begin{cases} 1 \text{ si } t_i > t_j \\ 0,5 \text{ si } t_i = t_j \\ 0 \text{ si } t_i < t_j. \end{cases} \tag{B.2}$$

Remarques sur cet estimateur non-paramétrique : comme remarqué dans Rutter [106], l'estimateur non-paramétrique de l'AUC donné en (B.1) est en fait une statistique de Mann-Whitney sur l'ensemble des couples (unité malade, unité non-malade). Cette statistique fait partie de la classe plus générale des statistiques de type U. Notons à ce propos que :

- pour déterminer un intervalle de confiance (IC), il faut supposer que les résultats des couples d'unités malades et non-malades sont indépendants. Ce n'est pas le cas ici, la corrélation provenant du patient pouvant présenter des unités malades et non-malades. Les formules classiques pour calculer les IC ne sont donc plus valides.
- la théorie des *bootstraps* est valide sur les U-statistiques (des références à ce sujet sont données dans [**?**]). Néanmoins, les unités tirées aléatoirement lors de l'échantillonnage doivent être indépendantes pour que la théorie du *bootstrap* reste valide. L'idée de Rutter est de tirer aléatoirement parmi les patients (les unités 'patients' sont bien indépendantes les unes des autres), et non pas directement parmi les cibles.

L'AUC pour une lecture est estimée par $\frac{1}{2000} \sum_{n=1}^{2000} \widetilde{A_n}$, où $\widetilde{A_n}$ est calculée comme en (B.1) sur l'échantillon de *bootstrap* n. De plus, on dispose de 2000 quantités $\widetilde{A_n}$.

Calcul des IC à 95%

Il existe en fait plusieurs méthodes pour déterminer des IC à partir d'un échantillonnage par *bootstrap*. Citons en trois :

- *Standard bootstrap method :* cette méthode présuppose que l'estimateur $\hat{A}$ suit une loi normale (ou s'en rapproche suffisamment). Les bornes de l'IC sont données par $\hat{A} \pm z_{\frac{\alpha}{2}}.\hat{\sigma}$, où $z_{\frac{\alpha}{2}}$ est le quantile d'ordre 97.5% d'une loi normale centrée réduite (de moyenne 0 et d'écart-type 1), et où $\hat{\sigma}$ est l'écart-type empirique déterminé sur l'échantillon de *bootstrap* (cf. Manly, [70] p.34) ;

- *Simple percentile method :* les intervalles de confiance à 95% de l'estimation de l'AUC peuvent être obtenus à l'aide des percentiles d'ordre 0.025 et 0.975 du vecteur des 2000 $\widetilde{A}_n$ (cf. Manly, [70] p.39) ;

- *Bias-corrected percentile method :* (cf. Manly, [70] p.44 pour plus de détail) c'est cette méthode qui a été retenue. Elle correspond à un raffinement de la méthode précédente et se déroule en plusieurs étapes :
  a) calcul de la proportion p d'échantillons vérifiant $\widetilde{A}_n \geq \hat{A}$, pour n allant de 1 à 2000 :
  b) détermination du quantile z_0 d'ordre $1-p$ d'une loi normale centrée réduite ;
  c) les bornes de l'IC sont les percentiles empiriques d'ordre $\phi(2z_0 - z_{\frac{\alpha}{2}})$ et $\phi(2z_0 + z_{\frac{\alpha}{2}})$, où ϕ est la fonction de répartition d'une loi normale centrée réduite, et $z_{\frac{\alpha}{2}}$ est le quantile d'ordre $97,5\%$.
  Notons que si $p = 50\%$, la *Bias-corrected percentile method* correspond exactement à la *Simple percentile method*.

Avec cette procédure d'estimation, les AUC ont été estimées (avec IC à 95% par Bias-corrected percentile method) par :

	Première lecture	Deuxième lecture	Lecture avec CADx
lecteur 1	82.4 [74.8 ;89.6]	85.8 [78.5 ;92.8]	88.1 [81.5 ;94]
lecteur 2	84.7 [78.3 ;90.6]	86.1 [79.2 ;92.1]	87.2 [80 ;93.9]
lecteur 3	88.9 [82.6 ;94.6]	90.5 [84 ;96.2]	92.7 [87.5 ;96.9]
lecteur 4	81.9 [73.3 ;89.7]	84.5 [75.8 ;91.9]	88.8 [82 ;94.5]
lecteur 5	80.2 [71.3 ;88.1]	80.3 [71.2 ;88.6]	84.8 [76.6 ;92]
lecteur 6	79 [70.8 ;86.9]	76.6 [66.5 ;85.9]	81.2 [72 ;89.4]
lecteur 7	79.2 [69.9 ;87.5]	84.3 [76.7 ;91.1]	87.8 [79.7 ;94.4]
lecteur 8	78.2 [66.8 ;88.3]	81.4 [71.2 ;90.8]	86.2 [77 ;94.1]
lecteur 9	79.2 [71.8 ;86.5]	77.6 [68.3 ;85.9]	82.7 [74.1 ;90.6]
lecteur 10	82 [75.7 ;87.8]	78.8 [71.1 ;86.4]	83.1 [74.9 ;90.7]
lecteur 11	64.2 [55.1 ;72.9]	75.8 [67.6 ;83.9]	79.5 [71.7 ;86.6]
lecteur 12	80.6 [73.1 ;87.6]	80.6 [72.6 ;88.4]	86.1 [78 ;93.1]

Par un procédé identique à celui exposé précédemment, on peut estimer la différence d'AUC entre deux lectures et fournir des intervalles de confiance. Si on note A et B les AUC respectives de deux lectures, il suffit de calculer, pour chaque simulation n, la différence $\widetilde{A}_n - \widetilde{B}_n$. La différence $A - B$ entre les AUC peut être estimée par $\frac{1}{2000}\sum_{n=1}^{2000}(\widetilde{A}_n - \widetilde{B}_n)$ et les intervalles de confiance à 95% peuvent être obtenus à l'aide de la *Bias-corrected percentile method* appliquée au vecteur des 2000 $\widetilde{A}_n - \widetilde{B}_n$.

Les différences d'AUC entre deux lectures ont été estimées (avec IC à 95% par *Bias-corrected percentile method*) par :

	Lecture 2 - lecture 1	Lecture CADx - lecture 1	Lecture CADx - lecture 2
lecteur 1	3.4 [-2.9 ;9.5]	5.7 [0.1 ;11.7]	2.3 [-2.9 ;7.2]
lecteur 2	1.4 [-4.7 ;6.9]	2.5 [-4 ;8.9]	1.2 [-5.8 ;8.3]
lecteur 3	1.6 [-2.2 ;4.9]	3.8 [-1.5 ;9.3]	2.2 [-2.5 ;7.1]
lecteur 4	2.5 [-3.4 ;8.7]	6.9 [0.4 ;13.7]	4.4 [-0.7 ;9.9]
lecteur 5	0.1 [-6.4 ;6.1]	4.6 [-1.4 ;10.7]	4.5 [-0.4 ;10]
lecteur 6	-2.4 [-10.3 ;4.6]	2.2 [-7 ;10]	4.6 [-3.5 ;12.2]
lecteur 7	5.1 [-3.3 ;14.5]	8.6 [-0.8 ;18]	3.4 [-2.2 ;9.4]
lecteur 8	3.2 [-7.3 ;13.8]	8 [-0.1 ;17]	4.8 [-2.9 ;13.7]
lecteur 9	-1.6 [-11.8 ;7.6]	3.5 [-5.7 ;12.2]	5.1 [-0.9 ;11.2]
lecteur 10	-3.2 [-9.9 ;3.4]	1.1 [-5.6 ;7.6]	4.3 [-4.6 ;13.5]
lecteur 11	11.6 [0.9 ;22.1]	15.3 [5.3 ;26.1]	3.7 [-0.8 ;8.5]
lecteur 12	0 [-4.3 ;4.5]	5.5 [1.2 ;10.1]	5.5 [1.6 ;9.7]

B.2 Modélisation des courbes ROC pour les lectures 1, 2 et avec CADx

B.2.1 Rappel sur les courbes ROC et notations

La variable Y désigne le résultat du test diagnostique, D le statut de la cible ($D = 1$ pour une cible malade et $D = 0$ pour une cible non-malade), X est un vecteur de covariables communes aux unités malades et non-malades (par exemple, X représente le couple (lecteur, lecture)), et X_D est un vecteur de covariables spécifiques aux unités malades.

Notons TPF_{X,X_D} et FPF_X les fonctions dites de 'survie' de Y chez les unités malades et non-malades respectivement, c'est-à-dire que pour un seuil s donné :

- $TPF_{X,X_D}(s) = P(Y \geq s \mid D = 1, X, X_D)$ est la probabilité que le score pour une unité malade de covariables X et X_D soit supérieur ou égal à s ou encore que $TPF_{X,X_D}(s)$ corresponde à la fraction de vrais positifs au seuil s (TPF : *True Positive Fraction*), i.e. la sensibilité au seuil s.
- $FPF_X(s) = P(Y \geq s \mid D = 0, X)$ est la probabilité que le score pour une unité non-malade de covariables X soit supérieur ou égal à s ou encore que $FPF_X(s)$ corresponde à la fraction de faux positifs au seuil s (FPF : *False Positive Fraction*), i.e. la différence 1 - spécificité au seuil s.

La courbe ROC_{X,X_D} est l'ensemble des couples $(FPF_X(y), TPF_{X,X_D}(y))$, pour $y = 0, ..., 5$ (la valeur 5 a été choisie arbitrairement comme valeur > 4, correspondant à un TPF et à un FPF de 0).

Etant donné X, $T_X = \{FPF_X(s) \mid s = 0, ..., 5\}$ est une collection de 6 fractions de faux positifs distinctes : $FPF_X(0) = 1 > FPF_X(1) > ... > FPF_X(4) > FPF_X(5) = 0$. Ainsi, pour chaque fraction de faux positifs $t \in T_X$, on peut définir ${FPF_X}^{-1}(t)$ comme étant l'unique seuil associé à cette fraction de faux positif t. Le seuil associé à $t_0 = FPF_X(0) = 1$

est 0, le seuil associé à $t_1 = FPF_X(1)$ est 1, etc... La courbe ROC_{X,X_D} peut aussi être vue comme l'application définie sur l'ensemble T_X des fractions de faux positifs (inclus dans $[0,1]$) par :

$$\mathrm{ROC}_{X,X_D}(t) = TPF_{X,X_D}\Big(FPF_X^{-1}(t)\Big),$$

pour $t \in T_X$. Autrement dit, $\mathrm{ROC}_{X,X_D}(t)$ est la fraction de vrais positifs (i.e. sensibilité) au seuil $FPF_X^{-1}(t)$, c'est-à-dire pour une fraction de faux positifs valant t (i.e. pour laquelle 1-spécificité vaut t).

Un exemple particulier et classique de courbes ROC sont les courbes dites "binormales". Elles admettent la forme paramétrique suivante : pour une fraction t de faux positifs,

$$\mathrm{ROC}(t) = \Phi(a + b \cdot \Phi^{-1}(t)),$$

où a et b sont deux paramètres réels de localisation et de forme et Φ est la fonction de répartition d'une loi normale centrée réduite, à savoir :

$$\begin{aligned} &\mathbb{R} \to]0,1[\\ &x \mapsto \Phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{]-\infty,x[} \exp(-\frac{\nu^2}{2}) d\nu, \end{aligned}$$

Z désignant une variable de loi centrée réduite.

La fonction Φ^{-1} réciproque de Φ s'appelle la fonction probit. Sur l'échelle du probit, les courbes ROC binormales peuvent être vues comme des droites d'ordonnée à l'origine a et de pente b.

B.2.2 Méthode de modélisation des courbes ROC

Présentation générale du modèle de régression ROC

Afin de comparer les performances diagnostiques des scores associés à la première lecture, à la deuxième lecture et à la lecture avec CADx, les courbes ROC ont été directement modélisées à l'aide d'un modèle de régression en fonction du lecteur et du type de lecture (1, 2 ou avec CADx). Ce modèle permet d'estimer et de comparer *directement* les courbes ROC des trois lectures (lecteur par lecteur et tous lecteurs confondus). L'idée globale sous-jacente de ce modèle est, afin d'estimer la courbe ROC en fonction du lecteur et du type de lecture, de comparer les scores entre unités malades et non-malades pour des mêmes valeurs de covariables (i.e. même lecteur, même lecture).

Une adaptation de la méthode proposée par Alonzo TA et Pepe MS en 2002 [3,89,90] a été utilisée. Le modèle général de régression ROC suivant a été considéré :

$$\mathrm{ROC}_{X,X_D}(t) = \Phi\Big(\alpha_0 + \beta \cdot \Phi^{-1}(t) + \gamma \cdot X + \gamma_D \cdot X_D\Big),$$

où t est une fraction de faux positifs comprise strictement entre 0 et 1, X un vecteur de covariables communes aux unités malades et non-malades (ici, le type de lecture et le lecteur) et X_D est un vecteur de covariables spécifiques aux unités malades. Les paramètres à estimer dans ce modèle sont α_0, β et les vecteurs de paramètres γ et γ_D.

Procédure d'estimation des paramètres

La procédure d'estimation des paramètres de ce modèle de régression ROC proposée par Alonzo TA et Pepe MS est basée sur les variables indicatrices binaires suivantes :

$$U_{i,t,x} = \mathsf{Ind}[Y_{i,x} \geq FPF_x^{-1}(t)]$$

pour une unité malade i, une FPF $t \in T_x$ (l'ensemble des fractions de faux positifs pour le vecteur de covariables x).

Autrement dit, $U_{i,t,x}$ indique si le score de l'unité malade i, étant donné le vecteur de covariables $X = x$, est supérieur ou égal au seuil associé à la fraction de faux positifs t quand $X = x$. Ce sont ces variables indicatrices binaires $U_{i,t,x}$ qui permettent de comparer les scores entre unités malades (à savoir les $Y_{i,x}$) et non-malades (à savoir les seuils $FPF_x^{-1}(t)$ associés à la fraction de faux positifs t) pour des mêmes valeurs de covariables (à savoir $X = x$).

Le point clé de la procédure d'estimation proposée par Alonzo TA et Pepe MS est que la variable binaire $U_{i,t,x}$ suit un modèle de régression binomiale de type probit (cf. section B.3.1). En effet, la probabilité de "*succès*" $P(U_{i,t,X} = 1 \mid X, X_D)$ s'écrit :

$$\begin{aligned} P(U_{i,t,X} = 1 \mid X, X_D) &= P(Y_{i,X} \geq FPF_X^{-1}(t) \mid X, XD) \\ &= \mathrm{ROC}_{X,X_D}(t) \\ &= \Phi\Big(\alpha_0 + \beta \cdot \Phi^{-1}(t) + \gamma \cdot X + \gamma_D \cdot X_D\Big). \end{aligned} \tag{B.3}$$

Elle est donc bien modélisée en fonction des covariables $\Phi^{-1}(t)$, X et X_D. La fonction de lien de ce modèle binomial est un lien probit (Φ^{-1}) et le prédicteur linéaire est $\alpha_0 + \beta \cdot \Phi^{-1}(t) + \gamma \cdot X$. Il est ainsi possible d'utiliser le cadre des régressions binomiales pour obtenir les estimations des paramètres α et β.

La procédure d'estimation se déroule de la manière suivante :

Etape 1 : on estime les fractions de faux positifs $FPF_x(y)$ pour $y \in \{1, 2, 3, 4\}$ et pour chaque vecteur de covariables $X = x$. Pour $X = x$, on dispose ainsi de l'ensemble des fractions de faux positifs $\hat{T}_x$.

Etape 2 : pour chaque vecteur de covariables $X = x$, pour chaque unité malade i et pour chaque fraction de faux positifs $\hat{t} \in \hat{T}_x$, on calcule $\hat{U}_{i,\hat{t},x} = \mathsf{Ind}[Y_{i,x} \geq FPF_x^{-1}(\hat{t})]$ et $\Phi^{-1}(\hat{t})$.

Etape 3 : on ajuste le modèle de régression binomiale de type probit donné en (B.3) sur les données :

$$\{(\hat{U}_{i,\hat{t},x}, \Phi^{-1}(\hat{t}), x, x_D) \mid i \text{ est une unité malade},, x \text{ et } x_D \text{ des vecteurs de covariables et } \hat{t} \in \hat{T}_x\}.$$

Il est nécessaire (pour les étapes 1 et 2) de choisir une méthode d'estimation des fractions de faux positifs $FPF_x(y)$.

Rappelons que dans notre modèle, X est le couple (lecteur, type de lecture). Les fractions de faux positifs ont été modélisées séparément pour chaque type de lecture en fonction

du lecteur à l'aide de trois régressions probit ordinales (une régression par type de lecture) :

$$\Phi^{-1}\Big(P\Big(Y_{\overline{D},\text{ lecteur } l}^{(\text{lecture } k)} \leq y\Big)\Big) = \mu_y^{(\text{lecture } k)} - \sum_{r=2}^{12} \lambda_r^{(\text{lecture } k)} \cdot \mathsf{Ind}(\text{lecteur } l).$$

Remarquons que les paramètres $\mu_y^{(\text{lecture } k)}$ et $\lambda_r^{(\text{lecture } k)}$ n'ont pas d'autre intérêt ici que d'estimer $FPF(l,k)(y)$ (pour les étapes 1 et 2) par :

$$\hat{P}\Big(Y_{\overline{D},\text{ lecteur } l}^{(\text{lecture } k)} \leq y\Big) = \Phi\Big(\hat{\mu}_y^{(\text{lecture } k)} - \sum_{r=2}^{12} \hat{\lambda}_r^{(\text{lecture } k)} \cdot \mathsf{Ind}(\text{lecteur } l)\Big).$$

Les estimations $\hat{\mu}_y^{(\text{lecture } k)}$ et $\hat{\lambda}_r^{(\text{lecture } k)}$ des paramètres des trois régressions ordinales ne sont donc pas présentées.

Dans cette annexe, nous nous limitons à une comparaison globale pour tous les lecteurs. Nous considérons donc l'unique covariable "type de lecture".

Détermination des intervalles de confiance

Une méthode de type *bootstrap* a été utilisée pour déterminer les intervalles de confiance des différentes estimations présentées. A l'instar de l'estimation non-paramétrique (cf. section B.1), pour la constitution de chaque jeu de données simulé, 30 pseudo-patients sont tirés aléatoirement avec remise (avec donc possibilité de doublon) et les données de l'ensemble des cibles de ces 30 patients constituent un jeu de données simulé.

Mille jeux de données au total ont été simulés. Sur chaque jeu de données, la procédure d'estimation des paramètres du modèle ROC décrite en B.2.2 a été appliquée pour obtenir mille vecteurs d'estimations des paramètres du modèle de régression ROC. La *Bias-corrected percentile method* a été utilisée pour la détermination des intervalles de confiance à 95% des estimations des paramètres du modèle de régression ROC, mais aussi pour la détermination des IC des estimations des autres quantités déterminées à partir des paramètres du modèle de régression ROC (ordonnées à l'origine et pentes des courbes ROC, AUC et différences d'AUC). Pour tester si une quantité q était significativement différente de 0, des p-valeurs ont été calculées en supposant que la distribution de l'échantillon des mille pseudo-estimations de cette quantité q était normale.

B.2.3 Modélisation des courbes ROC globalement pour l'ensemble des lecteurs

Modèle

Deux analyses ont été effectuées pour la modélisation des courbes ROC globalement pour l'ensemble des lecteurs. et le modèle mixte suivant a été utilisé. Un effet aléatoire "lecteur" sur l'intercept (le paramètre de localisation), la "pente" (le paramètre d'échelle), et le type de lecture (l'écart entre courbes ROC de la deuxième lecture et de l'autre lecture considérée) a été introduit.

$$\Phi^{-1}\Big(\mathrm{ROC}_k(t)\Big) = \alpha_0 + U(l) + (\alpha_k + U_k(l)) \cdot \mathsf{Ind}[k = \text{lecture 1 ou CADx}] + \Big(\beta + \beta_k \cdot \mathsf{Ind}[k = \text{lecture 1 ou CADx}] + V(l)\Big) \cdot \Phi^{-1}(t).$$

où k désigne le type de lecture, et l le lecteur, U, U_k et V correspondent aux effets aléatoires au niveau lecteur, respectivement sur l'intercept, le type de lecture et la "pente" de la courbe ROC. Les paramètres α représentent "la localisation" tandis que les paramètres β représentent "la pente" de la courbe.

Ainsi, selon ce modèle, la courbe ROC lors de la lecture k est donnée par :

$$\mathrm{ROC}_k(t) = \Phi\Big(\alpha_0 + U(l) + (\alpha_k + U_k(l)) \cdot \mathsf{Ind}[k = \text{lecture 1 ou CADx}] + \Big(\beta + \beta_k \cdot \mathsf{Ind}[k = \text{lecture 1 ou CADx}] + V(l)\Big) \cdot \Phi^{-1}(t)\Big).$$

Les paramètres de localisation et d'échelle pour la courbe ROC "globale" de la lecture 2 valent donc d'après ce modèle α_0 et β. Les paramètres de localisation et d'échelle pour la courbe ROC 'globale' de la lecture k (où k est la lecture 1 ou la lecture avec CADx) valent donc d'après ce modèle $\alpha_0 + \alpha_k$ et $\beta + \beta_k$.

Le paramètre α_k représente le décalage entre la courbe ROC de la lecture 2 et la courbe ROC de la lecture k (lecture 1 ou lecture avec CADx). Il permet ainsi de quantifier l'écart entre ces deux courbes ROC.

Résultats

Résultats pour la lecture avec CADx (avec pour référence la lecture 2), globalement pour l'ensemble des lecteurs :

	Estimate (95% IC)	Erreur Standard	p-value
(Intercept)	1.273 [1.012 ;1.568]	0.151	0
quant.cr	0.787 [0.66 ;0.942]	0.072	0
lecture.idCAD	0.172 [-0.018 ;0.405]	0.108	0.112

Résultats pour la lecture 1 (avec pour référence la lecture 2), globalement pour l'ensemble des lecteurs :

	Estimate (95% IC)	Erreur Standard	p-value
(Intercept)	1.263 [1.03 ;1.536]	0.133	0
quant.cr	0.78 [0.657 ;0.906]	0.062	0
lecture.id1	-0.051 [-0.169 ;0.072]	0.062	0.41

Estimation et IC des ordonnées à l'origine et des pente des courbes ROC, lectures 1, 2, CADx (globalement pour l'ensemble des lecteurs) :

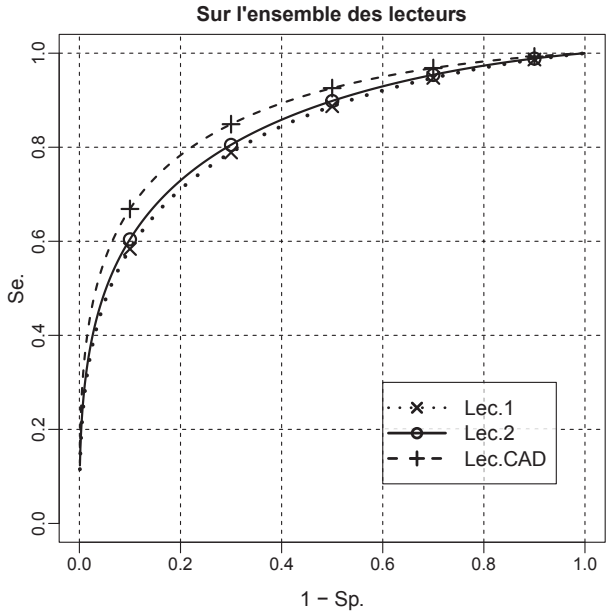


Figure *B.1* – *Graphique des courbes ROC modélisées lors des lectures 1, 2 et CADx, globalement pour l'ensemble des lecteurs*

	Ordonnée origine	Pente
Lecture 1	1.21 [0.99 ;1.47]	0.78 [0.66 ;0.91]
Lecture 2	1.27 [1.01 ;1.57]	0.79 [0.66 ;0.94]
Lecture CADx	1.44 [1.11 ;1.83]	0.79 [0.66 ;0.94]

AUC estimés, avec IC, à partir des courbes ROC modélisées, pour les lectures 1, 2, CADx (globalement pour l'ensemble des lecteurs) :

Lecture 1	Lecture 2	Lecture CADx
83 [77.9 ;88]	84.1 [78.1 ;88.7]	87.2 [81 ;92]

Différences estimées d'AUC, avec IC, à partir des courbes ROC modélisées, entre les lectures 1 vs 2, 1 vs CADx et 2 vs CADx (globalement pour l'ensemble des lecteurs)

Lecture 2 - Lecture 1	Erreur Standard	p-value
1.1 [-5.8 ;8.4]	3.7	0.763

Lecture CADx - Lecture 1	Erreur Standard	p-value
4.2 [-3.4 ;11.5]	3.8	0.2736

Lecture CADx - Lecture 1	Erreur Standard	p-value
3 [0 ;6.6]	1.7	0.0769

B.3 Propension des lecteurs à coder 0 pour des cibles non-malades, et à coder 4 pour des cibles malades

B.3.1 Introduction au modèle de régression linéaire généralisé pour critère binaire : régressions logistique et probit

Pour un exposé complet, on se référera au livre de McCullagh et Nelder [71].

Etant donné un critère de réponse Y binaire (*'succès'* vs *'échec'*), on cherche à étudier la probabilité de *succès* et comment des covariables d'intérêt notées $X = (X_j)_{j=1...k}$ agissent sur cette probabilité de *succès*, ou, autrement dit, quel est l'effet des covariables sur la probabilité de *succès* ?

Les modèles de régression linéaire généralisés permettent de modéliser la probabilité de *succès* en fonction des valeurs des covariables, notée $P(Y = succès|X = x)$. Cette probabilité est à valeurs dans $]0, 1[$. Deux modalités sont à spécifier pour ces modèles linéaires généralisés :

- la fonction de lien : c'est une application bijective bicontinue (de classe $\mathcal{C}^1$ en fait) de $]0\,;1[$ dans $\mathbb{R}$ correspondant à une transformation de l'échelle de la réponse (ici $]0, 1[$) à l'échelle du prédicteur linéaire $\mathbb{R}$. Deux fonctions de lien classiques sont les liens logistique et probit.
- le prédicteur linéaire : la spécification de l'effet des covariables est faite sur l'échelle du prédicteur linéaire, après transformation de la réponse $P(Y|X = x)$ par la fonction de lien. Le prédicteur linéaire est de la forme $\beta_0 + \sum_{j=1}^{k} \beta_j \cdot X_j$.

Fonction de lien logistique La fonction logistique est définie par :

$$\begin{aligned}]0; 1[&\to \mathbb{R} \\ p &\mapsto \log\Big(\frac{p}{1-p}\Big). \end{aligned}$$

La fonction réciproque, appelée anti-logit, est donnée par :

$$\begin{aligned} \mathbb{R} &\to]0; 1[\\ x &\mapsto \frac{\exp x}{1 + \exp x}. \end{aligned}$$

Notons que le rapport $\frac{p}{1-p}$ s'appelle la **cote** (*odds* en anglais). Il correspond au rapport de la probabilité de *succès* sur la probabilité d'*échec*. Il peut être vu comme le nombre de *succès* sur le nombre d'*échecs*. Par exemple, sur 25 réalisations, s'il y a eu 15 *succès* et 10 *échecs* observés, la proportion de *succès* est égale à $15/25 = 60\%$, et la cote correspondante est égal à $15/10 = 1,5$.

Ecriture du modèle logistique
Le modèle logistique s'écrit :

$$\text{logit}(P(Y = succès) \mid X)) = \log\Big(\frac{P(Y = succès) \mid X)}{1 - P(Y = succès) \mid X)}\Big) = \beta_0 + \sum_{j=1}^{k} \beta_j \cdot X_j. \qquad \text{(B.4)}$$

Les paramètres à estimer de ce modèle sont l'intercept β_0 et les $(\beta_j)_{j=1...k}$. On les estime par maximum de vraisemblance.

Une fois ces paramètres estimés par $(\hat{\beta}_j)_{j=1...k}$, pour une observation de vecteur de valeurs de covariables $(x_j)_{j=1...k}$, la probabilité prédite par le modèle décrit en (B.4) pour cette observation est donnée par :

$$\hat{P}(Y = \textit{succès}) \mid X) = \text{anti-logit}\Big(\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j \cdot x_j\Big) = \frac{\exp\Big(\hat{\beta}_0 + \sum\limits_{j=1}^{k} \hat{\beta}_j \cdot x_j\Big)}{1 + \exp\Big(\hat{\beta}_0 + \sum\limits_{j=1}^{k} \hat{\beta}_j \cdot x_j\Big)}.$$

Interprétation des paramètres
L'intercept β_0 correspond au 'niveau' de référence, c'est-à-dire que $\exp(\beta_0)$ correspond à la cote pour le vecteur de covariable $(0)_{j=1...k}$. Ainsi, si $k = 1$ et si X_1 est la fonction indicatrice de sexe masculin, $\exp(\beta_0)$ correspond au rapport de cote chez les femmes.

Pour une variable indicatrice X_j, β_j correspond en fait au logarithme d'un rapport de cotes relativement à deux modalités 0 et 1 de la variable X_j : $\frac{\text{Odds}_{X_j=1}}{\text{Odds}_{X_j=0}}$.

En effet, à partir de (B.4), les valeurs des covariables hormis X_j étant égales, ce rapport de cote vaut :

$$\frac{\text{Odds}_{X_j=1}}{\text{Odds}_{X_j=0}} = \frac{\exp\Big(\beta_0 + \sum\limits_{\substack{j=1...k \\ i\neq j}} \beta_i \cdot X_i + \beta_j \cdot 1\Big)}{\exp\Big(\beta_0 + \sum\limits_{\substack{j=1...k \\ i\neq j}} \beta_i \cdot X_i + \beta_j \cdot 0\Big)} = \exp(\beta_j).$$

Les variables dans le modèle décrit en (B.4) 'agissent' ainsi multiplicativement sur les cotes (donc additivement sur les logarithmes des cotes).

Notons aussi que, d'après ce modèle, l'effet des covariables est supposé *identique* quelles que soient les autres caractéristiques des observations non prises en compte dans le modèle. De plus, conditionnellement à X, les observations des variables de réponse Y sont supposées indépendantes. Ainsi, si plusieurs réalisations de Y correspondent à une même entité (le patient pour l'étude CADx), les hypothèses du modèle sont mises en défaut. Cela a un impact notamment sur les IC des paramètres estimés.

Ecriture du modèle logistique hiérarchique ou à effet mixte

Dans le cas où les observations ont une structure hiérarchique, où plusieurs réalisations j_i ont été observées pour une même entité i, un modèle logistique hiérarchique prenant en compte la structure des données permet de modéliser la probabilité de *succès* et l'effet des covariables sur cette probabilité. Ce modèle hiérarchique permet de prendre en compte la corrélation des réalisations répétées au sein de chaque unité.

Un exemple de modèle logistique hiérarchique peut s'écrire :

$$\text{logit}\Big(P(Y_{i,j_i} = \textit{succès} \mid X(i,j)\Big) = \beta_0 + \sum_{l=1}^{k} \beta_l \cdot X_l(i,j_i) + U(i), \tag{B.5}$$

où $U(i)$ est un effet aléatoire (ici sur l'intercept uniquement). Il est nécessaire dans ce genre de modèle de spécifier la distribution de l'effet aléatoire U. En général, U est supposé suivre une loi normale centrée (i.e. de moyenne nulle) et de variance σ^2. En plus des paramètres $(\beta_l)_{l=0\ldots k}$, σ doit être estimé.

Ce genre de modèle est bien moins contraignant que le modèle sans effet aléatoire. En effet, il autorise les probabilités de *succès* à être variables d'un individu i_0 à un autre individu i_1. Les paramètres $(\beta_l)_{l=0\ldots k}$ des effets dits "fixes" correspondent aux effets marginaux (i.e. en "moyenne" sur l'ensemble de la population étudiée) des covariables X_1. De plus, il est possible d'estimer des probabilités de *succès* marginales en prenant $U = 0$ dans (B.5).

On peut aussi rajouter dans (B.5) des effets aléatoires au niveau d'une covariable en plus d'un effet aléatoire sur l'intercept. Par exemple,

$$\text{logit}\Big(P(Y_{i,j_i} = \textit{succès} \mid X(i,j)\Big) = \beta_0 + \sum_{l=1}^{k} \beta_l \cdot X_l(i,j_i) + U(i) + V(i) \cdot X_1. \tag{B.6}$$

Dans le modèle décrit en (B.6), non seulement les probabilités de *succès* sont variables d'un individu à l'autre (avec l'effet aléatoire U), mais de plus l'effet de X_1 sur la probabilité de *succès* peut varier d'un individu à l'autre. En effet, pour deux individus i_0 et i_1, les logarithmes des rapports de cotes correspondant à la variable X_1 sont respectivement de $\beta_1 + V(i_0)$ et $\beta_1 + V(i_1)$. On remarque que l'effet moyen X_1 est donné par β_1 (en prenant $V = 0$).

B.3.2 Résultats : propension à coder 0 chez les cibles non-malades

On cherche ici à évaluer la propension des lecteurs à coder 0 selon les lectures pour les cibles non cancéreuses (statut.cible= 0). Dans les données, 46 cibles chez 25 patients sont non-cancéreuses. Le tableau ci-dessous renseigne sur la répartition des patients selon leur nombre de cibles non-cancéreuses étudiées. Ainsi, 10 patients ont une unique cible non-cancéreuse étudiée, 10 patients en ont exactement 2 cibles non-cancéreuses,...

Nb. de patients	10	10	4	1
Nb. de cibles par patients	1	2	3	4

Des modèles logistiques hiérarchiques à effet mixte ont été utilisés pour étudier la propension des lecteurs à coder 0 lors des lectures 1, 2 et avec CADx.

La propension $p_{i,j_i,l}$(lecture k) à coder $\{0\}$ ("le *succès*") vs $\{1,2,3,4\}$ (" l'*échec* ") pour la cible j_i, le lecteur l et la lecture k ($k \in \{1,2,$ ou CADx$\}$) est modélisée par :

$$\begin{aligned}\log\Big(\frac{p_{i,j_i,l}(\text{lecture } k)}{1-p_{i,j_i,l}(\text{lecture } k)}\Big) =& \beta_0 + T_i + U(j_i) + V(i) + \beta_{\text{junior}} \cdot \mathsf{Ind}(l : \text{lecteur junior ?}) \\ &+ (V_{l,\text{lecture 1}} + \beta_{\text{lecture 1}} \cdot \mathsf{Ind}(\text{lecture } 1(j_i,l)) \\ &+ (V_{l,\text{lecture CADx}} + \beta_{\text{lecture CADx}} \cdot \mathsf{Ind}(\text{lecture CADx}(j_i,l)).\end{aligned} \tag{B.7}$$

Ce modèle inclut un effet dit "fixe" :

- l'intercept β_0 (référence : lecteur expérimenté, lors de la lecture 2)
- un effet "fixe" type de lecture :
  → le paramètre $\beta_{\text{lecture 1}}$ correspond au log de l'odds-ratio lecture 1 vs lecture 2 (la référence) qui permettra de déterminer si, sur l'ensemble des lecteurs, la propension à coder 0 est différente entre la lecture 1 et la lecture 2 ;
  → le paramètre $\beta_{\text{lecture CADx}}$ correspond au log de l'odds-ratio lecture CADx vs lecture 2 (la référence) qui permettra de déterminer si, sur l'ensemble des lecteurs, la propension à coder 0 est différente entre lecture avec CADx et lecture 2 ;
  → le paramètre β_{junior} permet de prendre en compte le fait que la propension à coder 0 est globalement différente entre lecteurs juniors et lecteurs seniors.

Ce modèle inclut aussi un effet dit "aléatoire" :

- un effet aléatoire patient T_i sur l'intercept uniquement. Cet effet aléatoire T_i permet de prendre en compte le fait que plusieurs cibles proviennent d'un même patient ;
- un effet aléatoire cible $U(j_i)$ sur l'intercept uniquement, qui permet de prendre en compte le fait que plusieurs lectures ont été faites par plusieurs lecteurs sur une même cible. $\beta_0 + U(j_i)$ représente ainsi (le logit d') une propension moyenne à coder 0, tous lecteurs confondus, pour la lecture 2 ;
- un effet aléatoire lecteur au niveau de l'intercept (V_l) et du type de lecture ($V_{l,\text{ lecture 1}}$ et $V_{l,\text{lecture CADx}}$). Ces effets aléatoires permettent de prendre en compte le fait que différentes lectures sur différentes cibles ont été faites par un même lecteur, que, pour toutes les lectures, la propension à coder 0 varie selon les lecteurs (V_l) et enfin que la propension à coder 0 est variable d'une lecture à une autre pour un même lecteur ($V_{l,\text{lecture 1}}$ et $V_{l,\text{lecture CADx}}$).

Résultats du modèle (B.7) (réalisé sous *R*, package *lme4*) :

> summary(t1)

Generalized linear mixed model fit by the Laplace approximation

Formula : lecture.zero ~ (1 | idpat/idc) + (lecture.id | lecteur) + lecture.id + junior

Data : score0

AIC	BIC	logLik	deviance
1192	1257	-584	1168

Random effects :

Groups	Name	Variance	Std.Dev.	Corr	
idc :idpat	(Intercept)	7.41136	2.72238		
idpat	(Intercept)	0.44847	0.66968		
lecteur	(Intercept)	0.88134	0.93880		
	lecture.id1	0.53459	0.73116	-0.261	
	lecture.idCAD	0.15299	0.39115	0.623	0.321

Number of obs : 1656, groups : idc :idpat, 46 ; idpat, 25 ; lecteur, 12

Fixed effects :

	Estimate	Std. Error	z value	Pr($>\|z\|$)	
(Intercept)	-2.0783	0.6269	-3.315	0.000916	***
lecture.id1	**-0.1558**	0.2963	-0.526	**0.598940**	
lecture.idCAD	**1.9491**	0.2326	8.379	$<$ **2e-16**	***
junior1	**-1.3948**	0.5633	-2.476	**0.013287**	*

—

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Logarithme des odds-ratio, odds-ratio : estimation et IC à 95%

	estimate	odds-ratio
(Intercept)	-2.08 [-3.47 ;-0.77]	
lecture.id1	-0.16 [-0.77 ;0.45]	0.86 [0.46 ;1.57]
lecture.idCAD	1.95 [1.43 ;2.43]	7.02 [4.16 ;11.38]
junior1	-1.39 [-2.57 ;-0.01]	0.25 [0.08 ;0.99]

Remarquons que les intervalles de confiance ont été obtenus aussi à l'aide d'une méthode de *bootstrap* (ici du *bootstrap* dit paramétrique). Les p-values et les intervalles de confiance peuvent être légèrement discordants.

D'après ce modèle,
- la propension à coder 0 entre lecture 1 et lecture 2 est similaire avec une très légère tendance à coder moins de 0 lors de la lecture 1 (odds-ratio : $\exp(-0,16) \simeq 0,86$) ;
- la propension à coder 0 est bien plus importante pour la lecture avec CADx par rapport à la lecture 2 (odds-ratio : $\exp(1,95) \simeq 7,02$). Cet odds-ratio est significativement, d'un point de vue "statistique", différent de 1 au seuil de 5% (p-value $<$ 0.0001) ;
- l'estimation de β_{junior} indique une propension inférieure pour les lecteurs juniors à coder 0 (odds-ratio : $\exp(-1,39) \simeq 0,25$). En terme d'ampleur d'effet, les différences de propension à coder 0 sont importantes entre lecteurs juniors et seniors.

On peut noter aussi que la propension à coder 0 est très variable selon la cible, ce qui est indiqué par l'écart-type de l'effet aléatoire cible estimé à environ 2.72).

Probabilités prédites marginale (en "moyenne", exprimées en %) :

	Tous	Junior	Senior
Lecture 2	5.3 [1.7 ;14]	3 [0.8 ;9.6]	11.1 [3 ;31.6]
Lecture 1	4.5 [1.4 ;13]	2.6 [0.6 ;8.7]	9.7 [2.6 ;29.5]
Lecture avec CADx	28.4 [11.5 ;54.8]	17.9 [5.1 ;43.8]	46.8 [16.5 ;78.8]

B.3.3 Résultats : propension à coder 4 chez les cibles malades

On cherche ici à évaluer la propension des lecteurs à coder 4 selon les lectures pour les cibles cancéreuses (statut.cible=1).

Dans les données, 42 cibles chez 26 patients sont cancéreuses. Le tableau ci-dessous indique la répartition des cibles cancéreuses étudiées chez les patients. Ainsi, 12 patients ont une seule cible cancéreuse étudiée, 12 patients en ont 2 et 2 patients en ont 3.

Nb. de patients	12	12	2
Nb. de cibles par patients	1	2	3

Des modèles logistiques hiérarchiques à effet mixte ont été utilisés pour étudier la propension des lecteurs à coder 4 lors des lectures 1, 2 et avec CADx.

La propension $p_{i,j_i,l}$(lecture k) à coder $\{4\}$ vs $\{0,1,2,3\}$ pour la cible cancéreuse j_i du patient i, le lecteur l et la lecture k ($k = 1, 2$ ou CADx) est modélisée par un modèle similaire au modèle décrit en (B.7).

Résultats du modèle (B.7) (réalisé sous R, package *lme4*) : > summary(q1)
Generalized linear mixed model fit by the Laplace approximation
Formula : lecture.quatre ∼ (1 | idpat/idc) + (lecture.id | lecteur) + lecture.id + junior
Data : score4

AIC	BIC	logLik	deviance
1010	1073	-492.8	985.6

Random effects :

Groups	Name	Variance	Std.Dev.	Corr	
idc :idpat	(Intercept)	13.263389	3.64189		
idpat	(Intercept)	0.863732	0.92937		
lecteur	(Intercept)	1.502794	1.22588		
	lecture.id1	0.624309	0.79013	-0.589	
	lecture.idCAD	0.048393	0.21998	-0.976	0.398

Number of obs : 1512, groups : idc :idpat, 42 ; idpat, 26 ; lecteur, 12

Fixed effects :

	Estimate	Std. Error	z value	Pr(>| z |)	
(Intercept)	-1.2102	0.7958	-1.521	0.1283	
lecture.id1	**-0.5473**	0.3131	-1.748	**0.0805**	.
lecture.idCAD	**0.4756**	0.2268	2.097	0.0360	*
junior1	**0.1729**	0.5855	0.295	**0.7678**	

—
Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Logarithme des odds-ratio, odds-ratio : estimation et IC à 95%

	Estimate	odds-ratio
(Intercept)	-1.21 [-2.82 ;0.41]	
lecture.id1	-0.55 [-1.14 ;0.06]	0.58 [0.32 ;1.06]
lecture.idCAD	0.48 [0.05 ;0.89]	1.61 [1.05 ;2.44]
junior1	0.17 [-1.06 ;1.45]	1.19 [0.35 ;4.28]

D'après ce modèle,

- la propension à coder 4 était plus faible lors de la lecture 1 que lors de la lecture 2 (odds-ratio : $\exp(-0,55) \simeq 0,58$) ; cependant, cet effet n'atteint pas la significativité statistique (p-value de 0.08) ;
- la propension à coder 4 est plus importante pour la lecture avec CADx que pour la lecture 2 (odds-ratio : $\exp(0,48) \simeq 1,61$). Cet odds-ratio est supérieur à 1 de façon statistiquement significative (p-value de 0.036) ; - la propension à coder 4 était similaire chez les lecteurs juniors et seniors (odds-ratio : $\exp(0,17) \simeq 1,19$, $p = 0,77$).

Probabilités prédites marginale (en "moyenne", exprimées en %) :

	Tous	Junior	Senior
Lecture 1	16 [4 ;42.3]	17 [3.5 ;44.1]	14.7 [3.4 ;43.9]
Lecture 2	24.8 [6.3 ;55.3]	26.2 [6.2 ;58.6]	23 [5.6 ;60.2]
Lecture avec CADx	34.6 [10.4 ;65.4]	36.3 [10.4 ;68.4]	32.4 [9.1 ;69]

Bibliographie

[1] Cancer de prostate : quelle imagerie en 2012 ? Institut des sciences cognitives, Bron, France, juin 2012. Olivier Rouvière.

[2] Dossier cancer de la prostate. http://www.e-cancer.fr/cancerinfo/les-cancers/cancers-de-la-prostate, Agence nationale sanitaire et scientifique en cancérologie. Mis à jour le 1 février 2010.

[3] T.A. Alonzo and M.S. Pepe. Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, 3(3) :421–32, 2002.

[4] GL. Andriole, ED. Crawford, RL. Grubb 3rd, SS. Buys, D. Chia, and TR. Church. Mortality results from a randomized prostate-cancer screening trial. *N. Engl. J. Med.*, 360(13) :1310–9, 2009.

[5] Y. Artan, M.A. Haider, D.L. Langer, T.H. van der Kwast, A.J. Evans, Y. Yang, M.N. Wernick, J. Trachtenberg, and I.S. Yetik. Prostate Cancer Localization With Multispectral MRI Using Cost-Sensitive Support Vector Machines and Conditional Random Fields. *Image Processing, IEEE Transactions on*, 19(9) :2444–2455, September 2010.

[6] J. O. Barentsz, J. Richenberg, R. Clements, P. Choyke, S. Verma, G. Villeirs, O. Rouvière, V. Logager, and J.J. Fütterer. ESUR prostate MR guidelines 2012. *Eur Radiol.*, 22(4) :746–757, April 2012.

[7] J. O. Barentsz, J. Richenberg, R. Clements, P. Choyke, S. Verma, G. Villeirs, O. Rouvière, V. Logager, and J.J. Fütterer. ESUR prostate MR guidelines 2012. *Eur Radiol.*, 22(4) :746–757, April 2012.

[8] E. Bauvin, L. Remontet, P. Grosclaude, réseau FRANCIM, and CépiDc. Incidence and mortality of prostate cancer in france : trends between 1978 and 2000. *Progrès en Urologie : Journal de l'Association Française d'Urologie et de la Société Française d'Urologie*, 13(6) :1334–9, Décembre 2003.

[9] T. Bayes. An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 53 :370–418, 1763.

[10] H. Benoit-Cattin, G. Collewet, B. Belaroussi, H. Saint-Jalmes, and C. Odet. The SIMRI project : a versatile and interactive MRI simulator. *J Magn Reson.*, 173(1), 2005.

[11] D. Le Bihan, E. Breton, D. Lallemand, M.L. Aubin, J. Vignaud, and M. Laval-Jeantet. Separation of Diffusion and Perfusion in Intravoxel Incoherent Motion MR Imaging. *Radiology*, 168(2) :497–505, 1988.

[12] S. Brunelle, M. Marcy, and A. Ruocco. Confrontations anatomo-radiologiques entre IRM prostatiques et pièces de prostatectomie. In *Journées Françaises de Radiologie*, Octobre 2009.

[13] C. Burckhardt. Speckle in Ultrasound B-Mode Scans. *IEEE Trans., Sonics and Ultrasonics*, 25 :1–6, 1978.

[14] L. Calmels. Imagerie Fonctionnelle et Microscopique du petit animal orientée vers la cancérologie, 2010.

[15] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy. SVM and Kernel Methods Matlab Toolbox. Perception, Systèmes et Information, INSA de Rouen, Rouen, France, 2005.

[16] R. Caruana and A. Niculescu-Mizil. Data Mining in Metric Space : an empirical analysis of supervised learning performances criteria. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD '04)*, 2004.

[17] HP. Chan, K. Doi, S. Galhotra, CJ. Vyborny, H. MacMahon, and PM. Jokich. Image feature analysis and computer-aided diagnosis in digital radiography. 1. automated detection of microcalcifications in mammography. *Medical physics*, 14 :538–548, 1987.

[18] I. Chan, W. Wells 3rd, R. V. Mulkern, S. Haker, J. Zhang, K. H. Zou, S. E. Maier, and C. M. Tempany. Detection of Prostate Cancer by Integration of Line-scan Diffusion, T2-mapping and T2-weighted Magnetic Resonance Imaging ; A Multichannel Statistical Classifier. *Medical Physics*, 30(9) :2390–2398, September 2003.

[19] A. Cheikh, N. Girouin, M. Colombel, J.M. Marechal, A. Gelet, A. Bissery, M. Rabilloud, D. Lyonnet, and O. Rouvière. Evaluation of T2-weighted and dynamic contrast-enhanced MRI in localizing prostate cancer before repeat biopsy. *European Radiology*, 19 :770–778, 2009.

[20] D.A. Clausi. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Can. J. Remote Sensing,*, 28(1) :45–62, 2002.

[21] F.V. Coakley and H. Hricak. Radiologic anatomy of the prostate gland : a clinical approach. *Radiol. Clin. North Am.*, 38 :15–30, 2000.

[22] F. Cornud, X. Rebillard, A. Villers, M. Peyromaure, and M. Soulié. Place of contrast imaging in prostate cancer detection. *Sous-comité de Prostate du CCAFU, Prog Urol*, 16 :275–80, 2006.

[23] C. Cortes and V.N. Vapnik. Support-vector networks. *Machine Learning*, 20(3) :273–297, 1995.

[24] R. Cuignet. Contributions à l'apprentissage automatique pour l'analyse d'images cérébrales anatomiques, 2011. Rapport de thèse.

[25] A. V. D'Amico, R. Whittington, S. B. Malkowicz, D. Schultz, J. Fondurulia, M.H. Chen, J. E. Tomaszewski, A. A. Renshaw, A. Wein, and J.P. Richie. Clinical Utility of the Percentage of Positive Prostate Biopsies in Defining Biochemical Outcome After Radical Prostatectomy for Patients With Clinically Localized Prostate Cancer. *J Clin Oncol.*, 18(6) :1164–1172, March 2000.

[26] B.V. Dasarathy. *Nearest Neighbor (NN) Norms : NN Pattern Classification Techniques.* IEEE Computer Society, 1991.

[27] L. Dickinson, H.U. Ahmed, C. Allen, J.O. Barentsz, B. Carey, J.J. Fütterer, S.W. Heijmink, P.J. Hoskin, A. Kirkham, A.R. Padhani, R. Persad, P. Puech, S. Punwani, A. S. Sohaib, B. Tombal, A. Villers, J. van der Meulen, and M. Emberton. Magnetic Resonance Imaging for the Detection, Localisation, and Characterisation of Prostate Cancer : Recommendations from a European Consensus Meeting. *European Urology*, 59(4) :477–494, 2011.

[28] K. Doi. Computer-aided diagnosis in medical imaging : Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31 :198–211, 2007.

[29] M. Dundar, G. Funga, L. Bogonia, M. Macarib, A. Megibowb, and B. Raoa. A methodology for training and validating a CAD system and potential pitfalls. In *CARS - Computer Assisted Radiology and Surgery*, volume 1268, 2004.

[30] M. Dundar, J. Theiler, and S. Perkins. Incorporating spatial contiguity into the design of a support vector machine classifier. 2006.

[31] P.J. Effert, R. Bares, S. Handt, J.M. Wolff, U. Bull, and G. Jakse. Metabolic Imaging of Untreated Prostate Cancer by Positron Emission Tomography with sup 18 Fluorine-Labeled Deoxyglucose. *J. Urol.*, 155 :994–998, 1996.

[32] B. Efron. *The jacknife, the bootstrap, and the other resampling plans.* Society for Industrial and Applied Mathematics, Philadelphia, 1982.

[33] B. Efron and R.J. Tibshirani. Introduction to the bootstrap. *Chapman & Hall/CRC, Boca Raton*, 1993.

[34] M.R. Engelbrecht, G.J. Jager, R.J. Laheij, A.L. Verbeek, H.J. van Lier, and J.O. Barentsz. Local staging of prostate cancer using magnetic resonance imaging : a meta-analysis. *Eur. Radiol.*, 12 :2294, 2002.

[35] O. Faust, U. Acharya, and T. Tamura. Formal design methods for reliable computer aided diagnosis : A review. *Biomedical Engineering, IEEE Reviews in*, 99 :1, 2012.

[36] R.A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2) :179–188, 1936.

[37] F. Fraioli, G. Serra, and R. Passariello. CAD (computed-aided detection) and CADx (computer aided diagnosis) systems in identifying and characterising lung nodules on chest CT : Overview of research, developments and new prospects. *La radiologia medica*, 115(3) :385–402, 2010.

[38] J.J. Fütterer, M.R. Engelbrecht, H.J. Huisman, G.J. Jager, C.A. Hulsbergenvan De Kaa, J.A. Witjes, and J.O. Barentsz. Staging Prostate Cancer with Dynamic Contrast-enhanced Endorectal MR Imaging prior to Radical Prostatectomy : Experienced versus Less Experienced Readers. *Genitourinary Imaging*, 235 :541–549, 2005.

[39] K. Fukunaga and R.R. Hayes. Estimation of classifier performance. *IEEE transactions on pattern analysis and machine intelligence*, 11(10) :1087–1101, 1989.

[40] K. Fukunaga and L. Hostetler. k-nearest-neighbor Bayes risk estimation. *IEEE Trans. Information Theory*, 21(3) :285–293, 1975.

[41] M.L. Giger, K. Doi, and H. MacMahon. Image feature analysis and computer-aided diagnosis in digital radiography. 3. automated detection of nodules in peripheral lung fields. *Med Phys*, 15 :158–166, 1988.

[42] N. Girouin, F. Mège-Lechevallier, A. T Senes, A. Bissery, M. Rabilloud, J. M Maréchal, M. Colombel, D. Lyonnet, and O. Rouvière. Prostate dynamic contrast-enhanced mri with simple visual diagnostic criteria : is it reasonable ? *European Radiology*, 17 :1498–1509, 2007.

[43] D.F. Gleason. The Veteran's Administration Cooperative Urologic Research Group : histologic grading and clinical staging of prostatic carcinoma. *Tannenbaum M., editor. Urologic Pathology : The Prostate Lea and Febiger*, pages 171–198, 1977.

[44] G.P. Haas, N. Delongchamps, O.W. Brawley, C.Y. Wang, and G. de la Roza. The worldwide epidemiology of prostate cancer : perspectives from autopsy studies. *The Canadian Journal of Urology*, 15(1) :3866–3871, February 2008.

[45] M. A Haider, T.H. van der Kwast, J. Tanguay, A.J. Evans, A.T. Hashmi, G. Lockwood, and J. Trachtenberg. Combined T2-Weighted and Diffusion-Weighted MRI for localization of prostate cancer. *American Journal of Roentgenology*, 189(2) :323–328, August 2007.

[46] J.A. Hanley and B.J. McNeil. The meaning and use of the areau under a receiving operating characteristic (ROC) curve. *Radiology*, 143 :29–36, 1982.

[47] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on systems, man and cybernetics*, SMC-3(6) :610–621, November 1973.

[48] S. Haykin. *Neural networks : a comprehensive foundation*. Prentice hall, 1999.

[49] A. Heidenreich. Consensus criteria for the use of magnetic resonance imaging in the diagnosis and staging of prostate cancer : not ready for routine use. *Eur. Urol.*, 59(4) :495–7, 2011.

[50] H. Hotelling. The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2 :360–378, 1931.

[51] S. Ikonen, P. Kärkäinen, L. Kivisaari, J.O. Salo, K. Taari, T. Vehmas, P. Tervahartiala, and S. Rannikko. Endorectal magnetic resonance imaging of prostatic cancer : comparison between fat-suppressed t2-weighted fast spin echo and three-dimensional dual-echo, steady-state sequences. *Eur Radiol*, 11(2) :236–41, 2001.

[52] INSERM. Dossiers d'information : Cancer de la prostate. Technical report, Mis en ligne en avril 2012.

[53] G.J. Jager, E.T. Ruijter, C.A. van de Kaa, J.J. de la Rosette, G.O. Oosterhof, J.R. Thornbury, and J.O. Barentsz. Local staging of prostate cancer with endorectal MR imaging : correlation with histopathology. *AJR Am J Roentgenol.*, 166(4) :845–852, April 1996.

[54] M.W. Kattan, M.J. Zelefsky, P.A. Kupelian, P.T. Scardino, Z. Fuks, and S.A. Leibel. Pretreatment nomogram for predicting the outcome of three-dimensional conformal radiotherapy in prostate cancer. *J Clin Oncol.*, 18(19) :3352–3359, October 2000.

[55] A.P. Kirkham, M. Emberton, and C. Allen. How good is MRI at detecting and characterising cancer within the prostate ? *European Urology*, 50(6) :1163–1174, 2006.

[56] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1) :59–69, 1982.

[57] P. Kozlowski, S.D. Chang, E.C. Jones, K.W. Berean, H. Chen, and S.L. Goldenberg. Combined diffusion-weighted and dynamic contrast-enhanced MRI for prostate cancer diagnosis. Correlation with biopsy and histopathology. *Journal of Magnetic Resonance Imaging*, 24(1) :108–113, July 2006.

[58] G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5 :27–72, 2004.

[59] D.L. Langer, T.H. van der Kwast, A.J. Evans, J. Trachtenberg, B.C. Wilson, and M.A. Haider. Prostate cancer detection with multiparametric MRI : Logistic regres-

sion analysis of quantitative T2, diffusion-weighted imaging, and dynamic contrast-enhanced MRI. *Journal of Magnetic Resonance Imaging*, 30(2) :327–334, August 2009.

[60] Q. Li and K. Doi. Comparison of typical evaluation methods for computer-aided diagnostic schemes : Monte carlo simulation study. *Med. Phys.*, 34(3) :871–876, 2007.

[61] H.K. Lim, J.K. Kim, K.A. Kim, and K.S. Cho. Prostate cancer : apparent diffusion coefficient map with T2-weighted images for detection – a multireader study. *Radiology*, 250(1) :145–151, January 2009.

[62] H.T. Lin, C.J. Lin, and R.C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3) :267–276, 2007.

[63] X. Liu, D.L. Langer, M.A. Haider, Y. Yang, M.N. Wernick, and I.S. Yetik. Prostate Cancer Segmentation With Simultaneous Estimation of Markov Random Field Parameters and Class. *IEEE Transactions on Medical Imaging*, 28(6) :906–915, 2009.

[64] E.R. De Long, M. DeLong, and D.L. Clarke Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves : a non parametric approach. *Biometrics*, 44 :837–45, 1988.

[65] R. Lopes, A. Ayache, N. Makni, P. Puech, A. Villers, S. Mordon, and N. Betrouni. Prostate cancer characterization on MR images using fractal features. *Medical Physics*, 38(1) :83–95, 2011.

[66] J.B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations . In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press*, volume 1, 1967.

[67] A. Madabhushi, M.D. Feldman, D.N. Metaxas, J. Tomaszeweski, and D. Chute. Automated detection of prostatic adenocarcinoma from high-resolution ex vivo MRI. *Medical Imaging, IEEE Transactions on*, 24(12) :1611 –1625, December 2005.

[68] A. Madabhushi, J. Shi, M. Rosen, J.E. Tomaszewski, and M.D. Feldman. Comparing Classification Performance of Feature Ensembles : Detecting Prostate Cancer from High Resolution MRI. In *Computer Vision Methods in Medical Image Analysis (In conjunction with ECCV)*, volume 4241, pages 25–36. Springer Verlag, 2006.

[69] N. Makni. Méthodes d'identification, d'aide au diagnostic et de planification utilisant de l'imagerie multi-modalité pour les thérapies focales du cancer de la prostate., 2010. Rapport de thèse.

[70] BFJ. Manly. *Randomization, Bootsrap and Monte Carlo Methods in Biology*. Chapman and Hall, London, 1997.

[71] P. McCullagh and J.A. Nelder. Igeneralized linear models. *Chapman & Hall/CRC, Second Edition*, 1999.

[72] C.E. Metz. ROC methodology in radiologic imaging. *Investigative Radiology*, 21(9) :720, 1986.

[73] C.E. Metz and H.B. Kronman. Statiscal significance tests for binormal ROC curves. *J. Math. Psychol.*, 22 :218–243, 1980.

[74] BM. Mian, Y. Naya, K. Okihara, F. Vakar-Lopez, P. Troncoso, and RJ. Babaian. Predictors of cancer in repeat extended multisite prostate biopsy in men with previous negative extended multisite biopsy. *Urology*, 60(5) :836–40, 2002.

[75] S.S. Mohamed, M.M.A. Salama, M. Kamek, and K. Rizkalla. Region of interest based prostate tissue characterization using least square support vector machine LS-SVM. *Image analysis and recognition, Lecture Notes in Computer Science*, 3212 :51–58, 2004.

[76] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[77] M. Niemeijer, M. Loog, M.D. Abramoff, M.A. Viergever, M. Prokop, and B. van Ginneken. On Combining Computer-Aided Detection Systems. *IEEE Transactions on Medical Imaging*, 30(2) :215 –223, February 2011.

[78] D. Lyonnet O. Rouvière, R.P. Hartman. Prostate MR imaging at high-field strength : evolution or revolution ? *Eur Radiol.*, 16(2) :276–284, 2006.

[79] D. Lyonnet O. Rouvière, T. Vitry. Imaging of prostate cancer local recurrences : why and how ? *Eur Radiol.*, 20(5) :1254–66, 2010.

[80] N. Obuchowski. Nonparametric analysis of clustered roc curve data. *Biometrics*, 53 :567–578, 1997.

[81] N. Obuchowski. New methodological tools for multiple-reader ROC studies. *Radiology*, 243 :10–12, 2007.

[82] I. Ocak, M. Bernardo, G. Metzger, T. Barrett, P. Pinto, P.S. Albert, and P.L. Choyke. Dynamic Contrast-Enhanced MRI of Prostate Cancer at 3 T : A Study of Pharmacokinetic Parameters. *American Journal of Roentgenology*, 189(4) :W192–W201, October 2007.

[83] S. Ozer, D.L. Langer, X. Liu, M.A. Haider, T.H. van der Kwast, A.J. Evans, M.N. Wernick Y. Yang, and I.S. Yetik. Supervised and unsupervised methods for prostate cancer segmentation with multispectral MRI. *Medical Physics*, 37(4) :1873, 2010.

[84] A.R. Padhani and J.E. Husband. Dynamic contrast-enhanced MRI studies in oncology with an emphasis on quantification, validation and human studies. *Clin Radiol*, 56(8) :607–620, 2001.

[85] A.R. Padhani, G. Liu, D.M. Koh, T.L. Chenevert, H.C. Thoeny, T. Takahara, A. Dzik-Jurasz, B.D. Ross, M. Van Cauteren, D. Collins, D.A. Hammoud, G.J. Rustin, B. Taouli, and P.L. Choyke. Diffusion-weighted magnetic resonance imaging as a cancer biomarker : consensus and recommendations. *Neoplasia*, 11(2) :102–125, 2009.

[86] D.S. Paik, C.F. Beaulieu, G.D. Rubin, B. Acar, R.B. Jr Jeffrey, J. Yee, J. Dey, and S. Napel. Surface normal overlap : a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT. *IEEE Transactions on Medical Imaging*, 23(6) :661–675, 2004.

[87] AW. Partin, LA. Mangold, DM. Lamm, PC. Walsh, JI. Epstein, and JD. Pearson. Contemporary update of prostate cancer staging nomograms (partin tables) for the new millennium. *urology*, 58 :843–848, 2001.

[88] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information : criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 :1226–1238, 2005.

[89] M.S. Pepe. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 52(2) :352–9, 2000.

[90] M.S. Pepe. The statistical evaluation of medical tests for classification and prediction. *Oxford University Press*, 2003.

[91] M. Perrone. Improving regression estimation : averaging methods for variance reduction with extensions to general convex measure optimization, 1993.

[92] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999.

[93] P. Puech, N. Betrouni, N. Makni, A.S. Dewalle, A. Villers, and L. Lemaitre. Computer-assisted diagnosis of prostate cancer using DCE-MRI data : design, implementation and preliminary results. *International Journal of Computer Assisted Radiology and Surgery*, 4 :1–10, 2009.

[94] P. Puech, O. Rouvière, and F. Cornud. IRM fonctionnelle du cancer de la prostate. In *Journées Françaises de Radiologie*, volume 12, pages 157–172. Springer Verlag, 2010.

[95] S.F. Quinn, D.A. Franzini, T.A. Demlow, D.R. Rosencrantz, J. Kim, R.M. Hanna, and J. Szumowski. MR imaging of prostate cancer with an endorectal surface coil technique : correlation with whole-mount specimens. *Radiology*, 190(2) :323–327, February 1994.

[96] R.M. Rangayyan, F.J. Ayres, and J.E.L. Desautels. A review of computer-aided diagnosis of breast cancer : Toward the detection of subtle signs. *Journal of the Franklin Institute*, 344(2-3) :312–348, 2007.

[97] A. B. Rosenkrantz, X. Kong, B. E. Niver, D. S. Berkman, J. Melamed, J. S. Babb, and S. S. Taneja. Prostate cancer : Comparison of tumor visibility on trace diffusion-weighted images and the apparent diffusion coefficient map. *Genitourinary Imaging*, 196(1) :123–129, Jan 2011.

[98] O. Rouvière, A. Raudrant, R. Ecochard, C. Colin-Pangaud, C. Pasquiou, R. Bouvier, J.M. Maréchal, and D. Lyonnet. Characterization of time-enhancement curves of benign and malignant prostate tissue at dynamic MR imaging. *European Radiology*, 13(5) :931–942, May 2003.

[99] O. Rouvière. MR assessment of recurrent prostate cancer after radiation therapy. *Radiology*, 242(2) :635–6, 2007.

[100] O. Rouvière, A. Gelet, S. Crouzet, and J.Y. Chapelon. Prostate focused ultrasound focal therapy-imaging for the future. *Nat Rev Clin Oncol., In press*, 2012.

[101] O. Rouvière, N. Girouin, L. Glas, A. Ben Cheikh, A. Gelet, F. Mège-Lechevallier, M. Rabilloud, J.-Y. Chapelon, and D. Lyonnet. Prostate cancer transrectal HIFU ablation : detection of local recurrences using T2-weighted and dynamic contrast-enhanced MRI. *European Radiology*, 20(1) :48–55, 2010.

[102] O. Rouvière, R.P. Hartman, and D. Lyonnet. Prostate MR imaging at high-field strength : evolution or revolution ? *European Radiology*, 16(2) :276–284, February 2006.

[103] O. Rouvière, M. Papillard, N. Girouin, R. Boutier, M. Rabilloud, B. Riche, F. Mège-Lechevallier, M. Colombel, and A. Gelet. Is it possible to model the risk of malignancy of focal abnormalities found at prostate multiparametric MRI ? *European Radiology*, 22(5) :1149–1157, 2012.

[104] S. Rüping. A Simple Method for Estimating Conditional Probabilities for SVMs. Technical Report / Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen, 2004.

[105] S. Rueping. SVM classifier estimation from group probabilities. 2010.

[106] C.M. Rutter. Bootstrap estimation of diagnostic accuracy with patient-clustered data. *Academic Radiology*, 7(6) :413–419, 2000.

[107] B. Sahiner, H.P. Chan, and L. Hadjiiski. Classifier performance prediction for computer-aided diagnosis using a limited dataset. *Med. Phys.*, 35(4) :1559–1570, 2008.

[108] B. Sahiner, H.P. Chan, N. Petrick, R.F. Wagnerand, and L. Hadjiiski. Feature selection and classifier performance in computer-aided diagnosis : The effect of finite sample size. *Med. Phys.*, 27(7) :1509–1522, 2000.

[109] B. Schölkopf and A.J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. The MIT Press, 2002.

[110] FH. Schröder, J. Hugosson, MJ. Roobol, TLJ. Tammela, S. Ciatto, and V. Nelen. Prostate-cancer mortality at 11 years of follow-up. *N. Engl. J. Med.*, 366(11) :981–90, 2012.

[111] B. Senthilkumar and G. Umamaheswari. A Review on Computer Aided Detection and Diagnosis - Towards the Treatment of Breast Cancer. *European Journal of Scientific Research*, 52(4) :417–452, 2011.

[112] AK. Singh, J. Krueckerand, S. Xu, N. Glossop, P. Guion, and K. Ullman. Initial clinical experience with real-time transrectal ultrasonography-magnetic resonance imaging fusion-guided prostate biopsy. *BJU Int.*, 101(7) :841–5, 2008.

[113] L.-K. Soh. Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurrence Matrices. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2) :779–795, 1999.

[114] P. Sollich. Bayesian methods for support vector machines : Evidence and predictive class probabilities. *Machine Learning*, pages 46–21, 2002.

[115] S.S. Sonnad, C.P. Langlotz, and J.S. Schwartz. Accuracy of MR imaging for staging prostatecancer : a meta-analysis to examine the effect of technologic change. *Acad. Radiol.*, 8 :149, 2001.

[116] R. Souissi, P. Robert, and J.-S. Raynaud. Caractérisation de la microcirculation tumorale en IRM par imagerie paramétrique. In *SETIT 2007, 4th International Conference : Sciences of Electronic, Technologies of Information and Telecommunications*, pages 1–6, 2007.

[117] T.A. Stamey, N. Yang, A.R. Hay, and J.E. Mc Neal. Prostate specific antigen as a serum marker for adenocarcinoma of the prostate. *New Engl J Med*, 317(15) :909–916, October 1987.

[118] G. Stempfel and L. Ralaivola. Learning SVMs from Sloppily Labeled Data. *Artificial Neural NetworksŰICANN*, pages 884–893, 2009.

[119] C. Stephan, S. Wesseling, T. Schink, and K. Jung. Comparison of Eight Computer Programs for Receiver-Operating Characteristic Analysis. *Clinical Chemistry*, 49 :433–439, 2003.

[120] G.C. Sutton. Computer-aided diagnosis : A review. *British Journal of Surgery*, 76(1) :82–85, 2005.

[121] J.A. Swets and R.M. Pickett. *Evaluation of diagnostic systems : methods from signal detection theory*. Academic Press New York, 1982.

[122] C.H. Tan, J. Wang, and V. Kundra. Diffusion weighted imaging in prostate cancer. *Eur Radiol.*, 21(3) :593–603, 2011.

[123] A. Tanimoto, J. Nakashima, H. Kohno, H. Shinmoto, and S. Kuribayashi. Prostate cancer screening : the clinical value of diffusion-weighted imaging and dynamic MR imaging in combination with T2-weighted imaging. *J Magn Reson Imaging.*, 25(1) :146–52, 2007.

[124] P. Tiwari, M. Rosen, and A. Madabushi. A hierarchical spectral clustering and nonlinear dimensionality reduction scheme for detection of prostate cancer from magnetic resonance spectroscopy. *Med. Phys.*, 36(9) :3927–3939, sept 2009.

[125] P.S. Tofts, G. Brix, D.L. Buckley, J.L. Evelhoch, E. Henderson, M.V. Knopp, H.B.W. Larsson, T.-Y. Lee, N.A. Mayr, G.J.M. Parker, R.E. Port, J. Taylor, and R.M. Weisskoff. Estimating Kinetic Parameters from Dynamic Contrast-Enhanced T1-Weighted MRI of a Diffusible Tracer : Standardized Quantities and Symbols. *J. Magn. Reson. Imaging*, 10 :223–232, 1999.

[126] B. Tombal. Over- and underdiagnosis of prostate cancer : The dangers. *European Urology Supplements*, 5(6) :511–513, Apr 2006.

[127] D. Tuia, R. Flamary, M. Volpi, M. Della Mura, and A. Rakotomamonjy. Discovering relevant spatial filterbanks for VHR image classification. 2012.

[128] V. N Vapnik. *Statistical Learning Theory*. Wiley-Interscience, First edition, September 1998.

[129] S. Verma, B. Turkbey, N. Muradyan, A. Rajesh, F. Cornud, M.A. Haider, P.L. Choyke, and M. Harisinghani. Overview of dynamic contrast-enhanced MRI in prostate cancer diagnosis and management. *AJR Am J Roentgenol*, 198(6) :1277–88, 2012.

[130] S. Viswanath, B. Bloch, E. Genega, N. Rofsky, R. Lenkinski, J. Chappelow, R. Toth, and A. Madabhushi. A Comprehensive Segmentation, Registration, and Cancer Detection Scheme on 3 Tesla In Vivo Prostate DCE-MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 5241, pages 662–669. 2008.

[131] S. Viswanath, B.N. Bloch, M. Rosen, J. Chappelow, R. Toth, N. Rofsky, R. Lenkinski, E. Genega, A. Kalyanpur, and A. Madabhushi. Integrating structural and functional imaging for computer assisted detection of prostate cancer on multi-protocol. In *SPIE Medical Imaging*, volume 7260, pages 72603I–72603I–12, Miami, Florida, 2009.

[132] P.C. Vos, T. Hambrock, J.O. Barentsz, and H.J. Huisman. Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MRI. *Physics in Medicine and Biology*, 55(6) :1719, 2010.

[133] P.C. Vos, T. Hambrock, C.A. Hulsbergen van de Kaa, J.J. Fütterer, J.O. Barentsz, and H.J. Huisman. Computerized analysis of prostate lesions in the peripheral zone using dynamic contrast enhanced MRI. *Medical Physics*, 35(3) :888, 2008.

[134] G. Wahba, X. Lin, F. Gao, D. Xiang, R. Klein, and B. Klein. The Bias-Variance Tradeoff and the Randomized GACV. In *Advances in Neural Information Processing System*, volume 11, pages 620–626. MIT press, 1999.

[135] J.H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the american statistical association*, 58(301) :236–244, 1963.

[136] J.C. Weinreb, J.D. Blume, F.V. Coakley, T.M. Wheeler, J.B. Cormack, C.K. Sotto, H. Cho, A. Kawashima, C.M. Tempany-Afdhal, K.J. Macura, M. Rosen, S.R. Gerst, and J. Kurhanewicz. Prostate cancer : sextant localization at MR imaging and MR spectroscopic imaging before prostatectomy–results of ACRIN prospective multi-institutional clinicopathologic study. *Radiology.*, 251(1) :122–133, 2009.

[137] B.L. Welch. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 1/2 :28–35, 1947.

[138] M. Wiart, L. Curiel, A. Gelet, D. Lyonnet, J.Y. Chapelon, and O. Rouvière. Influence of perfusion on high-intensity focused ultrasound prostate ablation : a first-pass MRI study. *Magn Reson Med*, 58(1) :119–127, 2007.

[139] K. Yoshimitsu, K. Kiyoshima, and H. Irie. Usefulness of apparent diffusion coefficient map in diagnosing prostate carcinoma : correlation with stepwise histopathology. *Journal of Magnetic Resonance Imaging*, 27 :132–139, 2008.

[140] T Yoshizako, Wada, T Hayashi, K Uchida, M Sumura, N Uchida, H Kitagaki, and M Igawa. Usefulness of diffusion-weighted imaging and dynamic contrast-enhanced magnetic resonance imaging in the diagnosis of prostate transition-zone cancer. *Acta Radiol*, 49(10) :1207–13, 2008.

[141] K.K. Yu and H. Hricak. Imaging prostate cancer. *Radiol. Clin. North Am.*, 38 :59–85, 2000.

[142] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

[143] Y. Zhan, D. Shen, J. Zeng, L. Sun, G. Fichtinger, J. Moul, and C. Davatzikos. Targeted prostate biopsy using statistical image analysis. *IEEE. Trans. Medical Imaging*, 26(6) :779–788, 2007.

[144] Y. Zhu, S. Williams, and R. Zwiggelaar. Computer technology in detection and staging of prostate carcinoma : A review. *Med.Image Anal.*, 10 :178–199, 2006.

[145] M.H. Zweig and G. Campbell. Receiver operating characteristic (ROC) plots : a fundamental evolution tool in clinical medicine. *Clin Chem*, 39 :561–577, 1993.

TITRE EN FRANÇAIS

Aide au diagnostic du cancer de la prostate par IRM multi-paramétrique : une approche par classification supervisée.

RESUME EN FRANÇAIS

Le cancer de la prostate est la deuxième cause de mortalité chez l'homme en France. L'IRM multi-paramétrique est considérée comme la technique la plus prometteuse pour permettre une cartographie du cancer, ouvrant la voie au traitement focal, alternatif à la prostatectomie radicale. Néanmoins, elle reste difficile à interpréter et est sujette à une forte variabilité inter- et intra-expert, d'où la nécessité de développer des systèmes experts capables d'aider le radiologue dans son diagnostic.

Nous proposons un système original d'aide au diagnostic (CAD) offrant un second avis au radiologue sur des zones suspectes pointées sur l'image. Nous évaluons notre système en nous appuyant sur une base de données clinique de 30 patients, annotées de manière fiable et exhaustive grâce à l'analyse des coupes histologiques obtenues par prostatectomie. Les performances mesurées dans des conditions cliniques auprès de 12 radiologues, sans et avec notre outil, démontrent l'apport significatif de ce CAD sur la qualité du diagnostic, la confiance des radiologues et la variabilité inter-expert.

La création d'une base de corrélations anatomo-radiologiques est une tâche complexe et fastidieuse. Beaucoup d'études n'ont pas d'autre choix que de s'appuyer sur l'analyse subjective d'un radiologue expert, entâchée d'incertitude. Nous proposons un nouveau schéma de classification, basé sur l'algorithme du séparateur à vaste marge (SVM), capable d'intégrer, dans la fonction d'apprentissage, l'incertitude sur l'appartenance à une classe (ex. sain/malin) de certains échantillons de la base d'entraînement. Les résultats obtenus, tant sur des exemples simulés que sur notre base de données cliniques, démontrent le potentiel de ce nouvel algorithme, en particulier pour les applications CAD, mais aussi de manière plus générale pour toute application de *machine learning* s'appuyant sur un étiquetage quantitatif des données.

TITRE EN ANGLAIS

Computer-aided diagnosis of prostate cancer using multi-parametric MRI : a supervised learning approach.

RESUME EN ANGLAIS

Prostate cancer is one of the leading cause of death in France. Multi-parametric MRI is considered the most promising technique for cancer visualisation, opening the way to focal treatments as an alternative to prostatectomy. Nevertheless, its interpretation remains difficult and subject to inter- and intra-observer variability, which motivates the development of expert systems to assist radiologists in making their diagnosis. We propose an original computer-aided diagnosis system returning a malignancy score to any suspicious region outlined on MR images, which can be used as a second view by radiologists. The CAD performances are evaluated based on a clinical database of 30 patients, exhaustively and reliably annotated thanks to the histological ground truth obtained via prostatectomy. Finally, we demonstrate the influence of this system in clinical condition based on a ROC analysis involving 12 radiologists, and show a significant increase of diagnostic accuracy, rating confidence and a decrease in inter-expert variability.

Building an anatomo-radiological correlation database is a complex and fastidious task, so that numerous studies base their evaluation analysis on the expertise of one experienced radiologist, which is thus doomed to contain uncertainties. We propose a new classification scheme, based on the support vector machine (SVM) algorithm, which is able to account for uncertain data during the learning step. The results obtained, both on toy examples and on our clinical database, demonstrate the potential of this new approach that can be extended to any machine learning problem relying on a probabilitic labelled dataset.

MOTS-CLES

Imagerie par résonance magnétique multi-paramétrique, Cancer de la prostate, Systèmes d'aide au diagnostic, Apprentissage supervisé, Séparateurs à vaste marge

INTITULE ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE

INSERM, U1032, LabTau, Lyon, F-69003, France,
151, cours Albert Thomas 69424 LYON Cedex 03.
Université de Lyon, CREATIS ; CNRS UMR5220 ; Inserm U1044 ; INSA-Lyon ; Université Lyon 1,
7 Av. Jean Capelle, 69621 VILLLEURBANNE, France.