الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Projet de Fin d'Études

présentée par

ABDELAZIZ Radhia

&

NAB Zakaria

pour l'obtention du diplôme de master en Électronique spécialité Traitement d'Information et Système.

Thème

Système d'aide à la décision pour le diagnostic de la maladie de Parkinson à partir de la voix

Proposé par :

Promoteur : **M. YKHLEF Fayçal**

Co -promoteur : M. YKHLEF Farid

Année Universitaire 2014-2015

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Projet de Fin d'Études

présentée par

ABDELAZIZ Radhia

&

NAB Zakaria

pour l'obtention du diplôme de master en Électronique spécialité Traitement d'Information et Système.

Thème

Système d'aide à la décision pour le diagnostic de la maladie de Parkinson à partir de la voix

Proposé par :

Promoteur : **M. YKHLEF Fayçal**

Co -promoteur : M. YKHLEF Farid

Année Universitaire 2014-2015

Remerciements

*Nous tenons avant tout de remercier le bon **DIEU** qui nous a donné la volonté et le courage pour la réalisation de ce travail.*

*Nous remercions vivement « **M. YKHLEF Fayçal**» notre promoteur pour la précieuse assistance, sa disponibilité et son soutien qu'il nous a accordé tout au long de ce projet.*

*Nous remercions également notre Co-promoteur « **M. YKHLEF Farid »** de l'Université de SAAD DAHLEB de Blida, pour ses compétences, son ouverture d'esprit et sa grande disponibilité.*

Nous remercions la direction du Centre de Développement des Technologies Avancées, C.D.T.A et en particulier les dirigeants de la Division Architecture des Systèmes et Multimédias de nous avoir accueilli et d'avoir mis à notre disposition les conditions favorables pour la réalisation de ce présent travail.

Nous remercions aussi tous les membres de jury, qui ont accepté d'examiner et de juger notre modeste travail.

À tous les professeurs qui m'ont accompagné dans mon cursus et qui ont apporté un plus à ma formation universitaire

Dédicace

À mes chers parents
Aucune dédicace ne saurait être assez éloquente pour exprimer ce que vous méritez pour tous les sacrifices que vous n'avez cessé de me donner depuis ma naissance pour que je suive le bon chemin dans ma vie et mes études.

À mon cher Frère Abdelhadi
Tu étais toujours présent dans les moments les plus délicats de cette vie mystérieuse par tes sacrifices et ton profond attachement qui m'ont permis de réussir.

À mes chères sœurs Hanine et Ahlem
Mes fidèles compagnons, un remerciement particulier et sincère pour tous vos efforts fournis.

À mes chers frères et sœurs
Les mots ne suffisent guère pour exprimer l'attachement, l'amour et l'affection que je porte.

À tous les membres de ma famille
À tous mes amis
À tous ceux qui me sont chers
Veuillez trouver dans ce modeste travail l'expression de mon affection.

Radhia

Dédicace

Je dédie ce travail à mon cher père qui, par ses précieux conseils et son soutien, a su me guider vers le droit chemin et vers la voie de la réussite.

À ma très chère mère qui a sacrifié sa noble existence pour bâtir la mienne, et qui est pour moi le symbole du courage et du sacrifice.

À mes très chères sœurs.

À mon cher frère.

À tous ceux que j'aime.

Zakaria

ملخص:

لضمان تشخيص موثوق للمرض، أعد برنامج تشخيصي يساعد الطبيب لتقليل الأخطاء المحتملة التي يمكن أن تحدث خلال مرحلة التشخيص. بالإضافة إلى ذلك فإن البرنامج التشخيصي يساعد في تحديد نوعية المرض ودرجة الخطورة.

العمل المقدم في هذه الأطروحة يعتمد في تطوير نظام آلي يمكن من تشخيص مرض باركنسون انطلاقا من الموجة الصوتية . يستعمل البرنامج المطور أساليب تعتمد على تحديد مميزات الأصوات وتصنيفها باستخدام طريقة شعاع الدعم الآلي.وعليه يمكن تحديد و اتخاذ القرار النهائي ما إذا كان هو صوت صحي أو مصاب بالباركنسون.لقد استخدمنا قاعدة بيانات " باركنسون " للتأكد من صحة البرنامج.

مقاييس الأداء المستخدمة في عملنا : منحىROC ، ارتباك المصفوفة وخطأ التقدير.

كلمات المفاتيح: تصنيف، باركنسون ،SVM ، تصفية

Résumé :

Pour assurer un diagnostic fiable d'une maladie, un logiciel de diagnostic aide le médecin à réduire au minimum les erreurs possibles qui peuvent survenir pendant l'étape de diagnostic. Le logiciel analyse les différentes valeurs des paramètres pris en considération pour le dépistage de cette maladie.

Le travail que nous présentons dans ce mémoire consiste à développer un système automatique qui permet le diagnostic de la maladie de Parkinson à partir de l'onde acoustique de la voix. Le système développé utilise plusieurs techniques de sélection d'attributs et classifie les sons de la voix en utilisant la méthode Support Vector Machine (SVM). La décision finale consiste à indiquer s'il s'agit d'une voix saine ou parkinsonienne. Nous avons utilisé la base de données « PARKINSON DATASET » pour la validation de notre système. Les mesures de performances utilisées sont : le taux de classification, la matrice de confusion, la courbe ROC et l'aire sous cette courbe « AUC ».

Mots clés : Classification, SVM, Filter, Parkinson.

Abstract :

To ensure reliable diagnosis of diseases, a software system can help the Doctor to perform this operation by minimizing the possible errors which may occur during the diagnostic phase. The software analyzes the different values of parameters that are considered during the screening phase. The main mission of this final project consists in developing an automatic system that enables the diagnosis of Parkinson's disease from the acoustic wave of voice. The developed system uses several feature selection techniques and classifies the voice sounds using Support Vector Machines (SVMs). The final decision consists in indicating whether it is a healthy or Parkinson voice. We have used the PARKINSON DATASET to validate our system. The following metrics have been used: the classification accuracy, the confusion matrix, the ROC curve and the AUC.

Keywords: Classification, SVM, Filter, Parkinson

Liste des acronymes et abréviations

ACP: Analyse en Composantes Principales

APQ: Amplitudes Perturbatins Quotient

AUC: *Area Under Curve*

CV: Validation croisée

DAT: Transporteur de la Dopamine

DDA: *Difference of Differences of Amplitude*

DDP*: Difference of Differences of Periods*

DFA: *Detrended Fluctuation Analysis*

ECG: Electrocardiogramme

EDI: Environnement de Développement Intégré

EEG: Electroencéphalographie

EGG: Electroglottographie

Fs: *Feature Selection* (Opération de sélection d'attributs)

HNR: *Harmonics-to- Noise Ratio*

IRM: Imagerie par Résonance Magnétique.

IRMf: Imagerie par Résonance Magnétique Fonctionnelle.

LCR: Liquide Céphalo-Rachidien.

LRRK2: Leucine-Rich Repeat Kinase 2

MC: Matrice de confusion

MDVP: *Multidimensional Voice Program*.

MEEI: *Massachusetts Eye and Ear Infirmary*

MEG: Magnétoencéphalographie

MIBG: Méta-Iodo Benzyl Guanidine

MP: Maladie de Parkinson.

mRMR: *Minimum Redundancy and Maximum Relevance*

NF: Nombre d'attributs

NHR: *Noise-to-Harmonics Ratios*

NURR1: Nucléaire Récepteur lié 1 protéine

PPE: *Pitch Period Entropy*

PPQ*:* Points Period Perturbation Quotient

RBF: *Radial Basis Function*

ROC: *Receiver Operating Characteristic*

RAP: *Relative Average Perturbation*

RPDE: *Recurence Period Density Entrepy*

SN: Substantia Nigra

ST13: Suppression des Tumorigénicité 13

SVM: *Support Vector Machines*

TBC: Taux de Bonne Classification

TEMP: Tomographie d'Emission Mono Photonique

TEP: Tomographie par Emission de Positons

TFN: Taux de Faux Négatifs

TFP: Taux de Faux Positifs

TVN: Taux de Vrais Négatifs

TVP: Taux de Vrais Positifs

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

La maladie de Parkinson (MP) est une maladie neurologique chronique affectant le système nerveux central (responsable des troubles essentiellement moteurs). La MP est la deuxième maladie neurodégénérative la plus fréquente après la maladie d'Alzheimer (Charles 2005). Elle touche environ 1% de la population mondiale de plus de 55 ans (Emborg 2004). Elle affecte aussi bien les hommes que les femmes. L'âge de son apparition est de 57 ans, mais il arrive qu'elle débute pendant l'enfance et semble dans ce cas héréditaire (Benikhlef, Bendimerad et Settout 2013).

Cette maladie est complexe et multifactorielle. Néanmoins, les causes exactes restent mal connues. Le diagnostic de la MP repose sur un ensemble d'arguments cliniques, radiologiques et plus récemment biologique mais peut s'appuyer également sur des systèmes automatisés utilisant des outils d'expertise informatiques et mathématiques.

Dans ce cadre, une nouvelle approche de diagnostic se basant sur l'évaluation des troubles de la voix a récemment été mise au point et constitue un élément supplémentaire qui permet d'étayer le diagnostic.

L'évolution de cette maladie diffère d'un malade à l'autre; mais prend généralement quelques années avant de causer des problèmes majeurs ayant besoin à des ressources thérapeutiques dans les formes évoluées. Bien que les médicaments disponibles permettant une réduction significative des symptômes, en particulier dans les premiers stades de la maladie (Singh, Pillay et Choonara 2007), elles ne permettent pas cependant de la guérir.

De nos jours, il n'y a pas encore de traitements pour ralentir ou prévenir la progression de la dégénération des neurones (Gilles 2006). Mais lorsqu'il sera possible de le faire, un diagnostic précoce de la maladie se révèle crucial. Le traitement et l'analyse des images du cerveau sont sans doute la meilleure solution, car elle considère la source de la maladie en détectant le vieillissement et la mort des neurones au fur et à mesure

que la maladie progresse. Toutefois, cette approche a l'inconvénient d'être invasive, coûteuse et inaccessible pour une population assez large.

L'idée d'analyser la voix pour une détection précoce de la maladie se justifie par le fait que les muscles produisant la voix sont affectés par la maladie de Parkinson à un stade précoce. De plus, l'approche est simpliste et non coûteuse, car elle ne nécessite que l'enregistrement de la voix en utilisant des équipements élémentaires.

L'objectif de notre travail est de concevoir un système automatique qui permet la détection précoce de la MP à partir de la voix. Nous utilisons des plusieurs techniques de sélection d'attributs et nous classifions les sons de la voix en utilisant la méthode Support Vector Machine (SVM). La décision finale consiste à indiquer s'il s'agit d'une voix saine ou parkinsonienne. La base de données « PARKINSON DATASET » a été utilisée dans la partie expérimentale pour mesurer : le taux de classification, la matrice de confusion, la courbe ROC et l'aire sous cette courbe « AUC ».

Notre mémoire est divisé en trois chapitres:

Le premier chapitre regroupe des généralités sur la maladie de Parkinson, en particulier les méthodes de diagnostic de la maladie ainsi que les modalités de diagnostic automatique.

Le deuxième chapitre est consacré à l'étude des systèmes de diagnostic de la maladie de Parkinson à partir de la voix.

Le troisième chapitre donne les résultats et interprétations de notre système.

Ce présent document sera terminé par une conclusion générale et des perspectives.

Chapitre 1 Généralité sur la maladie de Parkinson

1.1 Introduction

La maladie de Parkinson (MP) est une affection dégénérative du système nerveux. Au cours de cette maladie, des cellules nerveuses dans certaines régions du cerveau se nécrosent et meurent. Il en résulte des troubles de la coordination des mouvements avec apparition des symptômes typiques de la maladie: ralentissement des mouvements, rigidité musculaire et tremblements. Dans ce chapitre, nous allons aborder succinctement une définition de la MP, son historique et les différents aspects de son diagnostic.

1.2 Définition et historiques

La MP est une maladie dégénérative qui se caractérise par le vieillissement de certains neurones (cellules nerveuses) qui sont situés dans le noyau du système nerveux nommé « locus niger ». Cette maladie doit son nom au Londonien, *James PARKINSON*, qui fut le premier médecin à décrire cette maladie, en 1817, dans un court mémoire intitulé « *An Essay on the Shaking Palsy* » (Khalil 1996). Dans cet ouvrage, docteur *James* décrit minutieusement les symptômes qu'il avait observés chez ses propres malades. Il fut, hélas, incapable de proposer un traitement susceptible d'aider un tant soit peu les malades.

1.3 Méthodes de diagnostic de la maladie de Parkinson

1.3.1 Diagnostic clinique

Le diagnostic clinique de la MP se base sur l'existence d'un certain nombre de symptômes recherchés lors d'un examen neurologique complet (Bernard 2008). On peut les classer en deux grandes catégories :

Symptômes classiques

Ils sont retrouvés dans la plupart des cas; mais leur coexistence n'est pas un impératif (une personne ne cumule pas toujours tous ces symptômes). On parle de la triade caractéristique.

> ***La lenteur :*** La personne marche à petits pas, est gênée dans les gestes courants (boutonner une veste, se raser, écrire…). Son visage devient également moins expressif, ses yeux clignent moins souvent.
>
> ***Le tremblement :*** Le plus souvent unilatéral, au niveau de la main ou du pied, il se manifeste au repos pour disparaître ou s'atténuer lors des mouvements volontaires.
>
> ***La rigidité :*** Il s'agit d'une augmentation du tonus musculaire. Elle peut, par exemple, diminuer le balancement des bras lors de la marche.

Autres symptômes

Ils peuvent être associés ou non aux signes classiques, leur présence permet de conforter le diagnostic de la MP.

> **Un changement d'élocution:** Le débit devient plus rapide et saccadé, la voix devient plus faible.
>
> **La rigidité du cou:** tardif, elle est liée à la raideur musculaire et peut s'étendre aux épaules.
>
> **Une difficulté à avaler :** Apparaissant tardivement, elle entraîne une accumulation de salive dans la bouche.

L'instabilité posturale : Due à un trouble de l'équilibre, elle entraîne une tendance à tomber vers l'avant ou l'arrière.

Le regard : Dans certaines formes de la maladie, il est légèrement fixe du fait de la difficulté de bouger les yeux.

Le syndrome frontal : Il se traduit par une difficulté à organiser les activités.

Les sphincters : Leur léger relâchement entraîne un besoin impérieux d'uriner voire, parfois, une incontinence urinaire.

La posture : Le corps se penche vers l'avant, comme si son centre de gravité avait été abaissé.

La démarche : Elle est lente et se caractérise par des pas plus petits, pouvant s'accélérer brusquement comme pour empêcher une chute vers l'avant.

1.3.2 Diagnostic radiologique

La radiologie permet d'apporter une rapidité au diagnostic de la MP ainsi que la différenciation de la MP des autres syndromes parkinsoniens. Les techniques utilisées sont résumées dans le Tableau 1.1.

Tableau 1. 1: Techniques radiologiques utilisées dans le diagnostic de la MP

Méthodes	Biomarqueurs	Évaluations	Sources
TEMP dite aussi SPECT	DAT SPECT	-Utile dans le diagnostic de la MP préclinique, -Inutile pour apprécier l'évolution.	(Booij et Knol 2007) (Schwingenschuh, et al. 2010)
TEP dite aussi PET	^{18}F-fluorodopa et DAT PET	-Très sensible pour détecter la MP préclinique, -Trop cher et pourrait être affecté par la lévodopa.	(Panzacchi, et al. 2008) (Bohnen, et al. 2006) (Hilke, et al. 2005) (STOESSL et Jon 2007)
Sonographie	Hyperéchogénicité de la SN	-Utile dans le diagnostic de la MP préclinique, -Ne permet pas d'apprécier la progression de la maladie.	(Auer et Dorothee 2009) (Berg, Godau et Walter 2008)
IRMf	Fer dans la SN ;	-Utiles dans le diagnostic de la	(Martin, Wieler et

	anisotropie fractionnelle en IRM	MP préclinique, en particulier combinée avec l'utilisation de biomarqueurs cliniques, mais d'autres études sont nécessaires.	Gee 2008) (Vaillancourt, et al. 2009)
La scintigraphie	MIBG	- Utile dans le diagnostic préclinique de la MP - Manque de sensibilité.	(Spiegel, et al. 2007)

Spécifications Médicales :

Transporteur de la dopamine (DAT);

Imagerie par Résonance Magnétique(IRM);

Imagerie par Résonance Magnétique Fonctionnelle (IRMf);

Méta-Iodo Benzyl Guanidine (MIBG);

Tomographie par Émission de Positons (TEP);

Substantia Nigra (SN);

Tomographie d'Émission monophotonique (TEMPS) ;

1.3.3 **Diagnostic biologique**

Cette méthode de diagnostic se base sur la recherche de substances normalement présentes et dont le taux est modifié au cours de l'évolution de la MP. Les techniques utilisées se basent sur la génomique, transcriptomique, proteomique, metabolomique (Bogdanov et al 2008) (Gasser 2009) (Hwang, et al. 2010) (Quinones et Kaddurah-Daouk 2009) (Schipper, et al. 2008) (Hennecke et Scherzer 2008). Les prélèvements utilisés ainsi que leur utilisation sont résumés dans le Tableau 1.2

***Tableau 1. 2**:* Techniques biologiques utilisées dans le diagnostic de la MP

Prélèvement	Biomarqueurs	**Évaluation**	**Source**
LCR	α –synucléine LRRK2, NURR1	-Utile dans le diagnostic de la MP préclinique, -Ne permet pas d'apprécier la progression de la maladie.	(Shi, et al. 2011) (Hong, et al. 2010) (Westermann, et al. 2008)
Sang	α-synucléine, ST13, urate	- Utile pour le diagnostic de la MP préclinique	(Brighina, et al. 2010) (Li, et al. 2007) (Le, et al. 2008) (Scherzer, et al. 2007)

Spécifications Médicales :

Liquide céphalo-rachidien (LCR);

Leucine-Rich Repeat Rinase 2 (LRRK2);

Nucléaire récepteur lié une protéine(NURR1);

Suppression des Tumorigénicité (ST13).

1.4 Modalités de diagnostic automatique de la maladie de Parkinson

Les systèmes automatiques de diagnostic de la MP sont basés sur plusieurs modalités physiologiques. On peut les résumer comme suit :

1.4.1 Électrocardiogramme (ECG)

La MP entraîne un dysfonctionnement du cœur dû à une perte de l'innervation de ce dernier, par conséquent, il y aura une variabilité de l'amplitude cardiaque qui sera appréciée grâce à l'ECG (Haapaniemi, et al. 2001).

1.4.2 **Electroencéphalographie (EEG)**

Un ralentissement du tracé EEG est constaté dans le cas de la MP. Cela est apprécié grâce à l'enregistrement du signal sur un support informatique (Soikkeli, et al. 1991).

1.4.3 **Magnétoencéphalographie (MEG)**

Dans le cerveau, il existe un champ magnétique induit par l'activité électrique des neurones. La MEG est une technique de mesure de champ magnétique. Un enregistrement des perturbations des signaux magnétiques générés par les oscillations des neurones dans la MP (oscillation anormale dans déférente bande de fréquence) (Stam 2010) est analysé à l'aide d'un logiciel**.** La MEG est une technique plus précise que l'EEG dans le diagnostic de la MP.

1.4.4 **Imagerie par résonance magnétique (IRM)**

Les développements récents dans les méthodes d'imagerie cérébrale sont sur le point de changer l'évaluation des personnes atteintes de la maladie de Parkinson (MP). Cela comprend un assortiment de techniques allant de tenseur de diffusion imagerie, l'IRM fonctionnel à l'état de repos, et la spectroscopie par résonance magnétique. En utilisant une approche multimodale qui détermine les différents aspects de la physiopathologie ou la pathologie du MP, il peut être possible de mieux caractériser les phénotypes de la maladie (Martin, Wieler et Gee 2008).

1.4.5 **Voix**

Les troubles de la voix, au même titre que la dysphonie, appartiennent aux manifestations de la dysarthrie parkinsonienne. En effet, la production de la voix met en jeu de nombreux organes et structures cérébrales que la MP peut affecter, entraînant des troubles et affectant l'articulation et la fluence. Ces troubles sont d'apparition plus tardive, mais entravent considérablement l'intelligibilité. La description du processus de production de la voix nous permettra d'en aborder les caractéristiques pathologiques (Amélie et Aurélie 2011).

a ***Traits acoustiques relatifs à la maladie de Parkinson***

Dans cette section, nous présentons les différentes caractéristiques de la voix qui peuvent être affectées dans le cas où une personne serait atteinte de la MP.

- **Fréquence fondamentale moyenne** $(\boldsymbol{F_0})$

Mesurée en Hertz (Hz), elle détermine la hauteur de la voix. Elle dépend du nombre de vibrations par seconde des cordes vocales (Maude 2012).

$$F_0(Hz) = {}^{1}/_{T} \qquad (1.1)$$

T: La période.

F_0: Fréquence fondamentale moyenne.

Les paramètres acoustiques qui peuvent être extraits à partir de F_0 sont donnés ci-après :

- **La plus haute fréquence fondamentale (Fhi)**

- **La plus basse fréquence fondamentale (Flo)**

- **Jitter**

Le Jitter est un paramètre important pour le diagnostic des troubles de la voix (Lhote 1982). Il représente l'altération de la périodicité de F_0 comme l'illustre la Figure 1.1.
Il existe différentes mesures de ce paramètre. On peut citer : Le Jitter absolu, Jitter(%), Jitter RAP, Jitter PPQ et Jitter DDP (Farrus et Hernando 2009).

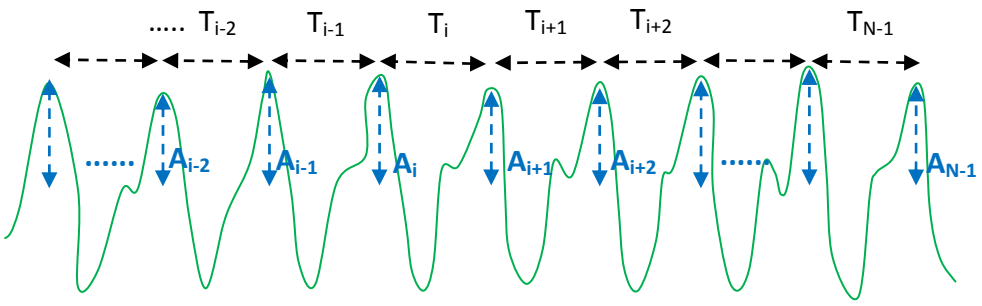

Figure 1.1 : *Paramètres Jitter (T_i) et Shimmer (A_i)*

- **Jitter(%)**

$$\text{Jitter}\ (\%) = \frac{1}{\mathrm{N}} \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}-1} |\mathrm{T_i} - \mathrm{T_{i-1}}|}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{T_i}} \quad (1.2)$$

T_i: Période de l'intervalle i^{ème}.

N: Nombre d'intervalles.

- **Jitter absolu**

$$\text{Jitter Abs} = \frac{1}{N-1} \sum_{i=1}^{N-1} |\mathrm{T_i} - \mathrm{T_{i-1}}| \quad (1.3)$$

- ***Jitter RAP*** *(Relative Average Perturbation)*

$$Jitter\ RAP = \frac{\sum_{i=2}^{N-1} |T_i - (T_{i-1} + T_i + T_{i+1})/3|}{(N-2)} \quad (1.4)$$

- ***Jitter PPQ*** *(5 Points Period Perturbation quotient)*

$$Jitter\ PPQ = \frac{\sum_{i=3}^{N-2} |T_i - (T_{i-1} + T_i + T_{i+1} + T_{i+2})/5|}{(N-4)} \quad (1.5)$$

- ***Jitter DDP*** *(Difference of Differences of Periods)*

$$Jitter\ DDP = \frac{\sum_{i=2}^{N-1} (T_{i+1} - T_i) - (T_i - T_{i-1})}{(N-2)} \quad (1.6)$$

- ***Shimmer***

Le Shimmer est la variation d'amplitude entre les périodes successives quand un individu tente de tenir la phonation à une fréquence et une intensité constantes (Lhote 1982) (Figure 1.1).

Il existe différentes mesures du Shimmer. Le Shimmer ordinaire, Shimmer dB, Shimmer factor, Shimmer APQ3, Shimmer APQ5, Shimmer APQ et Shimmer DDA (Farrus et Hernando 2009).

- ***Shimmer « ordinaire »***

$$Shimmer = \frac{\sum_{i=1}^{N-1} {A_i}/{A_{i+1}}}{(N-1) \sum_{i=1}^{N-1} A_i} \quad (1.7)$$

- ***Shimmer dB***

$$Shimmer\ dB = \frac{\sum_{i=1}^{N-1}\left|20log_{10}(^{A_i}/_{A_{i+1}})\right|}{N-1} \quad (1.8)$$

A_i: L'amplitude maximale sur l'intervalle i.

N: Le nombre d'intervalles.

- ***Shimmer APQ3 (3 points Amplitudes Perturbations Quotient)***

C'est la moyenne des variations d'amplitude sur trois cycles vibratoires du larynx rapportée à l'amplitude moyenne du signal.

$$Shimmer\ APQ3 = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}|A_i-(A_{i-1}+A_i+A_{i+1})/3|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \quad (1.9)$$

- ***Shimmer APQ5 (5 points Amplitudes Perturbations Quotient)***

C'est la moyenne des variations d'amplitude sur cinq cycles vibratoires du larynx rapportée à l'amplitude moyenne du signal.

$$Shimmer\ APQ5 = \frac{\frac{1}{N-4}\sum_{i=3}^{N-2}|(A_{i-2}+A_{i-1}+A_i+A_{i+1}+A_{i+2})/5|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \quad (1.10)$$

- ***Shimmer APQ (11 points Amplitudes Perturbations Quotient)***

C'est la moyenne des variations d'amplitude sur 11 cycles vibratoires du larynx rapportée à l'amplitude moyenne du signal.

$$Shimmer\ APQ = \frac{\frac{1}{N-10}\sum_{i=6}^{N-5}|(A_{i-5}+A_{i-4}+\cdots+A_{i+4}+A_{i+5})|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \quad (1.11)$$

- ***Shimmer DDA (Difference of Differences of Amplitude)***

$$Shimmer(DDA) = \frac{\sum_{i=2}^{N-1}|(A_{i+1}-A_i)-(A_i-A_{i-1})|}{(N-2)} \quad (1.12)$$

- ***Rapport bruit sur harmoniques (Noise to Harmonic Ratio NHR)***

Le NHR mesure le quotient entre signal vocal et bruit et permet d'évaluer la qualité vocale. Un rapport son/bruit élevé signifie que le bruit est faible et que le signal vocal est de bonne qualité (Maude 2012).

- ***Rapport harmonique sur bruit (Harmonics To Noise Ratio HNR)***

Le HNR, une mesure explorant la présence du bruit au cours de la phonation peut être calculée selon plusieurs méthodes (Yumoto, Gould et Baer 1982).

- ***Entropie récurrente de la densité de la période (Recurence Period Density Entrepy RPDE)***

Il s'agit d'une nouvelle méthode mise en place afin de mesurer la périodicité d'un signal après sa reconstruction dans un nouvel espace (T. Tsanas 2012).

$$H_{norm} = -(\ln T_{max}) - \sum_{t=1}^{T_{max}} P(t) \ln P(t) \qquad (1.13)$$

t : Délai.

P(t): Fonction de densité de periode de récurrence.

T_{max}: Plus grande valeur de récurrence.

- ***Analyse de fluctuation (Detrended Fluctuation Analysis DFA)***

Mathématiquement, ce paramètre acoustique est défini pour mettre en évidence les processus d'autosimilarités dans les séries temporelles. Pour le signal vocal, le DFA mesure le degré d'autosimilarité du bruit dans le signal de la parole (T. Tsanas 2012).

- **Mesures non linéaires de variation de F0 type 1 *(Spread 1)***

Spead 1 est le logarithme de la variance des périodes de pas blanchi (Kumar 2011).

- **Mesures non linéaires de variation de F0 type 2 *(Spread 2)***

Spead2 est l'entropie (estimée à l'aide des histogrammes) des périodes de pas blanchie, la fonction «entropie» calcule juste l'entropie de Shannon (Kumar 2011).

- ***Dimension de corrélation D2 (Correlation dimension D2)***

D2 est une mesure géométrique qui décrit le degré de corrélation entre deux points représentés dans l'espace des phases (Jiang, Zhang et McGilligan 2006) .Elle est ainsi très utile pour décrire des phénomènes irréguliers d'un signal donné. Ce paramètre a été largement utilisé par les chercheurs en raison de sa simplicité et sa convergence rapide dans les calculs numériques.

- ***Entropie de la période du pitch (Pitch Period Entropy PPE)***

Il s'agit d'un paramètre permettant d'évaluer la capacité d'une personne donnée à maintenir une fréquence fondamentale stable lors de l'examen de la voyelle tenue. En effet, toutes les personnes présentent une variation du pitch, même les personnes saines ne présentant aucune pathologie de la voix, c'est ce que l'on appelle une variation naturelle caractérisée par des vibrations lisses et des tremblements très fins.

1.5 Conclusion

L'étude accomplie dans ce premier chapitre nous a permis de découvrir le risque de cette maladie. Elle nous a aussi permis de comprendre les différentes méthodes de diagnostic, en particulier ceux qui sont reliés aux signaux vocaux. Dans le prochain chapitre, nous présenterons un système automatique de diagnostic de la MP qui est basé sur l'onde acoustique de la voix.

Chapitre 2 Aide au diagnostic de la maladie de Parkinson à partir de la voix

2.1 Introduction

La voix est le résultat d'une coordination du larynx, conduit vocal, la langue et les lèvres. Chez une personne atteinte de la malade de Parkinson, cette coordination est altérée. La voix a des rigidités, des faiblesses et des tremblements. Ces altérations représentent des indices de la maladie.

Dans ce chapitre, nous présentons les systèmes d'aide au diagnostic de cette maladie à pâtir de l'onde acoustique de la voix. Nous présentons aussi une description des étapes de traitement et une méthode d'évaluation globale des performances.

2.2 Système de diagnostic

o D'où est venue l'idée du diagnostic de la MP à partir de la voix?

Une nouvelle piste de recherche a été étudiée par un mathématicien du MIT (l'Institut technologique du Massachusetts), Max Little, qui s'est associé à deux autres chercheurs de l'université britannique d'Oxford. Son idée était de mettre au point un système qui permet de détecter aisément cette maladie en analysant la voix du patient (Tsanas, et al. 2012) (Figure 2.1).

Le système est capable de déceler les infimes changements de la voix qui représentent les premiers signes de cette pathologie.

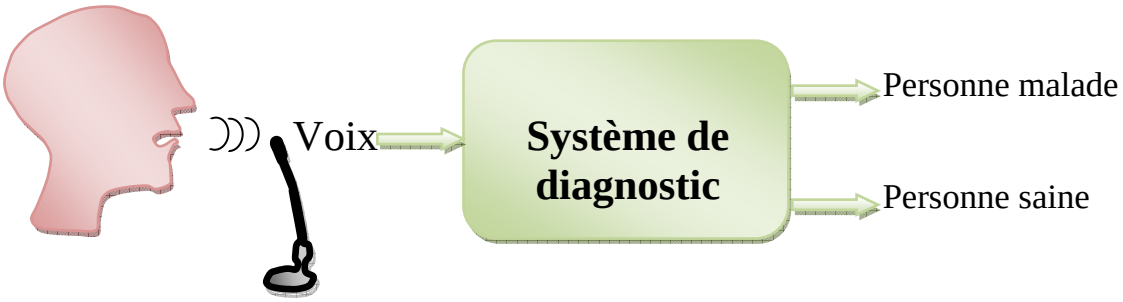

Figure 2.2: Détection de Parkinson à partir de la voix

Pour développer ce système, le chercheur a rassemblé des échantillons de voix des personnes de bonne santé et des personnes atteintes de cette maladie.

Il a ensuite développé un algorithme capable de détecter les particularités de ces échantillons, et de déceler les changements purement associés à la maladie.

Le travail que l'on va présenter dans ce mémoire consiste à développer une variante du système proposé par Dr Max Little. On essaye aussi d'utiliser une technique de sélection des attributs pour diminuer la charge de calcul. Par la suite, on compare les performances du système proposé avec ceux qui sont obtenus avec l'approche globale.

2.3 Description des étapes de traitement

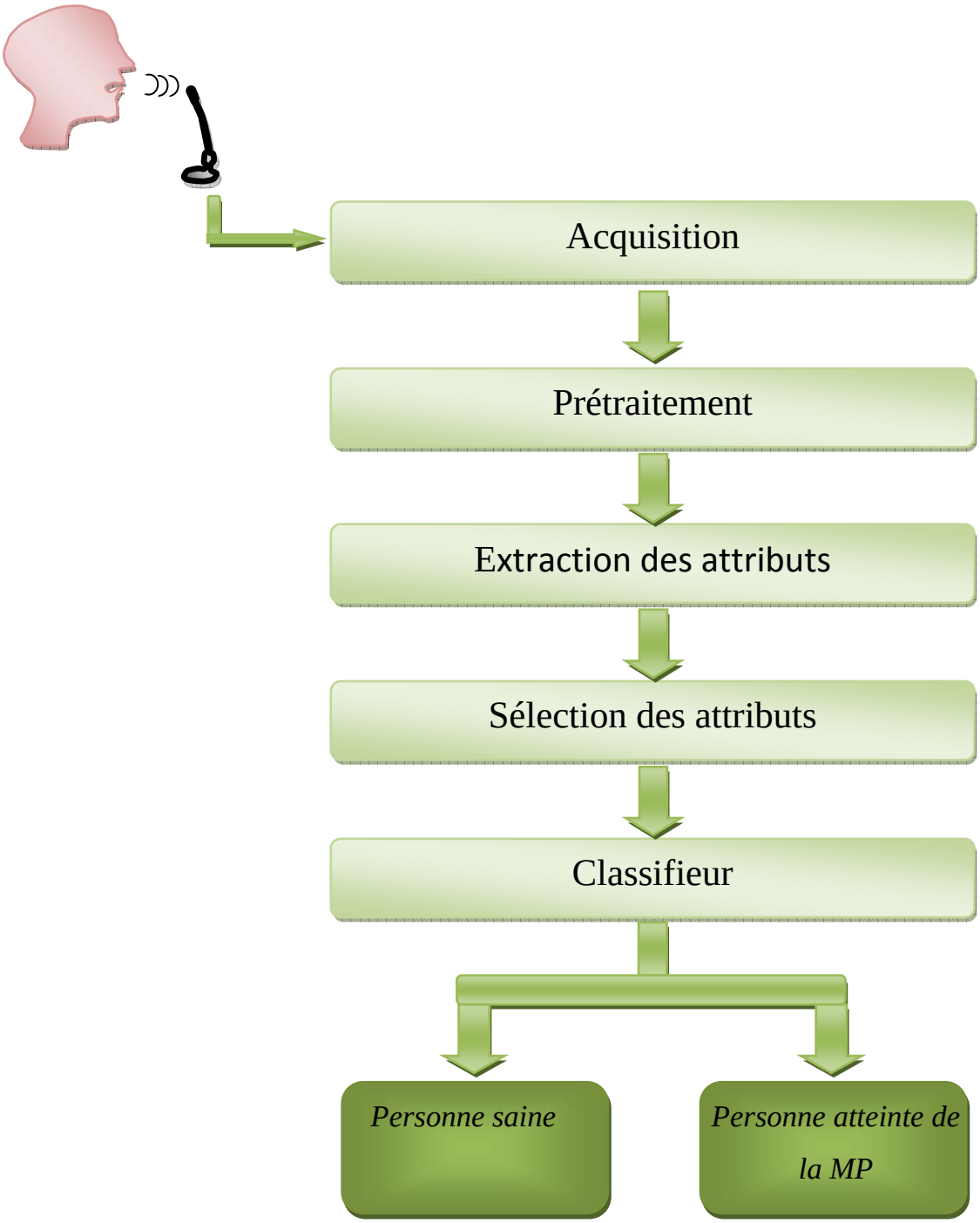

Figure 2.3: Organigramme de diagnostic

Le système de diagnostic est constitué cinq blocs (Figure 2.2) :

2.3.1 **Acquisition de la voix**

L'acquisition de la voix ce fait à partir des bases de données internationales. Une base de données est un ensemble de fichiers homogènes, cohérents et représentatifs. Elle doit être facilement utilisable en alliant la souplesse, la simplicité, la convivialité et la fiabilité. Il existe plusieurs bases de données dans ce domaine. On peut citer :

a ***MEEI***

MEEI "massachusetts Eye and Ear Infirmary (MEI) Voix Disorders Database" est une base de données des voix pathologiques constituée par l'équipe Géostat de l'institut technologique MIT aux États-Unis. Cette base de données est distribuée par Kay Elemetrics (Safaa 2013).

Elle contient environ 1400 enregistrements vocaux qui sont de deux types : une voyelle tenue (/a/) et une phrase bien spécifique. Ces échantillons sont obtenus à partir des enregistrements de 700 personnes. Il s'agit de la seule base de données qui est commercialement disponible.

Cette base de données est souvent utilisée pour beaucoup de travaux de recherche. Tous les enregistrements et les informations cliniques relatives aux personnes archivées dans la base ont été réalisés au sein de MEEI Voice and Speech Lab.

Ces échantillons sont obtenus à partir d'un total de 710 personnes dont 53 possédant des voix normales et 657 souffrant des pathologies de la voix. Chaque échantillon de voix normale a une durée égale à trois secondes. Un échantillon d'un /a/ tenu pour une voix pathologique est de durée égale à une seconde (Markaki 2011).

b ***Parkinson's dataset***

La base de données que nous avons utilisée dans notre projet est nommée ‘’*Parkinson Dataset*’’. Elle a été créée par Max Little, de l'Université d'Oxford, en collaboration avec le centre national de la voix de *Denver, Colorado* (Little, et al. 2009). Cette base de données est composée d'une série de mesures vocales biomédicales (attributs) de 31 personnes (12 hommes et 19 femmes) et l'âge des patients variés entre 46 et 85 ans dont 23 d'entre elles souffrent de la MP. Chaque colonne de la table représente un

attribut particulier et chaque ligne correspond à un enregistrement vocal spécifique (essai). L'ensemble total des enregistrements est égal à 195, chaque échantillon de voix à une durée égale à trois secondes. Il y a environ six enregistrements par patient. Les attributs sont présentés dans le Tableau 2.1.

c ***German database***

La base de données a été collectée dans le cadre d'un projet de collaboration entre le ministère de la santé, la clinique Saint-Caritas Theresia et l'Institut de phonétique de l'Université de la Sarre en Allemagne. Jusqu'à présent, plus que 95 pathologies vocales ont été inscrites. Les signaux enregistrés sont composés d'une phase complète et un ensemble de voyelles de la langue Allemande (Pützer et Koreman 1997).

2.3.2 Prétraitement

Les bons résultats qu'un classificateur automatique peut fournir reposent, en grande partie, sur la phase de prétraitement. Les données issues d'un mauvais prétraitement vont mettre en péril la qualité du classificateur. Cette phase consiste en une succession de traitement sur les données brutes afin d'extraire de l'information et de ne garder que celle qui est utile à la classification (Marref 2013).

2.3.3 Extraction des attributs

L'extraction des attributs pour la détection de la MP à partir de la voix consiste à calculer les paramètres représentés dans le Tableau 2.1.

Tableau 2. 3*:* Attributs de la MP

Nombre d'entités	Paramètres	Description
1	MDVP : Fo(Hz)	Moyenne de la fréquence Fondamentale (F_0)
2	MDVP : Fhi (Hz)	Maximum de F_0
3	MDVP : Flo (Hz)	Minimum de F_0
4	MDVP : Jitter(%)	Jitter en pourcentage (Variation de F_0)

5	MDVP : Jitter(Abs)	Jitter en valeur absolue (Variation de F_0)
6	MDVP:RAP	Amplitude relative de Perturbation (variation d'amplitude)
7	MDVP: PPQ	Les cinq points du quotient de la période de perturbation (Variation de F_0)
8	Jitter : DDP	Différence absolue moyenne des différences entre les cycles, divisée par la période moyenne (Variation de F_0)
9	MDVP : Shimmer	Shimmer local (variation d'amplitude)
10	MDVP : Shimmer (dB)	Trois points du quotient de la perturbation d'amplitude (variation d'amplitude)
11	Shimmer : APQ3	Trois points du quotient de la perturbation d'amplitude (variation d'amplitude)
12	Shimmer : APQ5	Cinq points du quotient de la perturbation d'amplitude (variation d'amplitude)
13	MDVP : APQ	Onze points du quotient de la perturbation d'amplitude (variation d'amplitude)
14	Shimmer : DDA -	Différence absolue moyenne entre les différences Consécutives d'amplitude des périodes consécutives (variation d'amplitude)
15	NHR	Rapport bruit sur harmoniques
16	HNR	Rapport harmonique sur bruit
17	Statut	État de santé 1 - Parkinson ; 0 personne Sainte
18	RPDE	Entropie récurrente de la densité de la période
19	DFA	Analyse de fluctuation
20	Spread1	mesures non linéaires de

		Variation de F_0 type 1
21	spread2	mesures non linéaires de Variation de F_0 type 2
22	D2	mesure des dynamiques non linéaires
23	PPE	mesures non linéaires de Variation de F_0

2.3.4 Sélection des attributs

La sélection des attributs est une approche qui permet de choisir le meilleur ensemble d'attributs utilisés en vue d'une classification. Cette étape a pour objectif de faciliter l'apprentissage et de réduire la complexité de calcul.

Les méthodes de sélection d'attributs sont divisées en trois catégories, selon la manière dont elles interagissent avec le classifieur (Mahdjane 2012).

a ***Méthode de filtrage (filter)***

Dans cette méthode la sélection des attributs est une étape indépendante de la construction du classifieur. C'est une étape de prétraitement des données (Mahdjane 2012). La Figure 2.3 représente son principe.

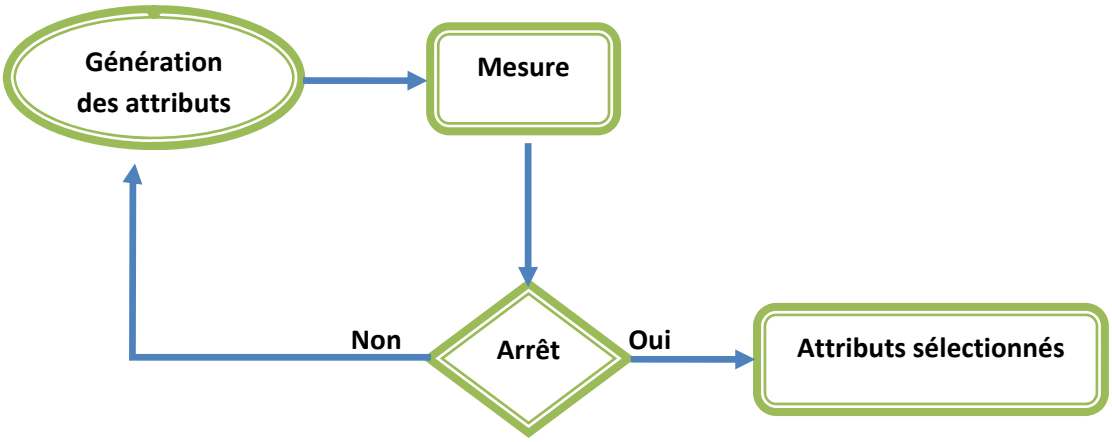

Figure2.4: Principe de l'approche par filtrage (Mahdjane 2012)

b ***Méthode enveloppante (wrapper)***

Dans cette méthode, le mécanisme de sélection interagit avec un classifieur pour trouver un sous-ensemble d'attributs qui sera optimal pour ce modèle d'apprentissage.

Elle a souvent de meilleurs résultats que la méthode de filtrage (Mahdjane 2012), mais au prix d'un temps de calcul plus important (Piyushkumar, et al. 2007).

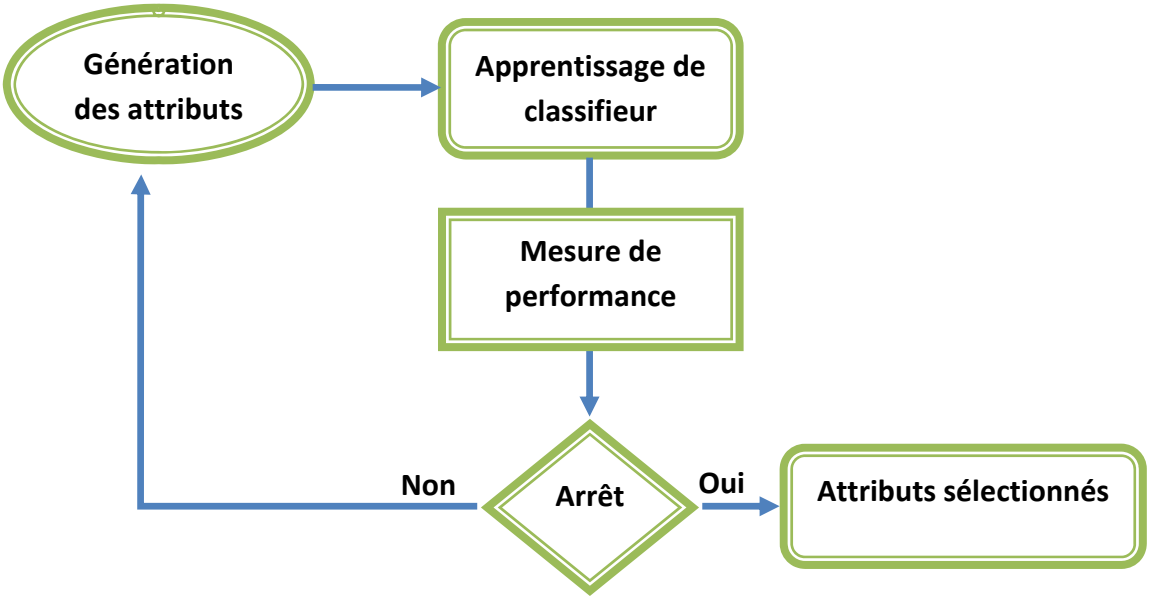

Figure2.5: Principe de l'approche enveloppante (Mahdjane 2012)

c ***Méthode intégrée (embedded)***

Cette méthode est proche des méthodes d'enveloppe, car elle combine le processus d'exploration avec un algorithme d'apprentissage. La différence avec les méthodes enveloppes est que le classifieur sert non seulement à évaluer un sous-ensemble candidats mais aussi à guider le mécanisme de sélection (Mahdjane 2012). Plus de détails sur cette méthode sont donnés dans (Singh, Pillay et Choonara 2007).

Dans notre travail, on s'intéresse seulement aux méthodes de type filter.

- **Techniques de filtrage d'ordonnancement des attributs**

Dans cette partie, nous donnons un bref aperçu de quelques techniques de filtrage d'ordonnancement des attributs. Nous nous intéressons aux :

- ✓ Fisher
- ✓ mRMR
- ✓ Chi-square

o **Le critère de *Fisher***

Permet de mesurer le degré de séparabilité des classes à l'aide d'une caractéristique donnée (Chouaib 2011). Il est défini par :

$$F_i = \frac{\sum_{Y-1}^{Y} n\left(u_Y^i - \mu^i\right)^2}{\sum_{Y-1}^{Y} n\left(\sigma_Y^i\right)^2} \quad (2.1)$$

n : L'effectif.

u_Y^i : La moyenne.

σ_Y^i : L'écart type du ième attribut au sein de la classe Y.

μ^i : La moyenne globale de l'i^{ème} attribut.

Le tableau 2.2 représente un exemple de 4 attributs et 2 classes avec 195 effectifs.

***Tableau 2. 4**: Fisher score* de 2 classes et 4 caractéristiques

Les classes Y	Classe 1				Classe 2			
Les attributs	attr1	attr2	attr3	attr4	attr1	attr2	attr3	attr4
n=1…N	f_1^1	f_2^1	f_3^1	f_4^1	f_1^1	f_2^1	f_3^1	f_4^1
Avec N=195	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	f_1^n	f_2^n	f_3^n	f_4^n	f_1^n	f_2^n	f_3^n	f_4^n

- **Chi-square**

Chi-square est utilisé pour évaluer les deux types de comparaison: les tests de qualité de l'ajustement et tests d'indépendance.

Dans la sélection des attributs, il est utilisé comme un test d'indépendance pour évaluer si l'étiquette de classe est indépendante d'une caractéristique particulière. Le score Chi-carré pour une fonction avec r différentes valeurs et classes Y est défini comme étant (Liu et Setiono 1995).

$$Q_i = \sum_{i=1}^{r} \sum_{j=1}^{Y} \frac{(n_{ij} - u_{ij})^2}{u_{ij}} \quad (2.2)$$

$$\text{Avec } \mu_{i,j} = \frac{n_{*j} n_{i*}}{n} \quad (2.3)$$

n : L'effectif.

n_{ij}: Le nombre d'échantillons avec la $i^{iéme}$ valeur d'attribut.

u_{i*} : Le nombre d'effectifs avec la valeur de la $i^{iéme}$ de l'attribut particulière.

u_{*j} : Le nombre d'effectifs dans la classe Y.

r : Nombre des attributs.

Tableau 2. 5*:* Chi-square de 2 classes et 4 caractéristiques

Les classes Y	Classe 1				Classe 2			
Les attributs $i = 1 \dots r$	Attr1	attr2	attr3	attr4	attr1	attr2	attr3	attr4
n=1…N	Q_1^1	Q_2^1	Q_3^1	Q_4^1	Q_1^1	Q_2^1	Q_3^1	Q_4^1
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	Q_1^n	Q_2^n	Q_3^n	Q_4^n	Q_1^n	Q_2^n	Q_3^n	Q_4^n

- **Minimum Redondance Maximum Relevance (mRMR)**

mRMR '' En Anglais : *Min-Redundancy, Max-relevance*" est une méthode de filtrage pour la sélection des attributs proposée par Peng et al en 2005. Cette méthode est basée sur des mesures statistiques classiques comme l'information mutuelle, la corrélation, etc.

L'idée de base est de profiter de ces mesures pour essayer de minimiser la redondance (mR) entre les variables et de maximiser la pertinence (MR) (Settouti et Hafa 2013).

$$Redondance(i) = \frac{1}{|F|^2}\sum_{j\in F} I(i,j) \quad (2.4)$$

$$Pertinence(i) = \frac{1}{|F|^2}\sum_{Y\epsilon F} I(i,Y) \quad (2.5)$$

$|F|$: La taille de l'ensemble des attributs.

$I(i,j)$: L'information mutuelle entre l'i^{ème} et la j^{ème} attribut.

$I(i,Y)$: L'information mutuelle entre l' i^{ème} attribut et l'ensemble des étiquettes de classes (Y).

Le score d'un attribut est la combinaison de ces deux facteurs tels que :

$$Score(i) = \frac{Pertinence(i)}{Redondance(i)} ou\ Score(i) = Pertinence(i) - Redondance(i) \quad (2.6)$$

- **Information mutuelle**

L'information mutuelle introduite par Shannon et al mesure la dépendance de deux variables aléatoires discrètes X et Y de densité de probabilité p(x) et p(y). Elle est définie par :

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x).p(y)} \quad (2.7)$$

$p(x,y)$: La densité de probabilité conjointe des deux variables X et Y.

L'information mutuelle n'est nulle si les variables sont indépendantes et croît lorsque la dépendance augments.

En divisant l'information mutuelle définie par équation (2.7) par la valeur maximale des entropies H(X) et H(Y), on obtient l'information mutuelle normalisée défini par :

$$NI(X,Y) = \frac{\sum_{x\in X}\sum_{y\in Y}(x,y)\log\frac{p(x,y)}{p(x).p(y)}}{\max\left(H(X),H(Y)\right)} \quad (2.8)$$

Avec :

$$H(X) = -\sum_{x\in X} p(x)\log p(x) \quad (2.9)$$

L'information mutuelle normalisée de l'équation (2.8) est ainsi utilisée pour mesurer la dépendance entre un attribut et leur Y de labels des classes des données (Kalakech 2011).

2.3.5 **Classification**

Les méthodes de classification ont pour but d'identifier les classes auxquelles appartiennent des objets à partir de certains paramètres descriptifs. Elles s'appliquent à un grand nombre d'activités humaines et conviennent en particulier au problème de la prise de décision automatisée. La procédure de classification sera extraite automatiquement à partir d'un ensemble d'exemples. Un exemple consiste en la description d'un cas avec la classification correspondante. Un système d'apprentissage doit alors, à partir de cet ensemble d'exemples, extraire une procédure de classification, il s'agit en effet d'extraire une règle générale à partir des données observées. La procédure générée devra classifier correctement les exemples de l'échantillon et avoir un bon pouvoir prédictif pour classifier correctement de nouvelles descriptions (Cherif et Ikram 2011).

Les méthodes utilisées pour la classification sont nombreuses, on peut citer : les réseaux de neurones, les méthodes des séparateurs à vastes marges (en Anglais : *Support Vecteur Machines* (SVM)). Nous présentons dans la suite de ce chapitre une étude détaillée de technique SVM.

- ***Support Vector Machines***

Les machines à vecteurs supports ou séparateurs à vaste marge sont des nouvelles techniques d'apprentissage statistique ; qui sont proposées par V. Vapnik en 1995

(V.Vapnik 1995). Elles permettent d'aborder des problèmes de discrimination, c'est-à-dire décidé à quelle classe appartient un échantillon, ou de régression (Marref 2013). Les SVM sont utilisées dans des nombreux problèmes d'apprentissage. Par exemple : la reconnaissance de formes, la catégorisation de texte et encore le *diagnostic médical* (Newton, Hse et Richard 2004) (Zidelmal, et al. 2007) (Abibullaev, Kang et Lee 2010).

- ***Principe de fonctionnement***

Le principe des SVMs est de trouver l'hyperplan optimal parmi l'ensemble des hyperplans possibles (Figure 2.5) permettant de classifier correctement les données et maximiser sa distance entre les vecteurs supports les plus proches.

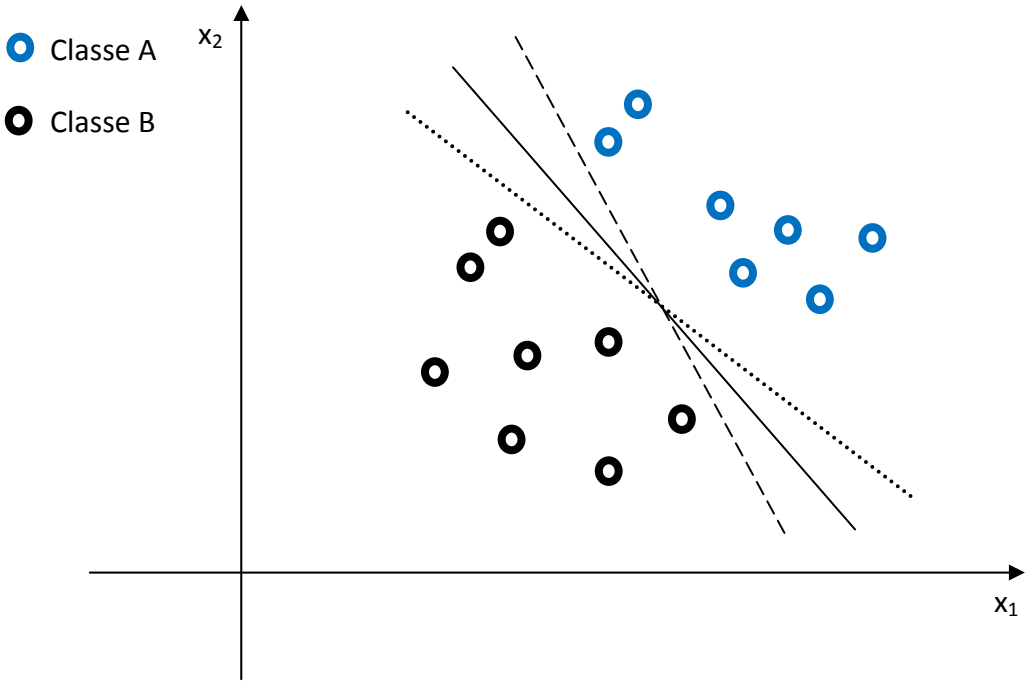

Figure2.6: Recherche de l'hyperplan optimal

Hyperplan

On appelle *hyperplan séparateur* un hyperplan qui sépare les points d'apprentissage de deux classes données (A et B). Les deux hyperplans (H_1) et (H_2) sont appelés hyperplans canoniques.

Vecteurs supports

La détermination de l'hyperplan est basée seulement sur les points situés sur la frontière entre les deux classes de données (parmi l'ensemble total d'apprentissage). Ces points sont appelés *vecteurs supports* (Figure 2.6).

Marge

Il existe une infinité d'hyperplans capable de séparer parfaitement les deux classes d'exemples. Le principe des SVM est de choisir celui qui va maximiser la distance minimale entre l'hyperplan et les exemples d'apprentissage (i.e. la distance entre l'hyperplan et les vecteurs supports), cette distance est appelée la marge.

Le fait d'avoir une marge plus large procure plus de sécurité lorsque l'on classe un nouvel exemple (Marref 2013).

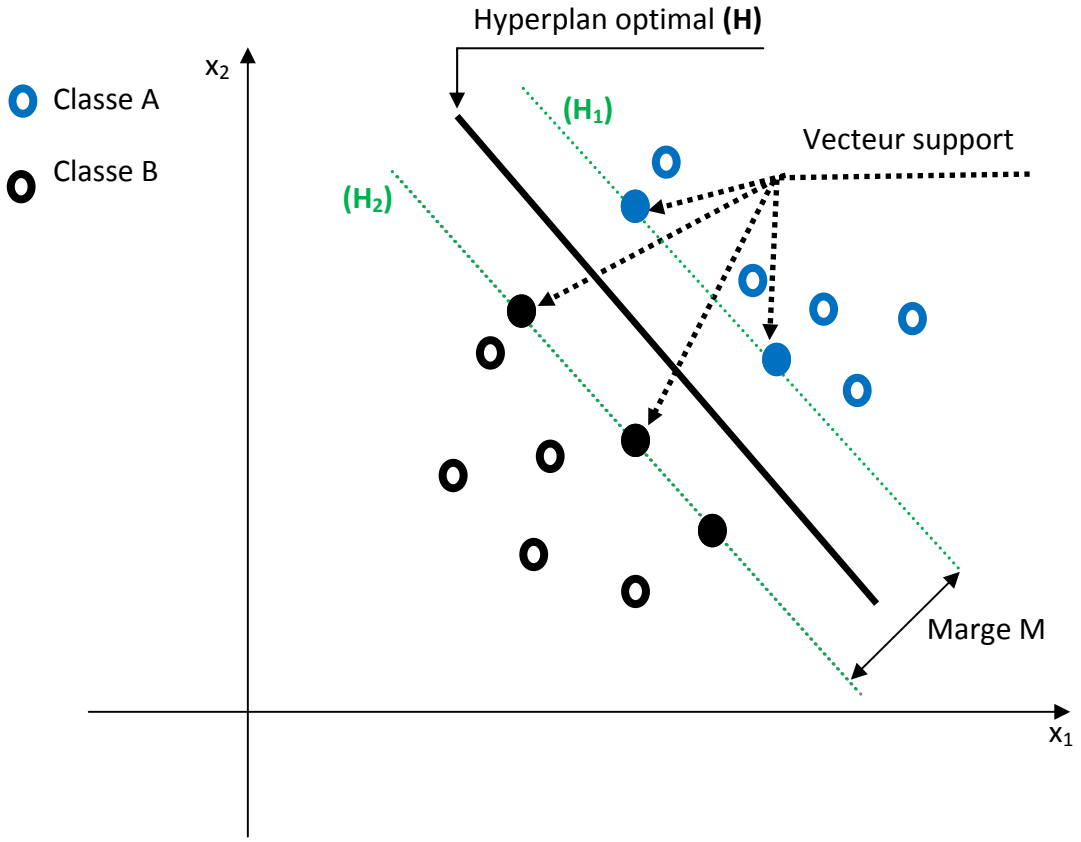

Figure 2.7: L'hyperplan H optimal, vecteurs supports et marge maximale. x_1 et x_2 représentent les attributs des classes A et B

Linéarité et non-linéarité

Parmi les modèles des SVMS, on constate les cas linéairement séparables et les cas non linéairement séparables (Figure 2.7). Les premiers sont les plus simples, car ils permettent de trouver facilement le classificateur linéaire. Dans la plupart des problèmes réels, il n'y a pas de séparation linéaire possible entre les données, le classificateur de marge maximale ne peut pas être utilisé, car il fonctionne seulement si les classes de données d'apprentissage sont linéairement séparables.

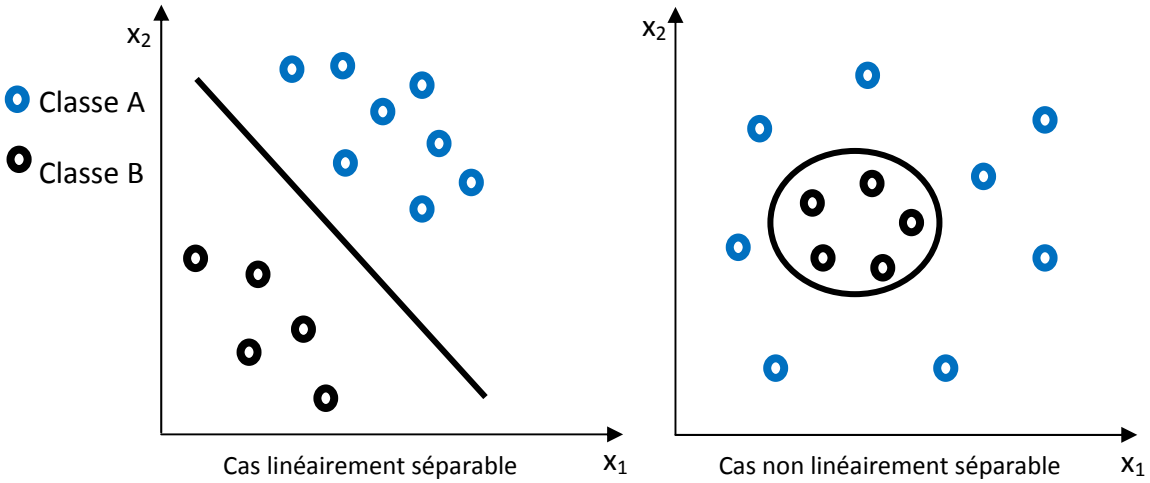

Figure2.8: Séparation linéaire et non linéaire

Cas linéairement séparable

Un classificateur est dit linéaire lorsqu'il est possible d'exprimer sa fonction de décision par une fonction linéaire en x (Figure2.8). Dans la suite, nous supposons que les exemples sont donnés dans un format vectoriel. Notre espace d'entrée x correspond donc à R^n ou n est le nombre de composantes des vecteurs contenant les données (Marref 2013).

$$h(x) =< w.x > +b = \sum_{i=1}^{n} w_i x_i + b \qquad (2.10)$$

Si les données sont linéairement séparables, alors il existe un hyperplan d'équation :

$$< w.x > + b = 0 \; et \; tel \; que$$

$$< w.x > + b \geq 1 \; si \; yi = +1 \qquad (2.11)$$

$$< w.x > + b \leq -1 \; si \; yi = -1 \qquad (2.12)$$

On peut combiner ces deux inéquations en une seule :

$$y_i(w.x + b) \geq 1 \quad pour \; i = 1, \ldots, n \qquad (2.13)$$

Ou x_i : Vecteur des attributs (x_1,x_2,x_3,…,x_n).

w_i : Vecteur des poids (w_1,w_2,w_3,…,w_n).

b : Seuil du séparateur linéaire.

y_i: Label de classe pouvant prendre la valeur $+1$ ou -1.

h : Fonction de décision.

i : La dimension des vecteurs d'entrée.

n: La taille de l'ensemble d'apprentissage.

La distance de l'hyperplan de séparation

$$d(x) = \frac{|<w.x>+b|}{\|w\|} \geq \frac{1}{\|w\|} \tag{2.14}$$

La distance de l'hyperplan à l'origine

$$\frac{b}{\|w\|} \tag{2.15}$$

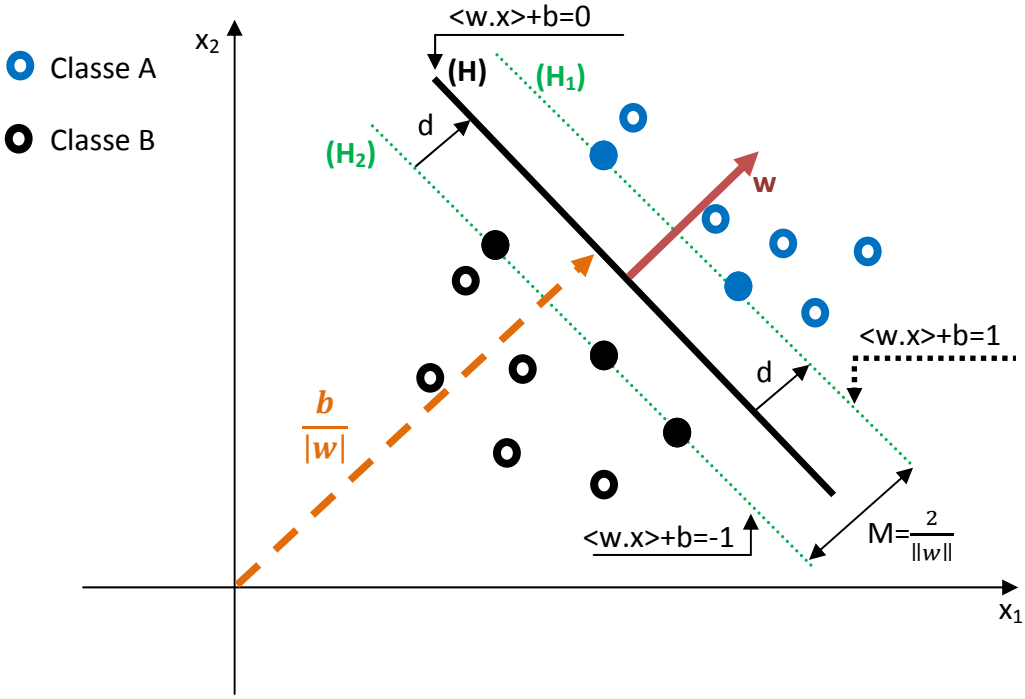

Figure2.9: Exemple graphique des données linéairement séparables

La marge géométrique représente la distance euclidienne prise perpendiculairement entre l'hyperplan et l'exemple x_i.

En prenant un point quelconque x_p se trouvant sur l'hyperplan, la marge géométrique peut s'exprimer par :

$$\frac{w}{\|w\|}.(x_i - x_p) \tag{2.16}$$

La marge est au moins égale à la distance entre les deux hyperplans H_1 et H_2 soit $\frac{2}{\|W\|}$ ou $\|w\|$ fait référence à la norme du vecteur w. Maximiser cette marge revient donc à minimiser $\|w\|$ (Marref 2013).

Minimisation quadratique sous contraintes

Maintenant que nous avons défini les notions de marges et d'hyperplans canoniques, nous pouvons formuler un problème d'optimisation mathématique tel que sa solution

nous fournisse l'hyperplan optimal qui permet de maximiser la marge (Boser, et al. 1992)

$$Minimiser\ \frac{1}{2}\|w\|^2 \qquad (2.17)$$

$$Tel\ que\ y_i(\langle w, x_i\rangle + b) \geq 1$$

Il s'agit d'un problème quadratique convexe sous contraintes linéaires de forme primale dont la fonction objective est à minimiser. Cette fonction objective est le carré de l'inverse de la double marge. L'unique contrainte stipule que les exemples doivent être bien classés et qu'ils ne dépassent pas les hyperplans canoniques.

Dans cette formulation, les variables à fixer sont les composants w et b. Le vecteur w possède un nombre de composantes égal à la dimension de l'espace d'entrée.

Généralement dans ce type de cas, on résout la forme duale du problème. Nous devons former ce que l'on appelle le Lagrangien (Pour plus de détails voir la référence (Boser, et al. 1992)).

Cas linéairement non séparable

Nous considérons ici le cas où des exemples sont mal classés par l'hyperplan optimal.

Cela peut résulter du bruit dans les données. Pour résoudre ce problème, Courte et Vapnik (Vapnik et Cortes 1995) ont introduit la notion de ''marge souple'' (*soft margin*) qui correspond toujours à la recherche d'un hyperplan de marge optimale, mais avec une règle d'exception qui autorise que quelques exemples soient à une distance plus faible de l'hyperplan que la marge correspondante.

Pour surmonter cette nouvelle contrainte, nous allons introduire une notion de « tolérance » faisant appel à une technique dite des **variables ressort** (*slack variables*) (Figure 2.9).

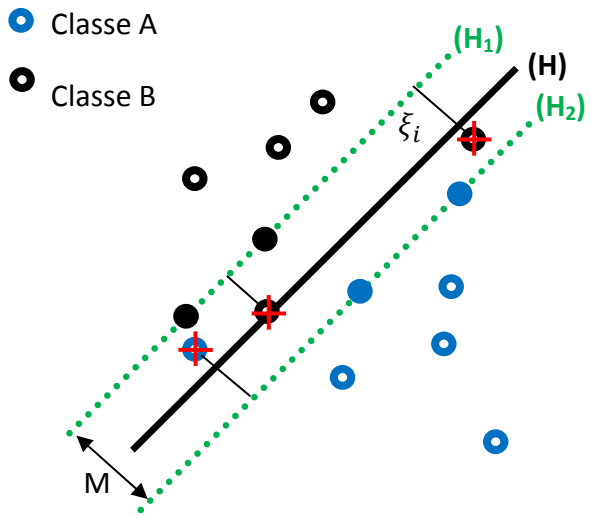

Figure2.10: Exemple graphique des données linéairement non séparable

Les vecteurs de support sont les points remplis et les points d'apprentissage mal classés sont marqués d'une croix.

Encore une fois nous pouvons formuler un problème d'optimisation mathématique tel que sa solution nous fournisse l'hyperplan optimal qui permet de maximiser la marge (V.Vapnik 1995).

$$Minimiser \ \frac{1}{2}\|w\|^2$$

$$tel\ que \quad y_i(w.x_i + b) \geq 1$$

La technique des variables ressort permet de construire un hyperplan en admettant des erreurs, mais en les minimisant, ce qui amène à assouplir les contraintes en introduisant les variables ressort $\xi i \geq 0$ dans la définition des contraintes : (V.Vapnik 1995).

$$w.x_i + b \geq +1 - \xi_i \ \text{ Si } y_i = +1 \qquad (2.18)$$

$$w.x_i + b \leq -1 + \xi_i \ \text{ Si } y_i = -1 \qquad (2.19)$$

Ce qui s'écrit $\quad y_i(w.x_i + b) \geq 1 - \xi_i \quad \forall i = 1 \ldots n \qquad (2.20)$

Quand une erreur de classification intervient, la variable ξi a une valeur plus grande que 1, donc ξi est une borne supérieure du nombre d'erreurs à l'apprentissage.

De là, un moyen naturel pour pénaliser les erreurs est de remplacer la fonction précédente à minimiser par :

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \qquad (2.21)$$

Autrement dit, on cherche à maximiser la marge en s'autorisant pour chaque contrainte une erreur positive ξi la plus petite possible (V.Vapnik 1995). **Le coefficient « C » est défini comme un paramètre de régularisation**, il donne un compromis entre la marge et le nombre d'erreurs admissibles. L'ajout du terme $C\sum_{i=1}^{n}\xi_i$ peut être considéré comme une mesure d'une certaine quantité de mauvaise classification. Ainsi, une faible valeur de C entraîne une faible tolérance. D'autres formulations existent, comme terme $C\sum_{i=1}^{n}\xi_i^2$ (V.Vapnik 1995).

Pour calculer les paramètres w_i *et* b, on utilise le Lagrange comme pour le cas précédent.

Cas non linéaire

Cependant, dans des nombreux cas, les échantillons d'entraînement ne sont pas linéairement séparables. La procédure consiste alors à introduire une fonction ϕ permettant de projeter les données dans un espace de dimension supérieure (*D*) où elles deviennent linéairement séparables, on a un exemple de plongement de $\Re^2$ *dans* $\Re^3$ en utilisant la fonction $\phi(x_1, x_2) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$ (Figure 2.10).

$$\Re^d \rightarrow \Re^D \; avec \; D \gg d$$

$$\vec{x} \rightarrow \phi(\vec{x})$$

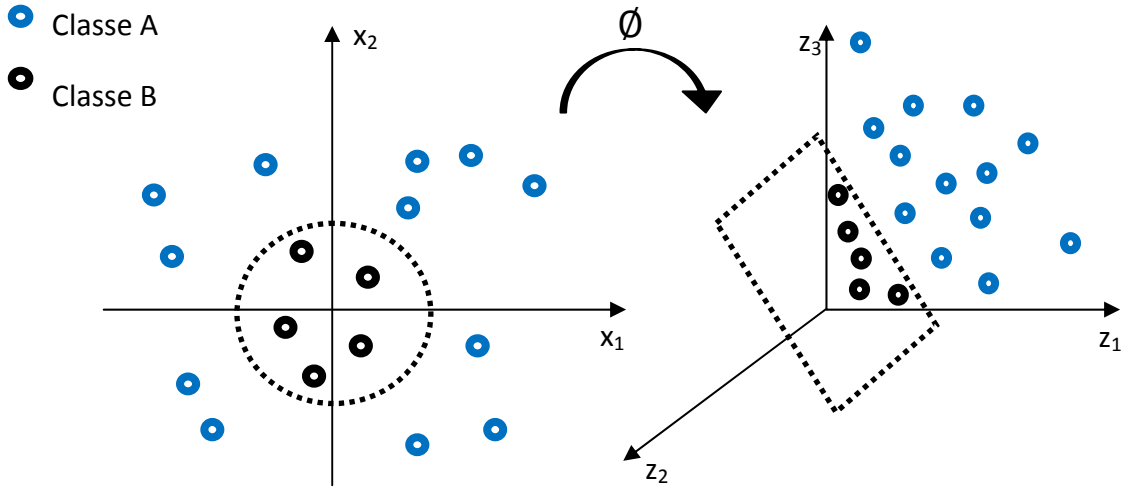

Figure2.11: Exemple de plongement de $\Re^2$ *dans* $\Re^3$

De la même façon que précédemment, on recherche dans ce nouvel espace l'hyperplan optimal donné cette fois par :

$$h(x) = w.\phi(x) + b \qquad (2.22)$$

Une observation importante est que la seule connaissance des produits scalaires entre points est suffisante pour trouver et calculer la fonction h .On n'a donc pas besoin de trouver une représentation $\phi(x)$ explicitement.

Il suffit d'être capable de calculer $\phi(x).\phi(x) = k(x,x)$. Le terme $k(x,x)$ appelé noyau. Vapnik (V. Vapnik 1998) a montré que toute fonction satisfaisant les conditions (symétrique, définie positive) peut être utilisée comme noyau. Parmi les noyaux les plus classiquement utilisés pour la classification :

Le noyau linéaire : $k(x,x) = \langle x,x \rangle$

Le noyau polynomial : $k(x,x) = (a * \langle x,x \rangle + b)^d$

Le noyau Gaussien (*En Anglais: Radial Basis Function* (RBF)) **:**

$k(x,x') = exp\left(-\frac{(x,x')^2}{2\sigma^2}\right)$

2.3.6 Évaluation des classifieurs

Évaluer les performances d'un système de classification est un enjeu de grande importance. Les performances globales de classification sont utilisées dans l'étape d'optimisation des hyper-paramètres du classifieur. Pendant longtemps, le critère retenu pour évaluer ses performances a été le taux de bonne classification, c'est-à-dire le nombre d'éléments d'une base de tests correctement classés (Oufella 2008).

L'évaluation de la classification est basée sur un tableau à deux dimensions, appelé matrice de confusion. Cette matrice trie tous les cas du modèle en catégories, en déterminant si la valeur prédite correspondait à la valeur réelle. Tous les cas dans chaque catégorie sont ensuite comptés et les totaux sont affichés dans la matrice. Les mesures que nous allons évoquer dans cette section utilisent la matrice de confusion.

- **Matrice de confusion**

Tableau 2. 6*:* Matrice de confusion

	Décision Positive	**Décision Négative**	
Étiquette Positive	Vrais Positifs (VP)	Faux Négatifs (FN)	Pos
Étiquette Négative	Faux positifs (FP)	Vrais Négatifs (VN)	Neg
Total (T)	Ppos	Pneg	N

Étiquette Positive : Patients malades.

Étiquette Négative : Patients sains.

Décision Positive : Test était positif.

Décision Négative : Test négatif.

Vrais Positifs (VP) : individus malades réagissent positivement au test.

Vrais Négatifs (VN) : individus sains réagissent négativement au test.

Faux Positifs (FP) : individus sains réagissent positivement au test.

Faux Négatifs (FN) : individus malades réagissent négativement au test.

Le Tableau 2.4 représente la matrice de confusion. Les paramètres de cette matrice sont décrits comme suit (Oufella 2008):

- Taux de Vrais Positifs (TVP)

$$\mathrm{TVP} = \frac{\mathrm{VP}}{\mathrm{Pos}} = \frac{\mathrm{VP}}{\mathrm{VP+FN}} \tag{2.23}$$

- Taux de vrais Négatifs (TVN)

$$\mathrm{TVN} = \frac{\mathrm{VN}}{\mathrm{Neg}} = \frac{\mathrm{VN}}{\mathrm{VN+FN}} \tag{2.24}$$

- Taux de Faux Positifs(TFP)

$$TFP = \frac{FP}{Neg} = \frac{FP}{FP+VN} \tag{2.25}$$

- Taux de Faux Négatifs (TFN)

$$\mathrm{TFN} = \frac{\mathrm{FN}}{\mathrm{Pos}} = \frac{\mathrm{FN}}{\mathrm{FN+VP}} \tag{2.26}$$

- Taux de bonne classification ou l'exactitude (TBC)

$$\mathrm{TBC} = \frac{\mathrm{VP+VN}}{\mathrm{VP+FN+VN+FP}} \tag{2.27}$$

- Précision (P)

$$\mathrm{P} = \frac{\mathrm{VP}}{\mathrm{PPos}} = \frac{\mathrm{VP}}{\mathrm{VP+FP}} \tag{2.28}$$

- Rappel (R)

$$\mathrm{R} = \mathrm{TVP} = \frac{\mathrm{VP}}{\mathrm{Pos}} = \frac{\mathrm{VP}}{\mathrm{VP+FN}} \tag{2.29}$$

- Taux de vrai Positif (Spécificité TVP)

$$\mathrm{TVP} = \frac{\mathrm{VN}}{\mathrm{Neg}} = \frac{\mathrm{VN}}{\mathrm{FP+VN}} \tag{2.30}$$

- Taux de Faux Positif (TFP)

$$\mathrm{TFP} = \frac{\mathrm{FP}}{\mathrm{Neg}} = \frac{\mathrm{FP}}{\mathrm{FP+VN}} \tag{2.31}$$

Maintenant que nous avons défini les paramètres liés à la matrice de confusion, nous allons représenter les performances des systèmes de classification à l'aide de la courbe ROC (En Anglais *Receiver Operating Characteristic*).

- **Courbe ROC**

La courbe ROC représente l'évolution des TVP en fonction des TFP quand on fait varier le seuil de décision (seuil=0 ,5 par défaut). Cette règle fournit une matrice de confusion MC1, et donc deux indicateurs TVP1 et TFP1.

Si nous choisissons un autre seuil (seuil=0.6 par exemple), nous obtiendrons MC2 et donc TVP2 et TFP2 Etc.… MCi, TVPi, TFPi.

L'idée de la courbe ROC est de faire varier le « seuil » de 1 à 0 et, pour chaque cas, calculer le TVP et TFP que l'on reporte dans un graphique : en abscisse le TFP et en ordonnée le TVP.

Construction

Classer les données selon un score décroissant.

Mettre en relation TFP (abscisse) et TVP (ordonnée) (Figure 2.11).

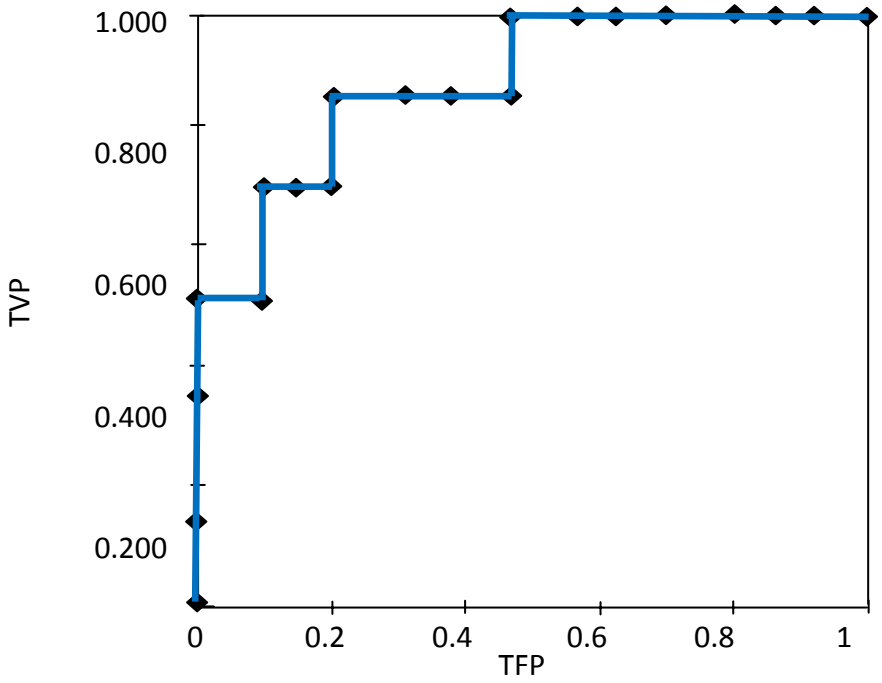

Figure2.12: Courbe ROC

- **AUC, l'aire sous la courbe**

L'aire sous la courbe (abrégée AUC, En Anglais *Area Under Curve*) est la mesure de l'aire de la surface située sous le tracé d'une fonction mathématique dessinée dans un repère. Formellement, cette valeur correspond à l'intégral de cette fonction (Figure 2.12).

Dans le meilleur des cas AUC = 1.

$$s_i = (TFP_i - TFP_{i-1}) \times \frac{TVP_i + TVP_{i-1}}{2} \tag{2.32}$$

Surface d'un trapèze

$$AUC = \sum_i s_i \tag{2.33}$$

AUC=somme (surface des trapèzes)

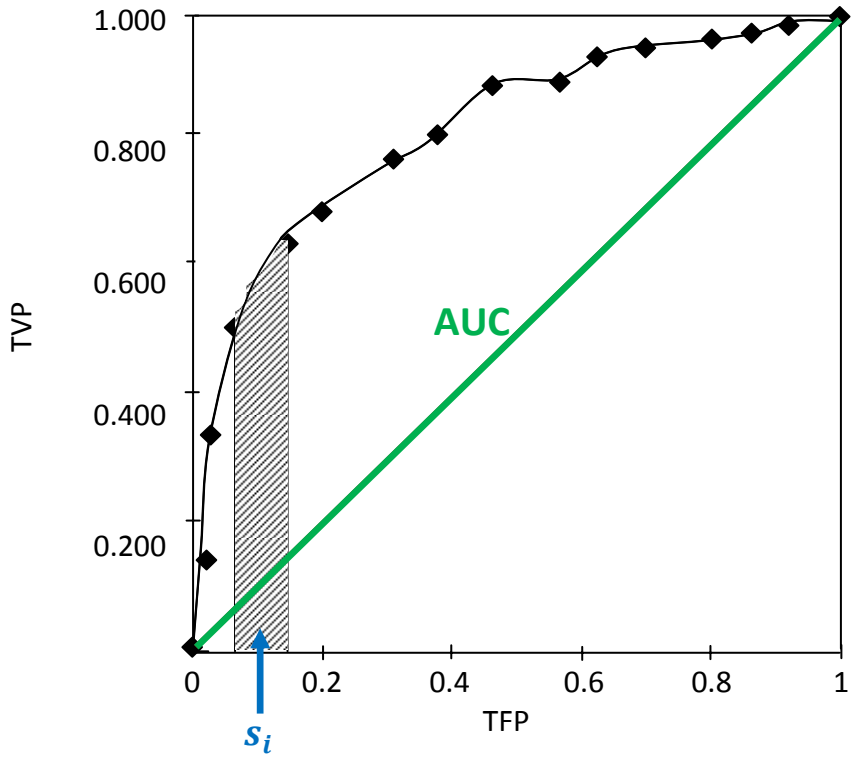

Figure2.13: Courbe ROC avec AUC

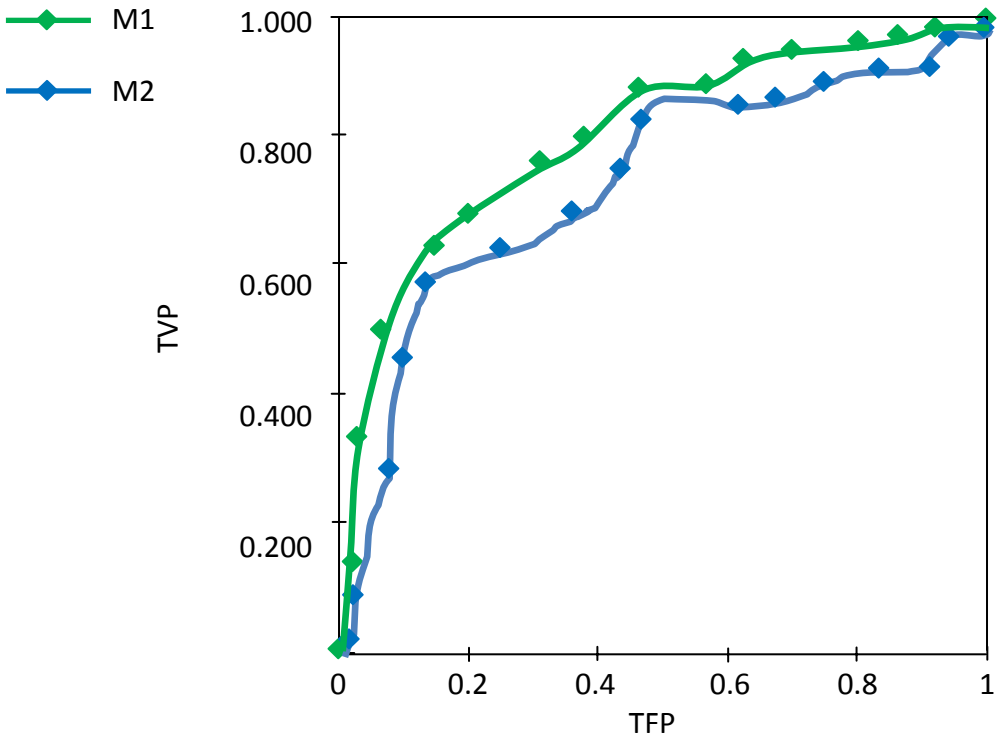

Figure2.14: Comparaison entre les deux courbes

Dans l'exemple de Figure 2.13, on peut observer que la courbe qui correspond au modèle M1 est toujours au-dessus de celle du modèle M2. Le modèle M1 sera toujours meilleur par rapport au modèle M2.

- **Stratégie de validation**

Il s'agit de calculer les erreurs décrites auparavant en se basant sur l'une des stratégies :

Validation croisée

Dans cette approche il y a trois techniques :

- ✓ ***Holdout***

On divise l'échantillon de taille N en un échantillon d'apprentissage et un échantillon de test. Ici il faut faire attention, car on doit réserver la plus grande partie pour l'apprentissage. Généralement cet échantillon contient plus que 60% de l'échantillon principal (Safaa 2013).

- ✓ ***K-fold validation croisée***

La validation croisée est une technique très populaire depuis 40 ans et la plus utilisée pour estimer le risque réel d'un estimateur. La validation croisée se décline en plusieurs sous méthodes. La plus répandue est la méthode « k-Fold » avec typiquement $k\epsilon[4,10]$.Si l'on a une base d'apprentissage A_P contenant P-éléments : $A_p = \{x_1, \dots, x_p\}$ la validation croisée consiste à appliquer les cinq étapes suivantes (Figure 2.14).

1. Diviser l'ensemble d'observations en k sous-ensembles de taille égale.
2. Pour i allant de 1 à k faire :
3. Retenir l'ensemble de numéros i pour le test de performance et faire l'apprentissage sur les k-1 ensembles restants.
4. À chaque itération, estimer l'erreur.

La validation croisée est utilisée dans la résolution d'un grand nombre de problèmes. Elle peut être vue comme une application particulière des méthodes dite de ré-échantillonnage (Cornec 2009).

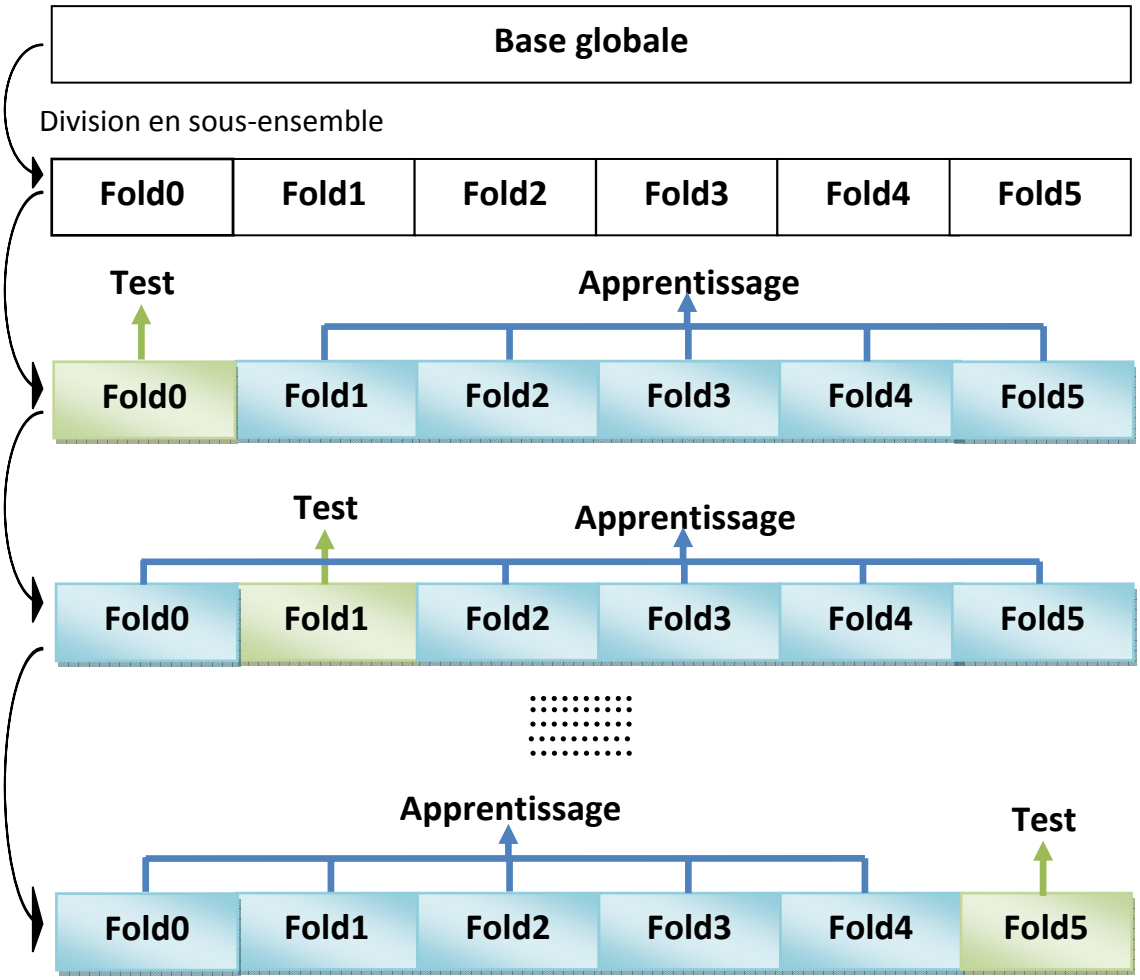

Figure2.15: Principe de la validation croisée

- ✓ ***Leave-one-out***

C'est un cas particulier du k-fold cross validation où k = N (Safaa 2013).

2.4 Conclusion

Dans ce chapitre, nous avons présenté en détail un système de diagnostic de la maladie de Parkinson à partir de la voix. Les étapes de traitement qui sont : l'acquisition, le prétraitement, l'extraction des attributs, la sélection des attributs et la classification ont été analysées. Nous nous sommes intéressés aux méthodes de sélection des attributs dits : Filter. La théorie de la classification à base des SVMs a été présentée dans cette deuxième partie de notre travail. La dernière section de ce chapitre a été consacrée aux méthodes d'évaluation des classifieurs. Le prochain chapitre va être consacré à la description du système de diagnostic que nous avons implémenté ainsi qu'aux résultats expérimentaux.

Chapitre 3 Système de diagnostic, résultats et interprétations

3.1 Introduction

Dans ce chapitre, nous exposerons les résultats obtenus par l'application des SVMs pour le diagnostic de la maladie de Parkinson. La base de données utilisée est issue du serveur UCI ''University of California at Irvin''. La première partie de notre travail consiste à évaluer les performances du modèle SVM en vue d'une classification binaire. Nous évaluerons leurs performances en matière de : taux de classification, matrice de confusion, courbe ROC et l'aire sous cette courbe « AUC ».

Dans la deuxième partie, et afin d'améliorer les performances de classification, une procédure de sélection d'attributs pertinents sera effectuée. Pour accomplir cette partie; nous utiliserons trois techniques de filtrage : ''Fisher'', ''mRMR '' et ''Chi-square''. Nous évaluerons l'influence de la sélection des attributs sur les SVMs et nous comparons les performances des solutions proposées.

3.2 Système de diagnostic

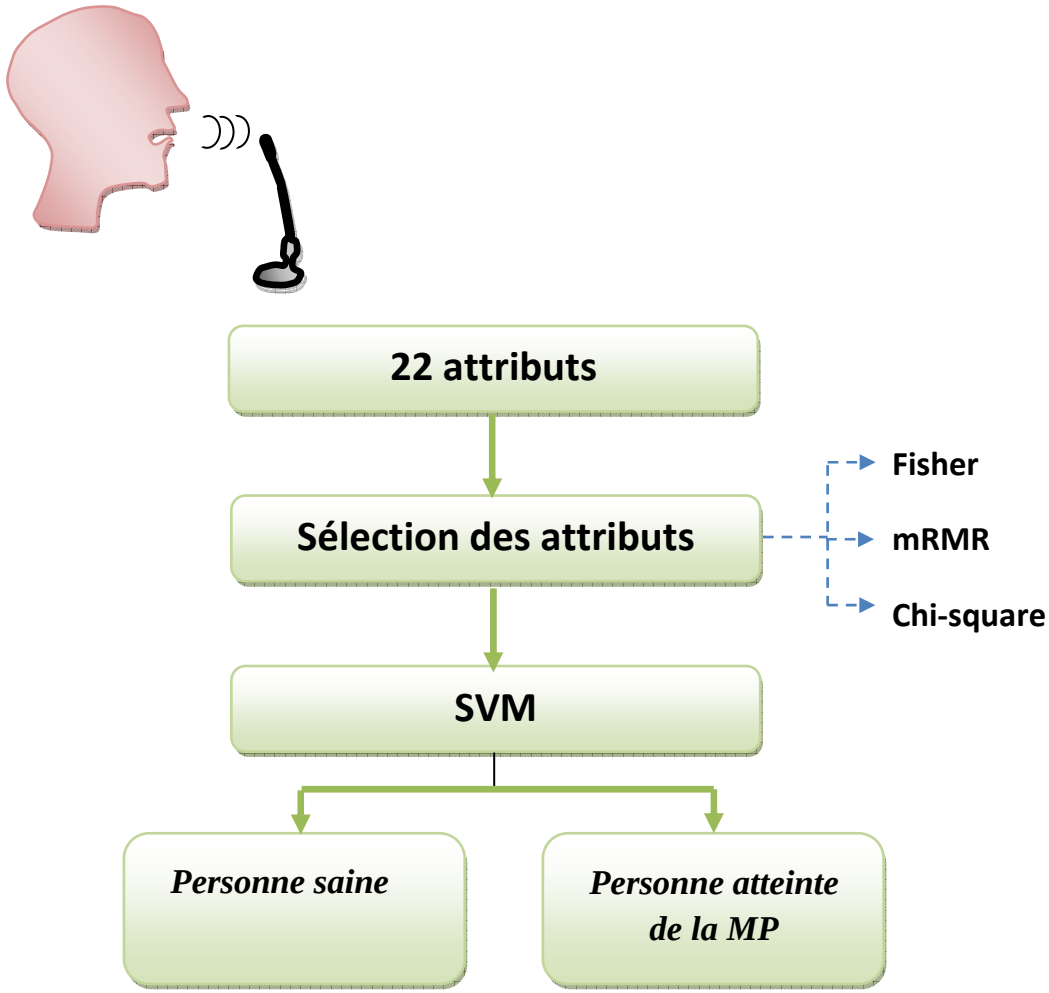

Figure 3.16: Système de diagnostic simplifié

Nous avons utilisé une base de données internationale contenant 22 attributs (voir la section 2.3 du chapitre 2). La sélection des attributs par la méthode de filtrage est basée sur trois techniques : Fisher, mRMR et Chi-square. Le SVM (classifieur) permet de distinguer les sujets atteints de la MP des sujets sain.

3.3 Logiciels Utilisés

o Environnement MATLAB

C'est en exploitant les fonctionnalités de l'Environnement de Développement Intégré (EDI) MATLAB que les principaux travaux de ce projet ont été réalisés.

C'est la version 2012 qui a été utilisée (MATLAB R2012a). Un grand nombre de fonctions dédiées aux calculs numériques, au traitement de signal et aux statistiques

est disponible. Le choix de l'environnement MATLAB a été fait en raison de la rapidité des calculs et la facilité de développement sous ce dernier.

a ***Statistics Toolbox***

Les toolboxes sont réellement des caisses à outils comportant une collection de fonctions relatives à plusieurs domaines scientifiques et techniques, nous avons utilisé la toolbox statistics. Cette toolbox fournit des algorithmes et des outils d'apprentissage statistique pour l'organisation, l'analyse et la modélisation des données.
Pour l'analyse des données multidimensionnelles, Statistics Toolbox inclut des algorithmes permettant la mise en œuvre des modèles. La boîte à outils fournit aussi des algorithmes d'apprentissage, supervisées ou non, y compris les machines à vecteurs de support (SVMs).

b ***Algorithmes de sélection d'attributs***

Nous avons utilisé la boite d'outils « *Feature Selection Algorithms* » développée par l'université d'ARIZONA. Elle est téléchargeable via le site de l'université (Feature Selection Algorithms s.d.).

3.4 Evaluation du système global

Nous avons subdivisé la base de données en deux parties.

- Une partie apprentissage qui contient 75% des données.
- Une partie de test qui contient 25% des données.

o ***Partie apprentissage***

La partie d'apprentissage consiste à :

- Construire le modèle de prédiction,
- Sélectionner le nombre d'attributs,
- Choisir la technique de sélection (Fisher, mRMR ou Chi-square),
- Confirmer la bonne sélection en utilisant la méthode 10-fold cross validation. Cette étape permet aussi de chercher les paramètres C et gamma qui donnent le meilleur taux de classification.

- ***Partie de test***

Avant l'opération de test. Le modèle SVM gaussien est ré-estimé en utilisant toutes les données d'apprentissage (75% de l'ensemble de données représentées sur la figure 3.2). Le test est effectué en utilisant les 25% des données restantes. La sélection des attributs consiste à utiliser l'une des techniques de sélection décrite dans la partie précédente.

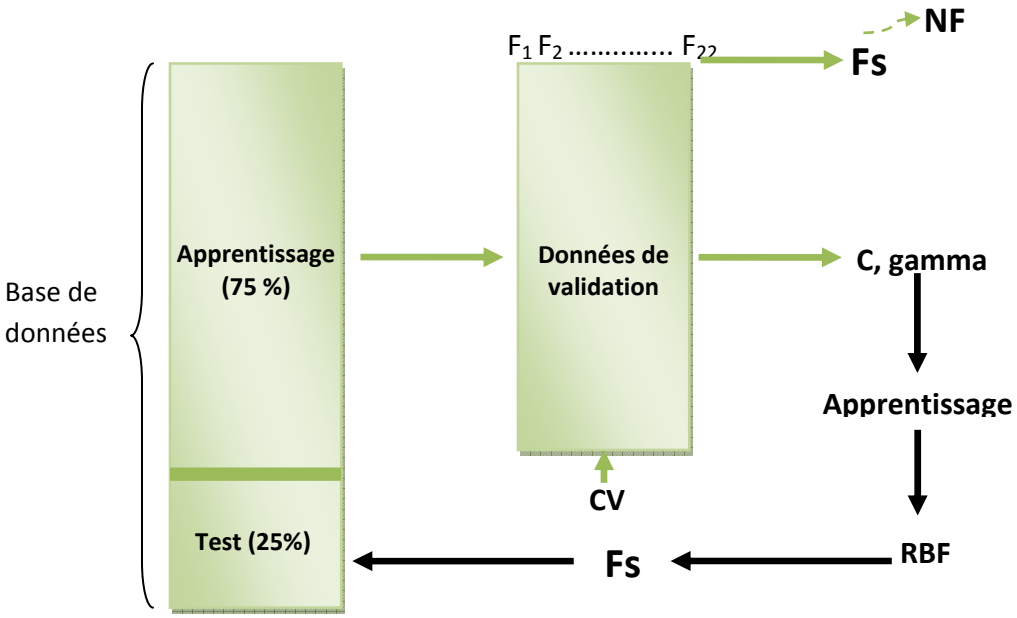

Figure 3.17: Schéma simplifié de la méthode d'évaluation du système global

NF : le nombre d'attributs.

F_s : la technique de sélection d'attributs.

CV : la validation croisée (10-fold cross validation).

RBF : *Radial Basis Function* (SVM avec un noyau gaussien).

3.5 Résultats et interprétations

3.5.1 Estimation des paramètres du modèle SVM

Pour évaluer un modèle SVM non linéaire, nous avons choisi le noyau Gaussien (RBF), nous devons tous d'abord optimiser les paramètres C et gamma.

Les intervalles pour C et gamma sont respectivement fixés entre [-5 5] et [15 3].

La validation croisée est appliquée pour chaque valeur de C et gamma de l'intervalle afin d'obtenir la précision. Parmi plusieurs valeurs de C et gamma, la combinaison qui donne la meilleure précision est retenue.

3.5.2 Sélection d'attributs

Le nombre d'attributs (NF) utilisé dans notre expérimentation varie entre 2 et 22. Pour chaque valeur de NF, l'optimisation des paramètres de notre système est effectuée en se basant sur trois techniques de sélection :

- Fisher
- mRMR
- Chi-square

Le but de cette partie est d'étudier l'influence de la sélection des attributs (nombre d'attributs et technique de sélection) sur un SVM non linéaire (SVM à noyau gaussien(RBF)).

3.5.3 Mesure de performances

Dans un milieu médical, pour évaluer les performances des systèmes de décision, on utilise les mesures suivantes : taux de classification, matrice de confusion, courbe ROC et l'aire sous cette courbe « AUC ».

Tableau 3. 7: Taux de bonne classification (en %) des trois les techniques de sélection.

Techniques / N° attributs	Fisher	mRMR	Chi-squar
02	0.8125	0.7917	0.8125
03	0.875	0.8542	0.7708
04	0.8958	0.8542	0.8958
05	0.8125	0.8333	0.8750
06	0.9167	0.9167	0.9167
07	0.8958	0.9167	0.8959
08	0.9167	0.8750	0.8958
09	0.9375	0.8542	0.9167
10	0.9375	0.8542	0.8750
11	0.9167	0.8750	0.9167
12	0.9375	0.8750	0.8958
13	**0.9583**	**0.9375**	0.8958
14	0.9583	0.8958	**0.9375**
15	0.9375	0.9167	0.9375
16	0.9583	0.8958	0.9375
17	0.9583	0.9375	0.9375
18	0.9375	0.9167	0.9375
19	0.9583	0.9375	0.9167
20	0.9375	0.9375	0.9375
21	0.9375	0.9375	0.9375
22	0.9375	0.9375	0.9375

A la lumière de nos résultats (Tableau 3.1), nous remarquons que :

Le nombre d'attributs permettant d'obtenir une **meilleure précision** est différent pour les trois techniques (13 attributs pour la technique Fisher, 13 attributs pour mRMR et 14 attributs pour Chi-square).

L'utilisation d'un nombre d'attributs supérieur à 13 ou 14 attributs (selon la technique) n'améliore pas la précision, au contraire cela va dans le sens de sa réduction.

La restriction du nombre d'attributs utilisés (13 ou 14 au lieu 22) permet un gain de temps et une limitation des calculs. La combinaison des différents attributs n'est pas identique pour les trois techniques et ce pour le même nombre d'attributs considérés.

A fin de sélectionner le meilleur sous ensemble des attributs, nous allons étudier l'évaluation de performances de la classification (TBC) en fonction du nombre des attributs sélectionnés par les différentes techniques. Les résultats obtenus sont représentés sur les Figures suivantes :

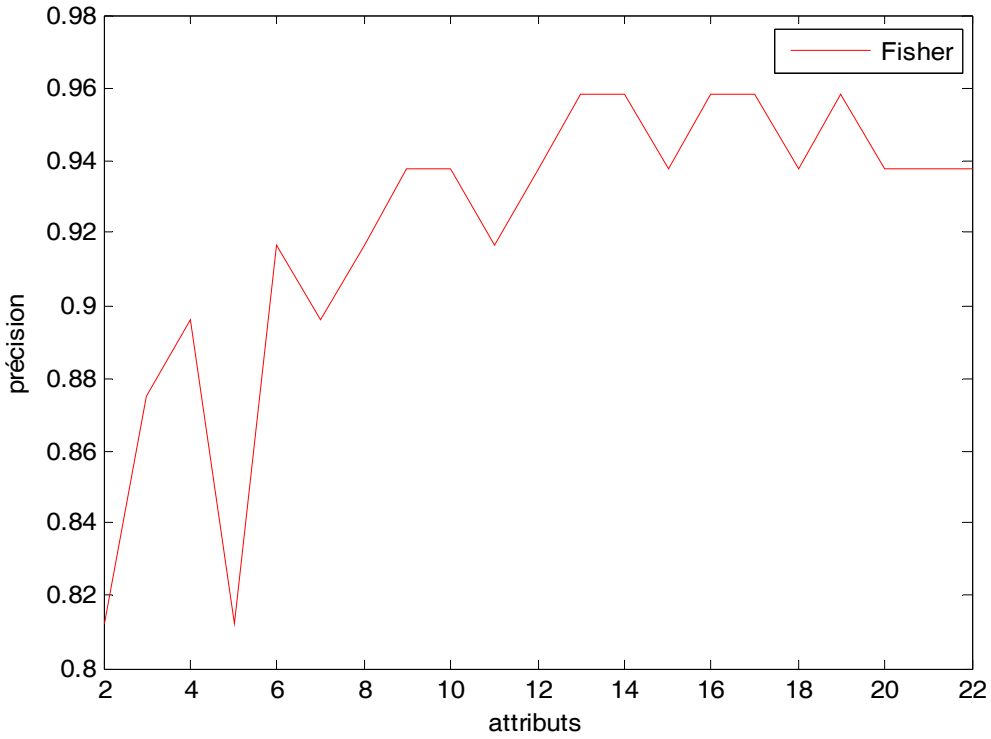

Figure 3.18: Evolution du taux de bonne classification en fonction du nombre d'attributs sélectionnés en utilisant la technique *Fisher*

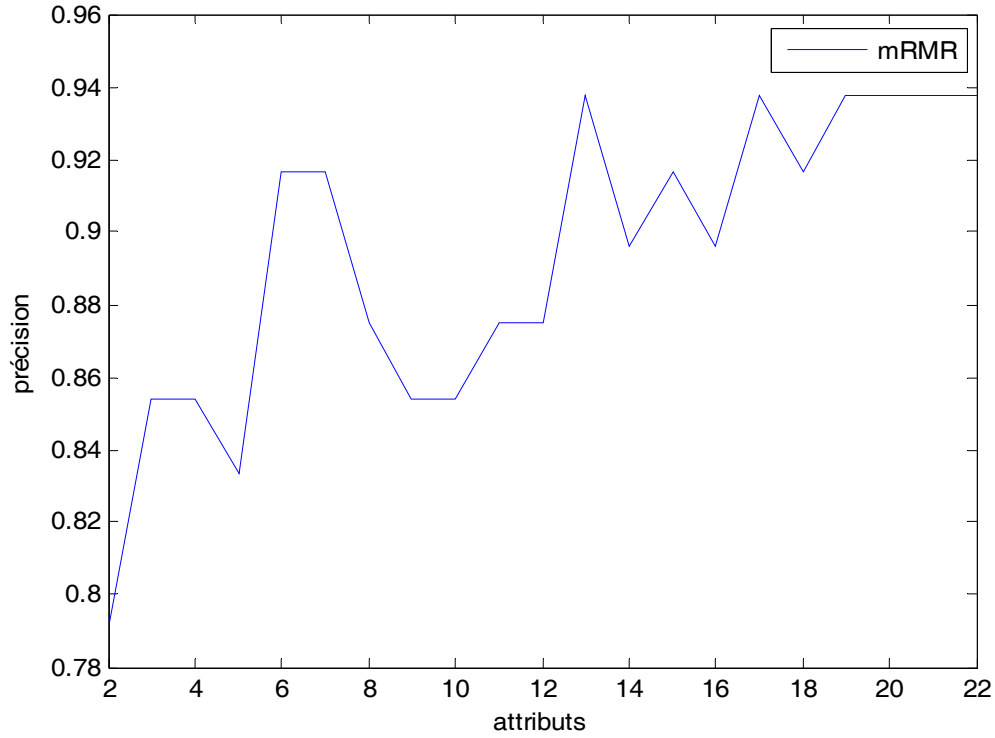

Figure 3.19: Evolution du taux de bonne classification en fonction du nombre d'attributs sélectionnés en utilisant la technique mRMR

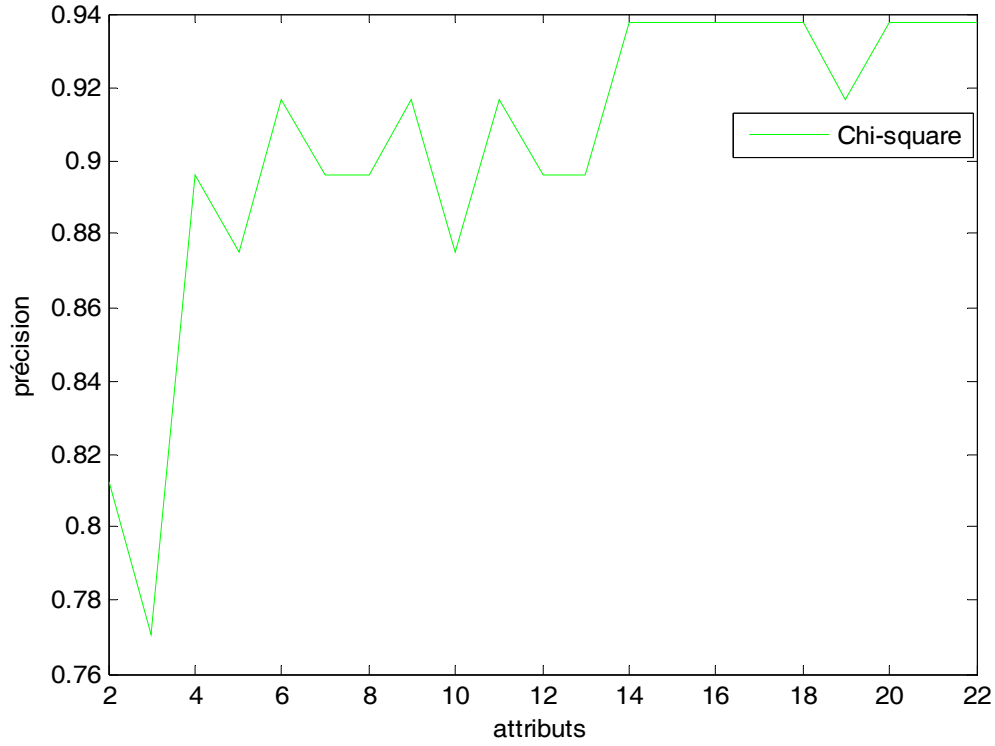

Figure 3.20: Evolution du taux de bonne classification en fonction du nombre d'attributs sélectionnés en utilisant la technique Chi-square

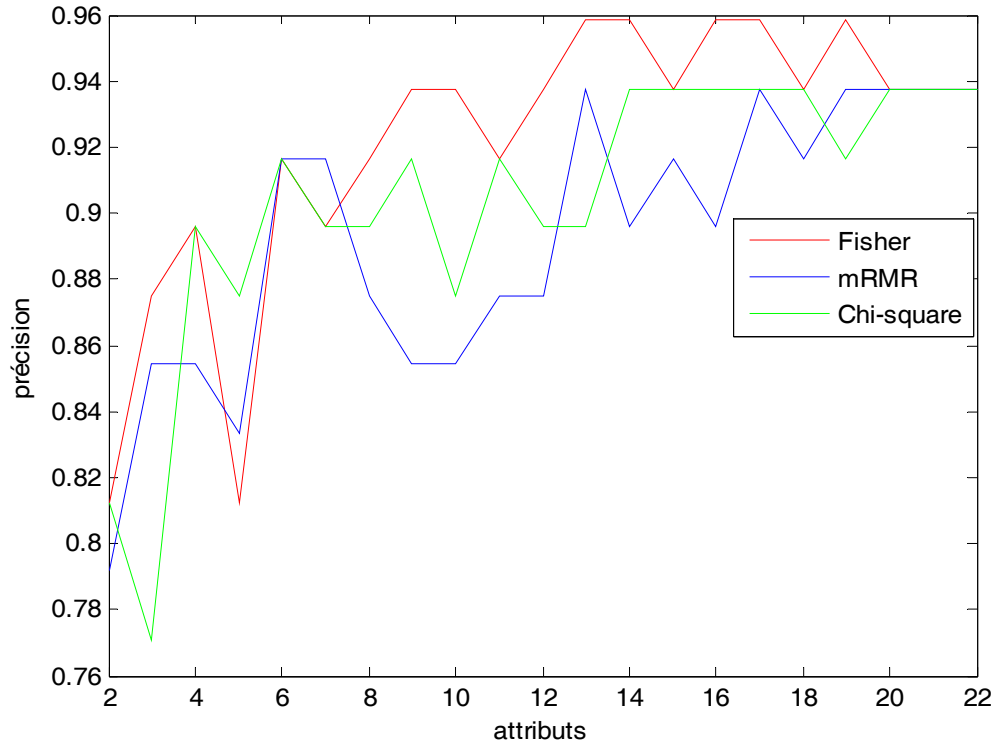

Figure 3.21: Evolution les taux de bonne classification en fonction du nombre d'attributs sélectionnés par les trois techniques (Fisher, mRMR et chi-square)

3.5.4 **Matrice de confusion, courbe ROC et AUC**

Dans cette partie nous utilisons la matrice de confusion, la courbe ROC et l'AUC pour évaluer et comparer les trois techniques que nous choisissons.

a ***Technique Fisher***

La première mesure à laquelle nous allons nous intéresser est le taux de bonne classification (TBC), nous calculons cette valeur pour 13 attributs. Par la suite, nous représentons la matrice de confusion et l'AUC.

Tableau 3. 8: Matrice de confusion obtenue avec la technique Fisher

	Décision Positive	**Décision Négative**
Étiquette Positive	35 (VP)	1 (FN)
Étiquette Négative	1 (FP)	11 (VN)
Total (T)	36	12

$$\text{TBC} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FN} + \text{VN} + \text{FP}} = \frac{35 + 11}{35 + 1 + 11 + 1} = \mathbf{0.9583}$$

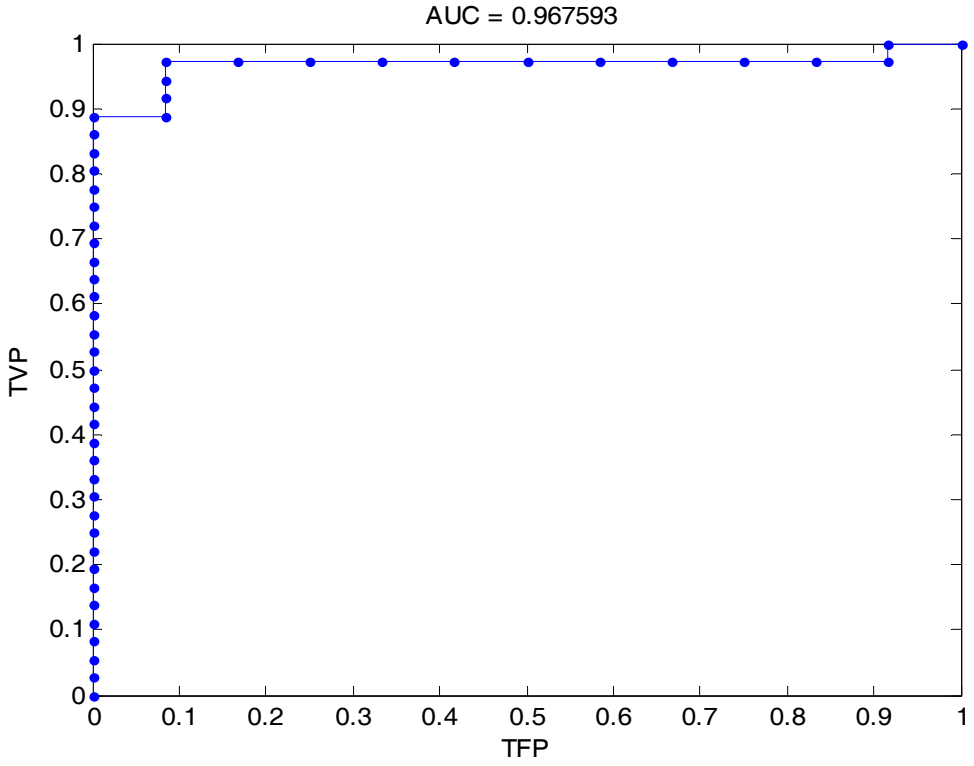

Figure 3.22: Courbe ROC et AUC obtenus par la technique Fisher.

b ***Technique mRMR***

La première mesure à laquelle nous allons nous intéresser est le taux de bonne classification (TBC), nous calculons cette valeur pour 13 attributs. Par la suite, nous représentons la matrice de confusion et l'AUC.

Tableau 3. 9: Matrice de confusion obtenue avec la technique mRMR

	Décision Positive	**Décision Négative**
Étiquette Positive	34 (VP)	1 (FN)
Étiquette Négative	2 (FP)	11 (VN)
Total (T)	36	12

$$\mathrm{TBC} = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{FN} + \mathrm{VN} + \mathrm{FP}} = \frac{34 + 11}{34 + 1 + 11 + 1} = 0.9375$$

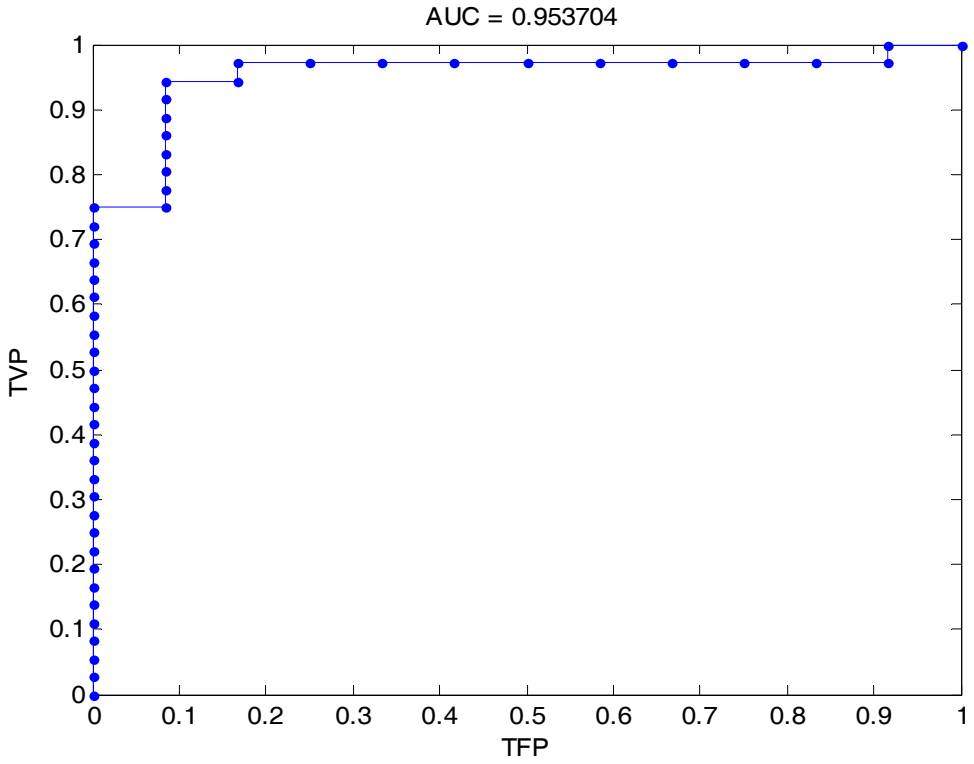

Figure 3.23: Courbe ROC et AUC obtenus par la technique mRMR

c ***Technique Chi-square***

La première mesure à laquelle nous allons nous intéresser est le taux de bonne classification (TBC), nous calculons cette valeur pour 14 attributs. Par la suite, nous représentons la matrice de confusion et l'AUC.

Tableau 3. 10: Matrice de confusion obtenue avec la technique Chi-square

	Décision Positive	**Décision Négative**
Étiquette Positive	35 (VP)	2 (FN)
Étiquette Négative	1 (FP)	10 (VN)
Total (T)	36	12

$$\text{TBC} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FN} + \text{VN} + \text{FP}} = \frac{35 + 10}{35 + 2 + 10 + 1} = 0.9375$$

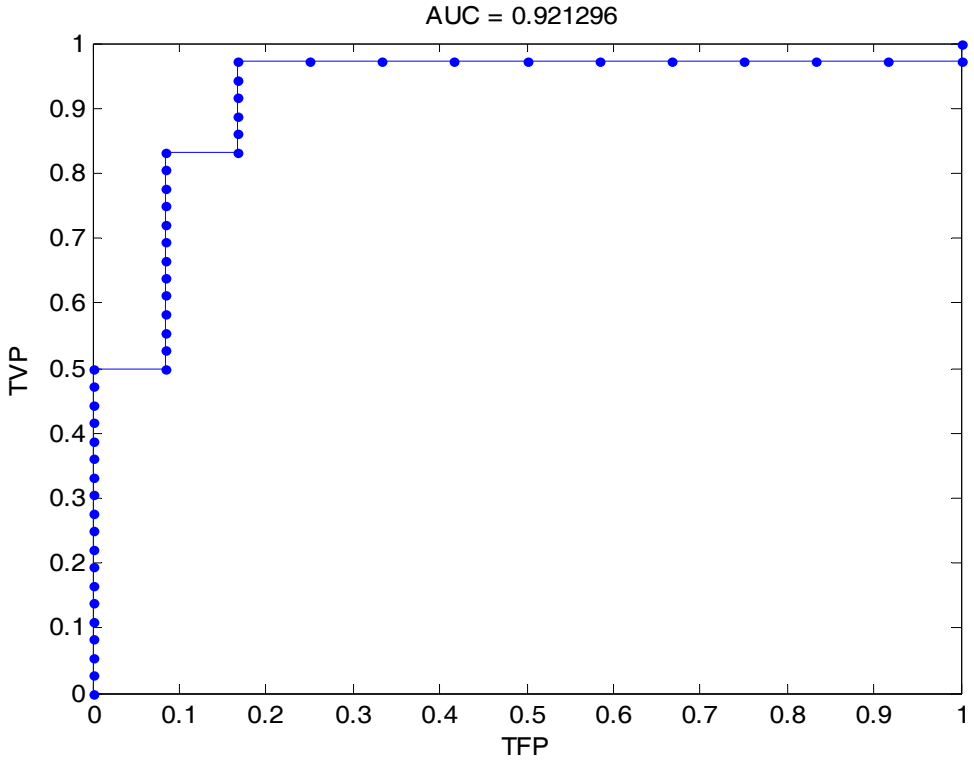

Figure 3.24: Courbe ROC Avec AUC technique de Chi-square

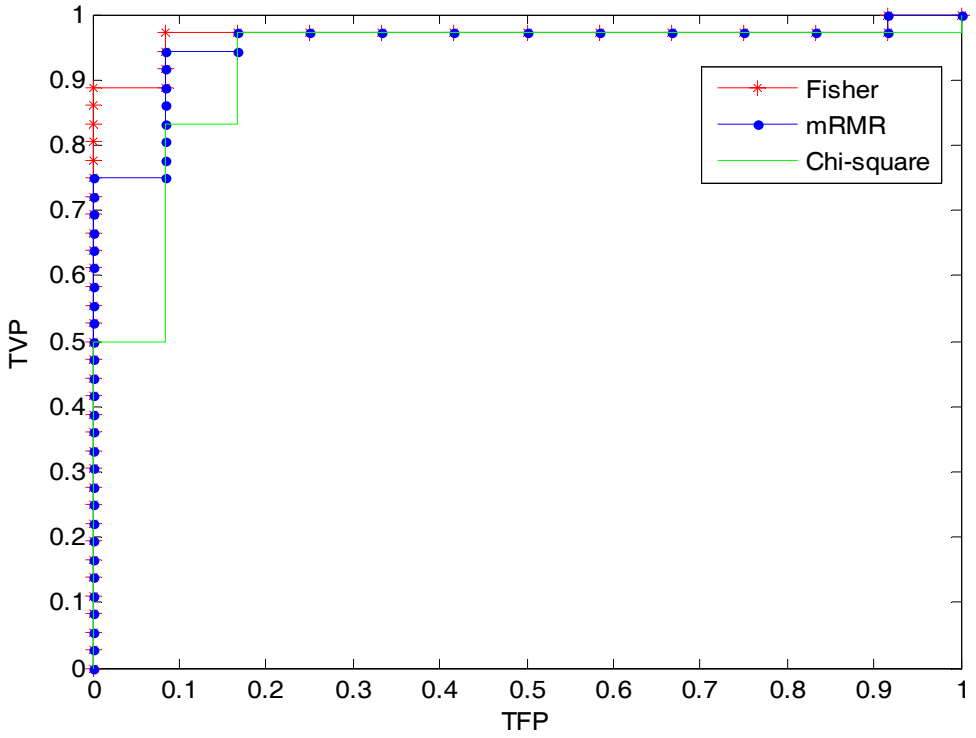

Figure 3.25: Courbes de ROC obtenues par les trois techniques

Nous avons effectué une comparaison entre trois techniques d'apprentissage en vue de la conception d'un système d'aide au diagnostic de la MP à partir de la voix.

Les résultats obtenus dans la Figure 3.10 rejoignent ceux obtenus dans le Tableau 3.1 et démontrent que la technique de Fisher demeure la meilleure technique.

3.6 Conclusion

Les tests que nous avons effectuées nous ont permis de sélectionner une technique d'apprentissage optimale qui eu sélectionnant un sous ensemble d'attributs pertinents qui donne une meilleure précision. Pour la sélection des attributs, nous avons étudié trois techniques de filtrage : Fisher, mRMR et Chi-square. Chaque technique donne le meilleur résultat en utilisant une combinaison propre a chacune une d'entre elle. Les différents tests sont évalués en utilisant les taux de classification, la matrice de confusion, la courbe ROC et l'AUC. Les résultats obtenus démontrent que la technique de Fisher demeure la meilleure technique pour une précision optimale avec un nombre réduit d'attributs.

Conclusions et perspectives

Le travail que nous avons présenté s'inscrit dans le cadre de l'apprentissage statistique et s'intéresse essentiellement au problème de la classification binaire. Notre objectif est de détecter la maladie de Parkinson à partir de l'onde acoustique de la voix.

Le classifieur utilisé (*Support Vector Machine* (SVM) avec un noyau RBF) a permis de mettre en œuvre notre système. Nous avons utilisé une base de données des voix Parkinsoniennes issues du serveur UCI (*University of California at Irvin*). Dans un premier temps, nous avons estimé les performances des SVMs en calculant le taux de bonne classification. Ensuite, une procédure de sélection d'attributs a été effectuée afin de réduire le volume de l'information à traiter et par conséquent de réduire le temps de calcul et la complexité de classification. Les techniques utilisées pour cette tâche sont ''Fisher, mRMR et Chi-square''. Ces techniques octroient pour chaque attribut un score de pertinence puis ordonnent l'ensemble des attributs dans un ordre décroissant. La sélection d'un sous ensemble d'attributs se fait par la validation croisée. Le sous ensemble choisi est celui pour lequel le taux de bonne classification est maximal. Une comparaison des techniques de sélection d'attributs est effectuée. Nous pouvons conclure que parmi les trois techniques utilisées, la technique de Fisher nous a permis d'obtenir la meilleure précision en utilisant seulement 13 attributs.

Enfin, le système testé dans notre étude pourra, dans le futur, faire l'objet d'une application Androïde qui serait accessible et largement utilisée en vue de détecter les premiers signes de la maladie de Parkinson par le biais d'un outil de l'utilisation courante (Exemple : Téléphone portable).

Bibliographie

Abibullaev, B, Won-Seok Kang, et Seung Hyun Lee. «Classification of cardiac arrhythmias using biorthogonal and support vector machines.» *6th International Conference on Networked Computing .* 2010. 24-34.

Amélie, L, et T Aurélie. *Impact d'une rééducation vocale intensive sur la dysprosodie parkinsonienne.* MEMOIRE présenté pour l'obtention du CERTIFICAT DE CAPACITE D'ORTHOPHONISTE, Lyon (France): Université Claude Bernard, 2011.

Auer, et P Dorothee. «In vivo imaging markers of neurodegeneration of the substantia nigra.» *Experimental gerontology* 44, n° 1 (2009): 4-9.

Benikhlef, S, El Bendimerad, et N Settout. *Extraction des caractéristiques pour la classification de la maladie de Parkinson.* Rapport d'activité ,hal-00846805, Tlemcen (Algérie): Université Abou Bekr Belkaid, 2013.

Berg, D, J Godau, et U Walter. «Transcranial sonography in movement disorders.» *The lancet neurology* (Centre de neurologie) 7, n° 11 (2008): 1044-1055.

Bernard, M. «Maladie de Parkinson une maladie aux multiples visages.» *Journal de la fondation recherchemédicale*, 2008.

Bogdanov, M, et al. «Metabolomic profiling to develop blood biomarkers for Parkinson's disease.» *Brain : a journal of neurology* 131, n° 2 (2008): 389-396.

Bohnen, I Nicolaas, L Roger, A Robert, et al. «Positron emission tomography of monoaminergic vesicular binding in aging and Parkinson disease.» *Journal of cerebral blood Flow & metabolism* 26, n° 9 (2006): 1198–1212.

Booij, J, et R.J Knol. «SPECT imaging of the dopaminergic system in (premotor) Parkinson's disease.» *Parkinsonism & related disorders* 13 (2007): 425-428.

Boser, E Bernhard, M Isabelle, Guyon, et Vladimi. «A training algorithm for optimal margin classifiers.» *In Proceedings of the fifth annual workshop on computational learning theory*, 1992.

Brighina, L, A Prigione, B Begni, et al. «Lymphomonocyte alpha synuclein levels in aging and in Parkinson disease.» *Neurobiology of aging* 31, n° 5 (2010): 884-885.

Charles, B. «Diagnostic automatique de la maladie de Parkinson.» Mémoire d'ingénieur en Telecom. Option: Traitement et applications de l'Image, Paris (France), 2005.

Cherif, H, et Ikram. *Classification des tracés TocoGraphiques (CTG) d'un foetus à l'aide de classifieurs multiples.* Mémoire de Master en informatique, Tlemcen (Algérie): Université Aboubaker Belkaid, 2011.

Chouaib, H. *Sélection de caractéristiques: méthodes et applications.* Thèse de Doctorat, Paris (France): Université Paris Descartes, 2011.

Cornec, Matthieu. *Validation croisée et modèles statistiques appliqués.* Thèse de Doctorat, Paris (France): Université Paris X, Nanterre, 2009.

Emborg, ME. «Evaluation of animal models of Parkinson's disease for neuroprotective strategies.» *Journal of neuroscience methods* 139 (2004): 122.

Farrus, M, et J Hernando. «Using Jitter and Shimmer in speaker verification.» *Journal of institution of engineering and technology* 3, n° 4 (2009): 247-257.

Feature Selection Algorithms. ://featureselection.asu.edu/software.php (accès le 08 08, 2014).

Gasser, T. «Genomic and proteomic biomarkers for Parkinson disease.» *Journal of the american academy of neurology* 72, n° 7 (2009): 27-31.

Gilles, A, interviewer par J Caroline. *Les maladies neurologiques* (4 mars 2006).

Haapaniemi, T, V Pursiainen, J Korpelainen, et H Huiku. «Ambulatory ECG and analysis of heart rate variability in Parkinson's disease.» *Journal of neurology, neurosurgery, and psychiatry* 70, n° 3 (2001): 305–310.

Hennecke, G, et CR Scherzer. «RNA biomarkers of Parkinson's disease: developing tool for new therapies.» *Biomarkers in medicine journal* 2, n° 1 (2008): 41-53.

Hilke, R, et al. «Nonlinear progression of Parkinson disease as determined by serial positron emission tomographic imaging of striatal fluorodopa F 18 activity.» *Archives of neurology* 62, n° 3 (2005): 378-382.

Hong, Z, M Shi, KA Chung, et al. «DJ-1 and alpha synuclein in human cerebrospinal fluid as biomarkers of Parkinson's disease.» *Brain journal of neurology* 133, n° 3 (2010): 713-726.

Hwang, H, J Zhang, KA Chung, et al. «Glycoproteomics in neurodegenerative diseases.» *Journal of mass spectrometry* 29, n° 1 (2010): 79-125.

Jiang, JJ, Y Zhang, et C McGilligan. «Chaos in voice, from modeling to measurement.» *Journal of voice* 20 (2006): 2-17.

Kalakech, M. *Sélection semi-supervisée d'attributs : Application à la classification de textures couleur.* Thèse de Doctorat, France: Université Lille1, 2011.

Khalil, R. «Histoir de la maladie de Parkinson.» *Histoire des sciences médicales* 30 (1996): 215-220.

Kumar, Magesh Kumar Udaya. *Classification of Parkinson's disease using multiPass lvq, logistic model tree, k-star for audio data set.* Master thesis, Län (Sweden), Dalarna university, 2011.

Le, W, T Pan, M Huang, et al. «Decreased NURR1 gene expression in patients with Parkinson's disease.» *Journal of the neurological sciences* 204, n° 2 (2008): 29-33.

Lhote, E. *La parole et la voix.* Vol. 37. Allemagne: Hamburger phonetische Beiträge, 1982.

Li, QX, SS Mok, KM Laughton, et al. «Plasma alpha synuclein is decreased in subjects with Parkinson's disease.» *Journal of experimental neurology* 204, n° 2 (2007): 583-588.

Little, M.A, P.E McSharry, E.J Hunter, et J Spielman. «Suitability of dysphonia measurements for telemonitoring of Parkinson's disease.» *IEEE transactions on biomedical engineering* 56, n° 4 (2009): 1015-1022.

Liu, H, et R Setiono. «Chi2: Feature selection and discretization of numeric attributes.» *Seventh international conference on artificial intelligence.* 1995. 388-391.

Mahdjane, K. *Détection d'anomalies sur des données biologiques par SVM.* Mémoire de magister, Tizi ouazou (algérie): Unisersité mouloud mammeri, 2012.

Markaki, M. «Voice pathology detection and discrimination based on modulation spectral features.» IEEE Transactions on audio, speech, and language processing, 19, n° 7 (2011): 1938 - 1948 .

Marref, N. *Apprentissage incrémental et machines à vecteurs supports.* Memoire de magister, BATNA (algérie): Université hadj lakhdar, 2013.

Martin, WR, M Wieler, et M Gee. «Midbrain iron content in early Parkinson disease: a potential biomarker of disease status.» *Journal of neurology* 70, n° 16 (2008): 1411-1417.

Maude, G. *Effets de la stimulation sousthalamique bilatérale sur la voix et la parole de patients parkinsoniens.* Mémoire de maîtrise en médecine, Lausanne (Suisse): Centre hospitalier universitaire vaudois , 2012.

Newton, H Hse, et A Richard. «Sketched symbol recognition using zernike moments.» *Proceedings of the 17th International Conference on Pattern Recognition.* 2004. 367-370.

Oufella, Y. *Évolution du concept de front ROC et combinaison de classifieur.* Mémoire de Master génie informatique, Rouen (France): Université de Rouen, 2008.

Panzacchi, A, et al. «A voxel-based PET study of dopamine transporters in Parkinson's disease.» *Neurobiology of disease* 31, n° 1 (2008): 102-109.

Piyushkumar, A, Mundra, C Jagath, et Rajapakse. «SVM-RFE with relevancy and redundancy criteria for gene selection.» *Pattern recognition in bioinformatics.* 2007. 242-252.

Pützer, M, et J Koreman. «A german database of patterns of pathological vocal fold vibration.» *Journal phonus*, 1997: 143-153.

Quinones, MP, et R Kaddurah-Daouk. «Metabolomics tools for identifying biomarkers for neuropsychiatric diseases.» *Journal neurobiology of disease* 35, n° 2 (2009): 165-176.

Safaa, M. *Analyse non linéaire de la parole pour la détection des voix pathologiques.* Mémoire d'ingénieur, Tunis (Tunisie): Ecole nationale d'ingénieurs, 2013.

Scherzer, CR, AC Eklund, LJ Morse, et al. «Molecular markers of early Parkinson's disease based on gene expression in blood.» *Proceedings of the national academy of sciences of the USA* 104, n° 3 (2007): 955-960.

Schipper, HM, CS Kwok, SM Rosendahl, et al. «Spectroscopy of human plasma for diagnosis of idiopathic Parkinson's disease.» *Biomarkers in medicine journal* 2, n° 3 (2008): 229-238.

Schwingenschuh, P, et al. «Distinguishing SWEDDs patients with asymmetric resting tremor from Parkinson's disease: a clinical and electrophysiological study.» *Journal of movement disorders* 25 (2010).

Settouti, N, et A Hafa. *Approche Filtre pour la sélection des gènes pertinents des données biopuces du Cancer du Côlon.* Mémoire de Master, Tlemcen (Algérie): Université Abou Bekr Belkaid, 2013.

Shi, M, J Bradner, AM Hancock, et al. «Cerebrospinal fluid biomarkers for Parkinson disease diagnosis and progression.» *Journal annals of neurology* 69, n° 3 (2011): 570-580.

Singh, N, V Pillay, et Y.E Choonara. «Advances in the treatment of Parkinson's disease.» *Journal of progress in neurobiology* 81, n° 1 (2007): 29-44.

Soikkeli, R, J Partanen, H Soininen, et A Pääkkönen. «Slowing of EEG in Parkinson's disease.» *Electroencephalography and Clinical Neurophysiology* 79, n° 3 (1991): 159-165.

Spiegel, J, D Hellwig, G Farmakis, W.H Jost, S Samnick, et K Fassbender. «Myocardial sympathetic degeneration correlates with clinical phenotype of Parkinson's disease.» *Movement disorders* 22, n° 7 (2007): 1004-1008.

Stam, CJ. «Use of magnetoencephalography (MEG) to study functional brain networks in neurodegenerative disorders.» *Journal of the neurological sciences* 289, n° 1 (2010): 128-134.

STOESSL, A, et Jon. «Positron emission tomography in premotor Parkinson's disease.» *Parkinsonism & related disorders* 13, n° 3 (2007): 421-424.

Tsanas, A, MA Little, PE McSharry, J Spielman, et LO Ramig. «Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease.» *IEEE transactions on bio-medical engineering* 59, n° 5 (2012): 1264-1271.

Tsanas, T. *Accurate telemonitoring of Parkinson's disease symptom severity using non linear speech signal processing and statistical machine learning.* Doctoral dissertation, Oxford (England): University of Oxford, 2012.

V.Vapnik. *The nature of statistical learning theoy.* Book Chapter : *Information Science and Statistics,* New York, 1995.

Vaillancourt, D.E, M.B Spraker, J Prodoehl, I Abraham, D.M Corcos, et D.M Little. «High-resolution diffusion tensor imaging in the substantia nigra of de novo Parkinson disease.» *Journal of neurology* 72, n° 16 (2009): 1378-1384.

Vapnik, V, et C Cortes. «Support-vector networks.» *Journal of machine learning* 20 (1995): 273-297.

Vapnik, V. *Statistical Learning Theory.* Wiley, New York, 1998.

Westermann, B, E Wattendorf, U Schwerdtfeger, et al. «Functional imaging of the cerebral olfactory system in patients with Parkinson's disease.» *Journal of neurology, neurosurgery, and psychiatry* 79, n° 1 (2008): 19-24.

Yumoto, E, WJ Gould, et T Baer. «Harmonics-to-noise ratio as an index of the degree of hoarseness.» *Journal of the acoustical society of America* 71, n° 6 (1982): 1544-1550.

Zidelmal, A, A Amirou, M Djeddi, et N Djouaher. «Application des SVMs basés sur l'algorithme SMO pour la détection d'anomalies cardiaques.» *4th International Conference: Sciences of Electric, technologies of information and telecommunications.* Tizi ouzou (algérie), 2007.