République algérienne démocratique et populaire
وزارة التـعليـم العالـي والبحـث العلمـــي
Ministère de l'enseignement supérieur et de la recherche scientifique
المركز الجامعي لعين تموشنت
Centre Universitaire Belhadj Bouchaib d'Ain-Temouchent
Institut des Sciences et de la Technologie
Département de Génie Electrique

Projet de fin d'études
Pour l'obtention du diplôme de Master en :
Domaine : SCIENCE ET TECHNOLOGIE
Filière : Génie électrique
Spécialité : électronique biomédicale
Thème

Sélection de variables pour la reconnaissance de la maladie de parkinson

Présenté Par :
1) MANSOUR SOUHILA
2) TABTI ZOHRA

Devant les jurys composés de :

Bendimered .M	C.U.B.B (Ain Temouchent)	Président
Badir. B.L	C.U.B.B (Ain Temouchent)	Encadreur
Berrakem.S	C.U.B.B (Ain Temouchent)	Examinateur
Bouchikhi. S	U.A.B.B (Tlemcen)	Co-encadreur

Année universitaire 2015/2016

Dédicace

Je dédie ce mémoire

A mes chers parents ma mère et mon père

Pour leur soutien, leur amour et leur encouragement.

Ames chers sœurs Yamina et Aziza qui m'ont toujours encouragé

et soutenu moralement.

Je dédie spécialement ce mémoire Amon frère Bouameur qui ma

toujours encourager de continuer mes études.

A Meriem Mimouna le petit bijou de la famille

Mansour Souhila

Dédicace

Je dédie ce modeste travail à deux personnes qui occupent mon cœur.mes parent chérie qui mon soutenu durant toute ma vie, et spécialement à l'étoile qui illumine mes nuits par ses conseils qui font de moi ce que je suis : ma mère qui ma toujours soutenu tout au long de mes études.

A mes deux bougies : Kheira et Hadjer et surtout à la lumière de la famille mon petit frère Mohamed Noureddine a qui je souhaite la réussite.

Tabti Zohra

Remerciement

Nous tenons à remercier avant tous, le bon dieu qui nous a donné la force et la capacité pour préparer ce projet.

Nous tenons à remercier notre encadrant « Mml. Badir B.L» pour le suivi d'intérêt qu'elle a apporté à ce mémoire.

Nous remercions notre co-encadrant « Mml. Bouchikhi.S »pour son aide précieuse et pour le temps qu'elle nous a consacré.

Mes remerciements à nos professeurs ; « Mm Bendimered.M » et « Mml. Berrakem.S » qui ont participé à juger ce travail.

Finalement nous disons merci, merci et mille fois merci à toutes les personnes qui nous ont aidés pour réaliser cette tache.

Résumé

La sélection de variables en classification se pose généralement lorsque le nombre de variables est élevé, Dans ce mémoire, nous proposons des méthodes innovantes pour réduire la taille initiale des données afin de sélectionner les ensembles de variables pertinents pour une classification supervisée.

Notre travail s'inscrit dans le domaine d'aide au diagnostique médical.

Dans ce manuscrit nous nous intéressons à la détection et la reconnaissance de la maladie de Parkinson.

Notre première contribution concerne à proposer deux classifieurs supervisée le SVM (Support Vector Machine), et le KNN (k plus proche voisins) pour évaluer la pertinence des sous-ensembles.

Notre seconde contribution consiste à proposer deux approches de classification par le biais de deux méthodes de sélection de variables relieff et fisher qui sont destinée pour la sélectionner des variables les plus pertinents .

Nos expérimentations nous ont guidées vers l'identification de la maladie de Parkinson en utilisant une sélection de variables basée sur deux approches de classification supervisées (svm & knn).

Mots-clés : Sélection de variables, Classification supervisée, KNN, SVM, Relieff, Fisher maladie de parkinson.

Abstract

The selection of variables in classification arises generally, when the number of variables is high. In this study, we propose some innovative methods to reduce the initial dimension of data in order to select the whole pertinent variables for a supervised classification. Our research work fits into domain help of medical diagnosis. Therefore, in this manuscript, we are interested in the detection and recognition of Parkinson' disease. Our first contribution is concerned with proposing two supervised classifiers: the SVM and the KNN in order to evaluate subset's pertinence. Our second contribution consists in proposing two distinct classification approaches by means of two variable selection methods, namely: 'relieff' and 'fisher' that are intended for the selection of the most pertinent variables. Our experiments have led us to the identification of Parkinson's disease when using two supervised classifiers based upon variables' selection.

Key-words: Variable's selection- Supervised classification- KNN- SVM- Relieff- Fisher- Parkinson's disease.

ملخص

اختيار تصنيف المتغيرات عادة ما تنشا عندما يكون عدد المتغيرات مرتفع.في هذا البحث, نقترح أساليب مبتكرة للحد من الحجم الأولي للبيانات لتحديد مجموعات من المتغيرات ذات الصلة للتصنيف الذي يخضع للإشراف.

عملنا يدخل في مجال التشخيص الطبي .

في هذه المخطوطة نحن نركز على الكشف والتعرف على مرض الشلل الاهتزازي (Parkinson).

أول مساهمة تتعلق باقتراح اثنين من المصنفين تحت إشراف الدعم لناقل الآلة (SVM), و اقرب الجيران (KNN) بهدف تقييم أهمية المجموعات الجزئية.

مساهمتنا الثانية هي اقتراح منهجيتي التصنيف من خلال طريقتين لاختيار المتغيرات Fisher و Relief التي تهدف إلى تحديد المتغيرات الأكثر أهمية.

وجهت تجاربنا لتحديد مرض الشلل الاهتزازي (Parkinson) باستخدام اثنين من المصنفين تحت إشراف استنادا إلى التحديد متغير.

كلمات البحث: اختيار متغير، تصنيف الذي يخضع للإشراف،KNN,SVM , Relief, Fisher,مرض الشلل الاهتزازي.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

MP Maladie de parkinson.

K-NN (Kppv) K plus proche voisins.

SVM Support Vector Machine.

TC Taux de Classification.

SE Sensibilité.

SP Spécificité.

SV Sélection de Variables.

VN Vrai Négatif.

VP Vrai Positif.

FP Faux Positif.

FN Faux Négatif.

Introduction générale

Dans le domaine médical la résolution des problèmes d'aide au diagnostique se base sur le traitement de données extraites à partir des données acquises dans le monde réel.
Parmi les maladies qui sont en voie de développement ces dernières décennies est la maladie de Parkinson, pour cela on a conçu un système d'aide à la prise de décision concernant cette maladie.
Le présent travail que nous présentons s'inscrit dans le contexte de la sélection de variables plus particulièrement les gènes de la maladie de parkinson qui est devenu une maladie d'actualité, tout en permettant le développent d'outils d'aide au diagnostic pour la prédiction de l'état du patient (malade ou sain).
Cette méthode de sélection a pour but de réduire le nombre de variables afin de laisser les plus informatives ayant un poids fort, dans l'optique de ce travail on se base plus particulier sur deux approches de classification supervisée KNN et SVM linéaire pour l'identification de la maladie de parkinson afin d'assurer une bonne performance de ce système.
Nos principales contributions s'inscrit particulièrement autour de la proposition de deux classifieurs supervisée pour la reconnaissance de la maladie de Parkinson d'une part et d'autre part l'utilisation des méthodes de la sélection de variable, dans le but d'augmenter les performances pour sélectionner les variables les plus pertinents.
Les travaux menés dans le cadre de ce projet de fin d'étude et les résultats obtenus sont structurés en quatre chapitres.
Le premier chapitre est une introduction sur les notions médicales de la maladie objet de notre étude. On introduit tout d'abord les symptômes de cette maladie ainsi que les principales causes de cette dernière.
La deuxième chapitre décrit les différents concepts théoriques des outils et approches de classification supervisée utilisés pour développer notre système d'aide au diagnostique.
Le troisième chapitre est consacré pour la présentation de différentes techniques et méthodes de sélection de variables que nous avons utilisés.

Le quatrième chapitre est dédié à la contribution relative à la mise en œuvre d'un système de reconnaissance de la maladie visée. Tout en présentant les résultats obtenues d'une part et d'une autre part nous terminons ce chapitre par une comparaison entre les résultats obtenues par les deux classifieurs supervisés utilisés en matière d'aide à la prise de décision.
Enfin, une conclusion générale synthétise le contenu de ce manuscrit et présente les perspectives de ce thème de recherche.

1 .Introduction

La maladie de parkinson devient une des maladies d'actualité qui affectent la population. C'est la deuxième maladie neurodégénérative la plus fréquente après la maladie d'Alzheimer. Lorsque les individus vivent plus longtemps et que la maladie de parkinson est une maladie de l'âge mur ou de la vieillesse, et il n'existe pas de guérison de cette maladie, elle peut dans certains cas survenir très précocement, parfois touche les adultes moins de 40/50 ans (sujets jeunes).

Le diagnostic de cette pathologie consiste à classer le patient suivant deux situations «parkinsonien ou sain» donc il est important de faire appel aux systèmes de classification et de la sélection des données.

2. Notion médicale

2.1 Définition

La maladie de Parkinson est une maladie chronique neurologique qui se manifeste principalement par des troubles du mouvement. Elle s'explique par la perte de cellules dans une partie du cerveau que l'on nomme la substance noire (ganglion basal). Ces cellules sont responsables de la production d'une substance chimique appelé dopamine, qui agit comme un message entre les cellules du cerveau impliquées dans le contrôle du mouvement. On estime qu'au moment où le diagnostic est prononcé, environ 80 % des cellules produisant la dopamine ont déjà cessé de fonctionner. La diminution de la dopamine entraîne l'apparition des symptômes de la maladie de Parkinson [16] [5].

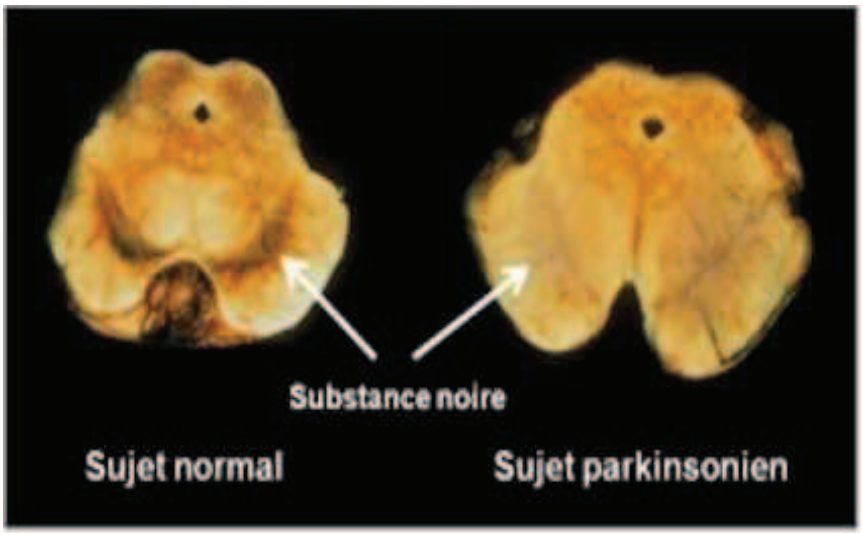

Figure I.2.1 - Coupes de mésencéphale humain illustrant la dépigmentation de la substance noire chez un sujet parkinsonien.

3. Les symptômes de la maladie de parkinson

Les principaux symptômes de cette maladie sont :

La perte de dopamine peut entraîner l'apparition de divers symptômes *moteurs* (du mouvement), notamment :

• Tremblements : de forme plus grave, qui affect les deux cotés du corps, peut altérer la santé s'il interfère avec l'alimentation, provoquant une perte de poids importante, d'abondantes transpiration et des insomnies.
• Rigidité musculaire : est le problème majeur que pose cette maladie .si le tremblement affecte généralement le ou les bras, mais rarement la tête ou les jambes, la rigidité peut atteindre les muscles de la face, le cou, les épaules, le corps, les bras et les jambes.
• Lenteur des mouvements : (acinésie).
• Problèmes d'équilibre [16][2].

Autres symptômes possibles

• Difficulté à écrire (tendance à écrire de plus en plus petit et interdit au sujet certain emploi)
• Élocution lente
• Posture voûtée : Conséquence de la chute de la tête, la salive s'accumule dans la partie frontale de la bouche et, avec la pesanteur, finit par couler continuellement. Le malade avale difficilement les aliments solides et liquides, il perd du poids.

• Expressions faciales réduites (visage moins mobile).
• Troubles de la marche (tendance à traîner les pieds).
• Douleurs musculaires [4].

Aux symptômes moteurs, peuvent s'ajouter des symptômes *non moteurs* notamment

• Constipation
• Troubles du sommeil
• Incontinence urinaire (urgence et fréquence)
• Étourdissements en se levant
• Fatigue
• Dépression : tristesse, manque d'énergie, perte d'intérêt
• Troubles de la mémoire
•dermite séborrhéique (peau plus grasse, particulièrement en bordure du nez et des arcades sourcilières, et aussi cuir chevelu plus gras) [4].

4. Les causes de la maladie de Parkinson

Bien que jusqu'à présent la cause da la maladie de parkinson idiopathique ne soit pas connue, les recherches ont mis en lumière certains de ces mécanismes. On a constaté que

ou produits chimiques s'ils sont administrés a des sujets normaux,provoquent en moins de 2 semaines tous les symptômes de la maladie de parkinson. Le fait que des substances chimiques causent ces symptômes chez le sujet en bonne santé indique que la cause immédiate de la véritable maladie de parkinson doit être une substance chimique, non identifiée jusqu'à présent, qui s'accumule dans le ganglion basal, et qui endommage ou détruit lentement les cellules de cette région du cerveau en provoquant progressivement, tremblement, rigidité et lenteur.

Il est également possible qu'une mauvaise alimentation en sens du ganglion basal est responsable de l'accumulation persistante de substances toxique. Cependant, il est clair que l'artériosclérose ou le durcissement des artères cérébrales, en lui-même, ne porte pas de responsabilité directe, étant donné que des millions d'individus souffrent d'artériosclérose cérébrale avancée, mais ne sont jamais atteints d'une maladie de parkinson.

Il est possible qu'un foie déficient soit responsable de l'accumulation de substances toxiques, ou qu'il existe prédisposition génétique a cette maladie .Il faudra encore que des recherche approfondies se poursuivent pour résoudre ce problème.

Même ainsi, le faite que les symptômes de la maladie de parkinson puissent être provoqués par des substances chimiques donne l'espoir que grâce à la recherche, un brillant chimiste découvre une substance qui agisse comme antidote pour prévenir ou arrêter le progrès des symptômes [16] [5].

5. Diagnostic

La maladie de Parkinson se développe généralement progressivement, comme il peut se passer plusieurs mois, voire d'années, avant que les symptômes ne deviennent assez gênants pour en faire part à son médecin.

Le diagnostic repose sur la description des symptômes par le patient lui-même. C'est pourquoi il est important de consulter un neurologue, le spécialiste de cette maladie, en cas de doute.
Le médecin recherche habituellement deux symptômes ou plus, parmi les principaux que sont la lenteur du mouvement, la raideur ou le tremblement. Un examen clinique est habituellement suffisant. Cependant, il peut être nécessaire de réaliser d'autres tests ou examens et des méthodes d'imagerie médicale, comme la tomographie par émission de positons ou la tomographie d'émission monophotonique, permettent de mettre en évidence la dénervation striatale mais ne permettent pas d'identifier la cause [1] [5].

6. Evolution et pronostic

« Combien de temps un parkinsonien peut vivre dans cet état ? » C'est une question très importante dont La réponse est difficile à formuler, car elle varie d'un patient à un autre. Ceux qui ont la chance d'être entouré de soins affectueuse peuvent vivre jusqu'à 75 ,80 ou 85 ans. A cet âge les artères principales s'obstruent et le malade entre dans un coma dont il ne sortira pas. Certains patients, même totalement infirmes et confinés au lit, peuvent, dans de bonnes condition de soins, vivre encore 5 ans et plus.

La lenteur des mouvements tend à préserver le cœur et la pression sanguine, elle isole l'organisme, en mettant a l'abri de l'usure et de la tension, moins que le patient soit agité déprimé ou affecté de sérieuse tremblement .La duré habituelle de la maladie est de 10 a20 années ou plus selon l'âge du patient au début de sévérité des symptômes. Le tableau (I.6) résume l'évolution typique de la maladie de Parkinson chez un sujet qui est non traité. Ce tableau présente un « portrait » très simplifié de l'évolution typique : la maladie commence d'un seul côté, puis atteint les deux côtés et, éventuellement, des problèmes d'équilibre apparaissent. Cependant, les traitements de la médecine moderne (médicaments ou chirurgies) font en sorte que peu de personnes atteignent aujourd'hui le stade 5 [3] [5].

Stade 1	Les symptômes sont unilatéraux et comprennent au moins deux des trois symptômes suivants : tremblement de repos, raideur et akinésie.
Stade 2	Les symptômes commencent à devenir bilatéraux et peuvent alors inclure des problèmes d'élocution, une posture déformée et des difficultés à marcher.
Stade 3	Les symptômes bilatéraux s'aggravent et des problèmes d'équilibre peuvent apparaître. L'autonomie de la personne n'est généralement pas affectée.
Stade 4	L'invalidité est présente, mais l'autonomie de la personne n'est généralement pas affectée. La bradykinésie est plus prononcée de même que les fluctuations, si elles sont présentes.
Stade 5	Le personne est confinée à un fauteuil roulant ou doit rester alitée.

Tableau I.6 - L'évolution « typique » de la maladie de parkinson [5].

7. La prévalence de la maladie de parkinson

La fréquence de la maladie varie considérablement en fonction de l'âge. Elle est rare avant 50 ans, mais sa fréquence augmente fortement à partir de l'âge de 60 ans. Ainsi, après 60 ans, la prévalence est Comprise entre 12 et 15 pour 1 000 personnes. Des projections estiment que le nombre de personnes, atteintes de MP aura doublé en 2030.

La MP est habituellement plus fréquente chez les hommes que chez les femmes,

En tenant compte des différences d'espérance de vie entre les hommes et les femmes, le risque d'atteindre la MP a été estimé, aux Etats-Unis, comme étant de 2 % chez les hommes et de 1,3 % chez les femmes.

Des différences de prévalence et d'incidence de la MP ont été observées dans différents pays, Une étude collaborative incluant quatre études européennes (Espagne, France, Hollande, Italie) ayant utilisé une méthodologie et des critères de diagnostiques similaires n'a pas mis en

évidence de différence de prévalence entre ces pays. En revanche, un méta analyse de six études retrouve une prévalence plus faible en Afrique qu'en Europe ou en Amérique du nord. D'après une revue d'études menées en Asie, la prévalence de la MP serait légèrement plus faible en Asie par rapport aux pays occidentaux, Toutefois, il est difficile de savoir si ces différences sont dues à des facteurs environnementaux ou des différences d'ordre méthodologique. Une étude utilisé la même méthodologie pour estimer la prévalence de la MP dans le Mississipi (Etats-Unis) chez des Noirs américains et des Caucasiens et au Nigéria., tandis que la prévalence était plus faible au Nigéria.

Il semble donc que dans cette région du nord des Etats-Unis, aucun facteur de risque environnemental de la MP ne soit intervenu. En revanche, dans une étude finlandaise l'incidence de la MP augmentait entre 1971 et 1992 chez les hommes.

Peu de données sur la fréquence de la MP sont disponibles en France (Figure I.7). Une étude en population générale parmi des personnes âgées de 65 ans et plus en Gironde et en Dordogne estime, en 1994, une prévalence de 1 400 cas pour 100 000 personnes dans cette classe d'âge. Une autre étude réalisée en 2000 à partir des données de remboursement de l'Assurance Maladie, incluant les bénéficiaires du régime général âgés de 65 ans et plus, rapporte une prévalence de la MP de 1 250 cas pour 100 000 personnes [6].

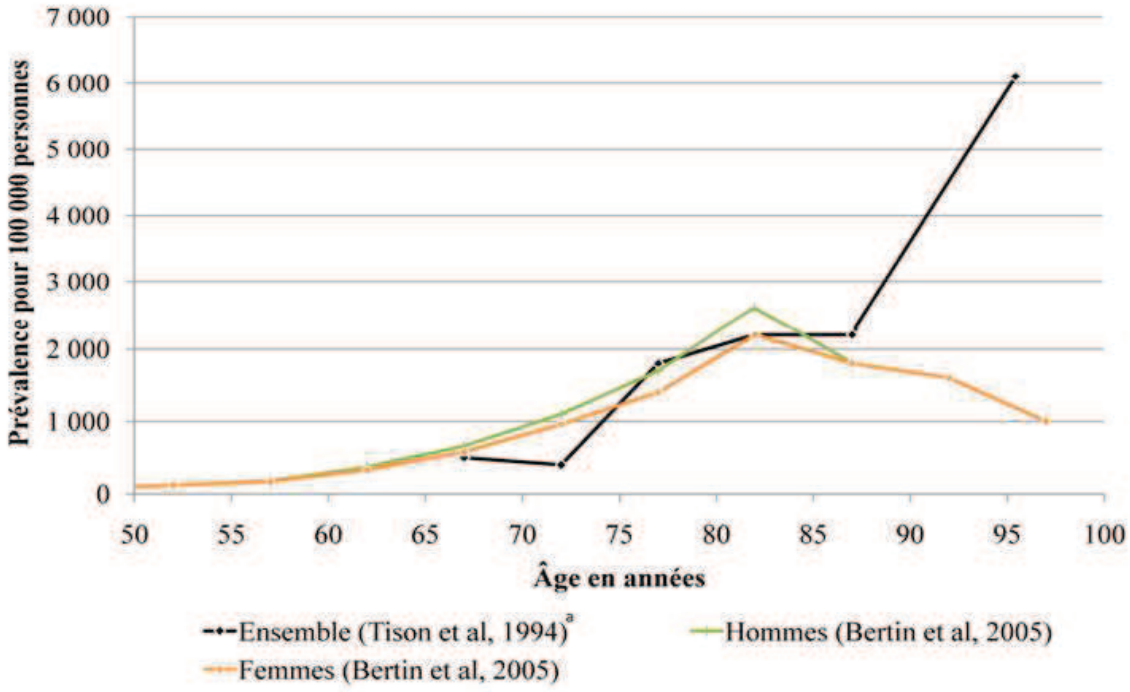

Figure I.7-Prévalence de la MP par âge et par sexe en France [6].

8. Conclusion

Dans ce chapitre, nous avons présenté la maladie de Parkinson, qui est l'une des maladies neurodégénérative, est une pathologie complexe qui nécessite une prise en charge pluridisciplinaire, au sein de laquelle nous tenons une place importante, Ce qui nécessite l'utilisation d'un système d'aide au diagnostic pour faciliter la prise de décision qui va être aborder dans le prochain chapitre .

1. Introduction

La classification a pris aujourd'hui une place importante en analyse des données Exploratoire et décisionnelle. Dans les domaines d'applications que des développements méthodologiques.
D'une façon générale, plus on a de grande base de données, il semble intéressant de travailler dans un contexte de classification.
Il existe de nombreuses méthodes de classification. Une bonne connaissance du problème est nécessaire pour choisir la bonne méthode à utiliser. Le choix de la méthode dépend notamment du problème posé, de la nature des données, des propriétés de la fonction à estimer. De plus, la difficulté intrinsèque du problème dépend de la qualité des données.

La première partie du chapitre 2est embardée différentes techniques de classification, Nous détaillerons, plus particulièrement, les méthodes de la classification supervisée qui seront utilisées dans notre problématique traitée de fin d'étude.

2 .Approche de classification

2.1 Définition

La classification est l'une des techniques les plus anciennes d'analyse et de traitement de données, Une classe est un ensemble d'éléments qui sont semblables entre eux et qui sont dissemblables à ceux d'autres classes. La classification repose sur des objets à classer. Les objets sont localisés dans un espace de variables (ont dit aussi attributs, caractéristiques ou critères). Il s'agit de les localiser dans un espace de classes. Ce problème n'a de sens que si on pose l'existence d'une correspondance entre ces deux espaces. Résoudre un problème de classification, c'est trouver une application de l'ensemble des objets à classer, décrits par les variables descriptives choisies, dans l'ensemble des classes. L'algorithme ou la procédure qui réalise cette application est appelé classifieur [15].

On retrouve trois approches de classification :

2.2 Classification non supervisé

Est une méthode d'apprentissage automatique [15].Il s'agit pour un logiciel de diviser un groupe hétérogène de données, en sous-groupes de manière que les données considérées comme les plus similaires soient associées au sein d'un groupe homogène et qu'au contraire les données considérées comme différentes se retrouvent dans d'autres groupes distincts l'objectif étant de permettre une extraction de connaissance organisée à partir de ces données.

La recherche d'une partition des données revient à regrouper celles-ci selon une certaine mesure de similarité ou de dissimilarité Contrairement à l'apprentissage supervisé, dans l'apprentissage non-supervisé il n'y a pas de signe qui explicite les étiquettes [20] [14].

2.3 Classification Semi supervisé

L'apprentissage semi-supervisé est une classe de techniques d'apprentissage qui utilise un ensemble de données étiquetées et non-étiquetés. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non-supervisé qui n'utilise que des données non-étiquetées. Il a été démontré que l'utilisation de données non-étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage [20].

2.4 Classification Supervisée

(Un ensemble de points étiquetés) c'est une technique d'apprentissage automatique où l'on cherche à produire automatiquement des règles à partir d'une base de données d'apprentissage. (En général des cas déjà traités et validés).

Dans le cas supervisé, les classes d'appartenance des données sont connues [15]. La recherche des frontières entre les classes peut être effectuée par la recherche des fonctions discriminantes.

La classification supervisée ou la classification inductive a pour objectif de chercher à expliquer et à prédire l'appartenance de documents à des données connues a priori [11].

Dans le cadre de ce projet de fin d'étude, nous entamerons l'apprentissage supervisé pour des problèmes de classification. Dans la suite, nous présenterons les principaux algorithmes de classification supervisée proposés dans la littérature. Il ne s'agit pas de faire une présentation exhaustive de toutes les méthodes mais seulement de préciser les méthodes les plus classiques que nous utiliserons dans le cadre de notre travail en fonction de leurs propriétés particulières.

3 .Les K plus proches voisins

3 1.définition

Les K plus proches voisins , connus en anglais sous le nom K-Nearest Neighbor(K-NN) est une méthode d'apprentissage non paramétrique qui ne nécessite pas de construction de modèle [28], c'est l'échantillon d'apprentissage, associé a une fonction de distance et d'une fonction de choix de la classe en fonction des classes des voisins les plus proches, qui constitue le modèle.

Pour prédire la classe d'un exemple donné, l'algorithme cherche les K plus proches voisins de ce nouveau cas et prédit la réponse la plus fréquente de ces K plus proches voisins [12]. Le principe de décision consiste tout simplement donc a calcules la distance de l'exemple inconnu a tous les échantillons fournis. L'exemple est alors affecté à la classe majoritaire représenté parmi ces K échantillons. La méthode utilise deux paramètres : le nombre K et la fonction de similarité pour comparer le nouvel exemple aux exemples déjà classés [13].

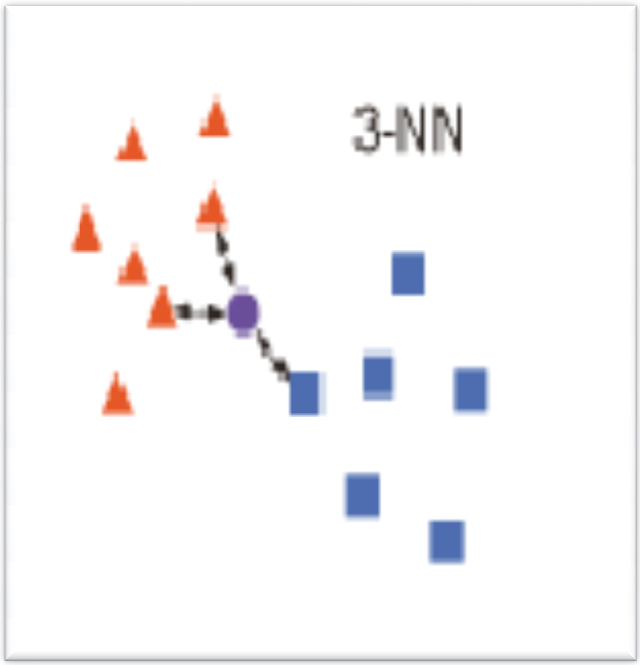

Figure II. 3.1 - Exemple de classification par KNN.

3.2 Le choix de k

Le paramètre k doit être déterminé par l'utilisateur : k∈ℕ, il est utile de choisir k impair pour éviter les votes égalitaires. Le meilleur choix de k dépend du jeu de donnée. La fixation du paramètre k est délicate, une valeur très faible va engendrer une forte sensibilité. Un K trop grand va engendrer un phénomène d'uniformisation des décisions.

Pour remédier à ce problème, il faut tester plusieurs valeurs de k et choisir le k optimal qui donne un meilleur taux de classification [15].

3.3 Algorithme de KNN

Algorithme KNN

Début

Paramètre : le nombre K de voisins

Données : un échantillon de n exemple d'apprentissage Ω = (ω_1, ..., ω_n)

la classe d'un exemple ω est Y (ω), Y= $\{c_1, c_2, \dots, c_n\}$

Entrée : un enregistrement X

Pour chaque exemple ω **faire**

Calculer la distance d(X,ω) ;

fin

KNN=les k plus proche voisins de X qui minimise la distance d ;

Pour chaque {$\omega \epsilon$ KNN} **faire**

Calculer les scores des classes ;

fin

Attribuer Y (X) a la classe ayant le plus grand score ;

Sortie : la classe de X est Y ((X)) =c_j ;

fin [21]

4 .Les Séparateurs à Vaste Marge

4.1. Principe de SVM

Les Séparateurs à Vaste Marge ou Support Vector Machines (SVM) sont une classe de techniques d'apprentissage introduite par (Vladimir Vapnik)au début des années 90 dans son livre « The nature of statistical learning theory » représente une des méthodes récente de classification supervisée [17] [18].

Le SVM utilisées dans différents domaines de recherche et d'ingénierie tel que le diagnostic médical, le marketing, la biologie, la reconnaissance de caractères manuscrits et de visages humains [22].

Le but de SVM est de déterminer si un élément appartient à une classe ou pas[27]. Nous disposons d'un ensemble de données et nous cherchons à séparer ces données en deux groupes. Le premier est l'ensemble de données appartenant à une classe, ces données sont étiquetées par malade et un autre ensemble qui contient les éléments qui n'appartiennent pas à la classe donc étiquetées non malade.

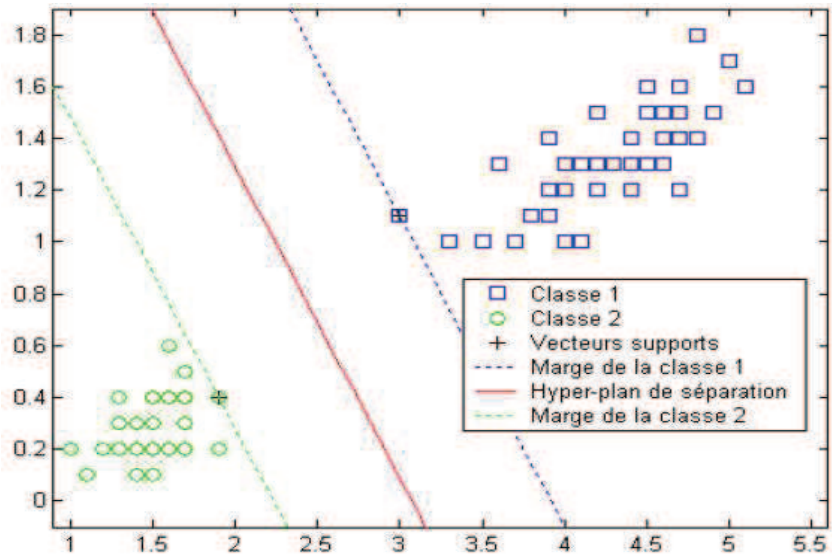

Figure II. 4.1 - principe de Séparateurs à Vaste Marge.

Il existe deux types de classifieur SVM mais dans le cadre de ce projet de fin d'étude, nous entamerons le SVM linéaire pour résoudre notre problème.

4 .1.1 SVM non linéaire

Dans la plupart des problèmes, l'hypothèse de linéarité est trop restrictive et le séparateur optimal doit pouvoir prendre une forme plus compliquée. La méthode du noyau est un moyen élégant et efficace pour traiter ce problème on transfère les données de l'ensemble de départ $\vec{x}$ vers un ensemble de dimension supérieure $\Phi(\vec{x})$ dans lequel le problème devient séparable linéairement [18].

La résolution de SVM de remplacer les vecteurs de l'ensemble des données qui interviennent seulement dans le produit scalaire $\vec{x}_i.\vec{x}_j$ par nouvel espace caractéristique $\Phi(\vec{x}_i).\Phi(\vec{x}_j)$.donc le produit scalaire devient :

$K(\vec{x}_i,\vec{x}_j) = \Phi(\vec{x}_i).\Phi(\vec{x}_j)$

et la nouvelle fonction de classification est donnée par :

$f(\vec{x}) = \sum_{i=1}^{N} \lambda_i \; y_i \, K\left(\vec{x}_i,\vec{x}_j\right) + b$ [17]

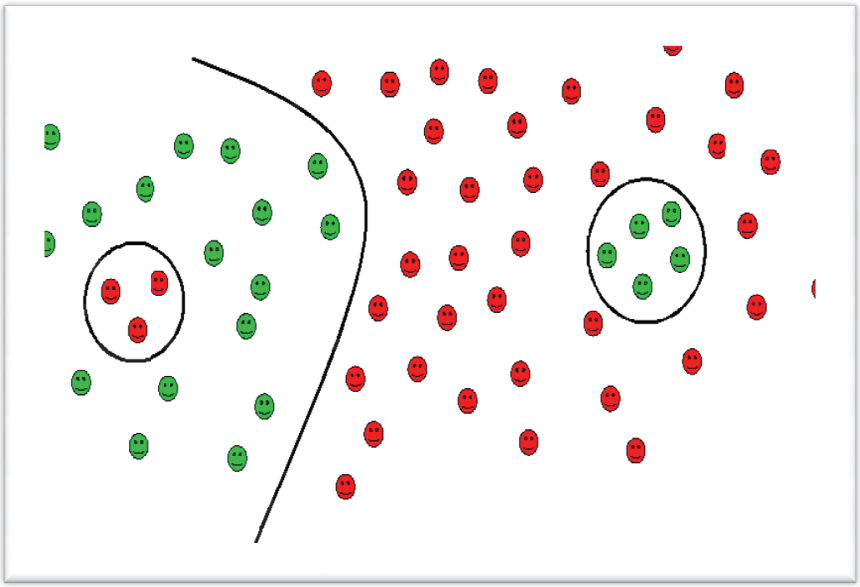

Figure II. 4.1.1 - Problème d'un SVM non linéaire [19].

4 .1.2 SVM linéaire

A partir d'un ensemble d'apprentissage $(\vec{x}_i, y_i)$ ou $\vec{x}_i$ est un ensemble des données et $y_i \in \{-1, +1\}$ celui des classes, les SVM consistent `a trouver l'hyperplan séparateur Optimal qui maximise la distance entre l'hyperplan et les deux classes Cette distance est appelée la marge [18].

L'hyperplan est défini par $\vec{\omega} . \vec{x} + b$ ou $(\vec{\omega}, b)$ désignent les paramètres de l'hyperplan (respectivement un vecteur normal au plan et le biais). Le classifieur est donné $y = sign(\vec{\omega} . \vec{x} + b) \in \{-1, +1\}$.

Néanmoins, ce dernier doit satisfaire :

$$\begin{cases} \vec{\omega} . \vec{x} + b \geq 1 & si \ y_i = +1 \\ \vec{\omega} . \vec{x} + b \leq 1 & si \ y_i = -1 \end{cases} \quad [26]$$

La distance d'un point `a l'hyperplan est : $\frac{|w\,x+b|}{\|w\|}$

La marge entre les deux classes vaut : $\frac{2}{\|w\|}$ [17]

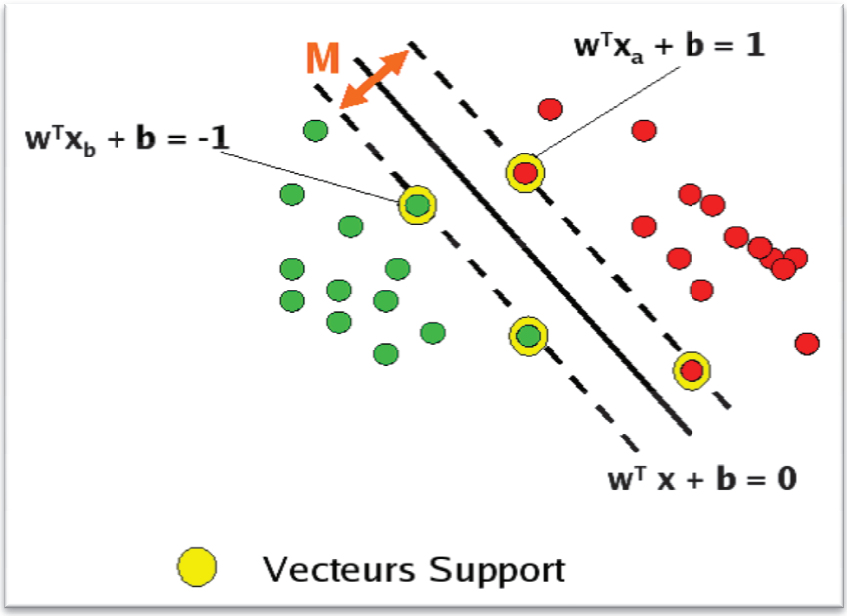

Figure II.4.2 -Principe d'un SVM linéaire [19].

5.Évaluation de classification

La performance est un aspect de comportement, qui est évaluée par le calcul du pourcentage de sensibilité (SE), la spécificité (SP) et taux de classification (TC).

5.1 La sensibilité

La Sensibilité (SE%) : [SE = 100 * VP / (VP + FN)] on appelle sensibilité (Se) la capacité de donner un résultat positif quand la maladie est présente en représentant ceux qui sont correctement détectés parmi tous les événements réels [8].

5.2 La spécificité

La Spécificité (SP %) : [SP = 100 *VN / (VN + FP)] on appelle spécificité la capacité de donner un résultat négatif quand la maladie est absente. Elle est représentée pour détecter les cas non parkinsoniens [9].

5.3 Taux de classification

Taux de classification (TC %) : [TC = 100 * (VP + VN) / (VN + VP + FN +FP)] est le taux de reconnaissance [15].

– VP : parkinsonien classé parkinsonien

– FP : parkinsonien classé non parkinsonien

– VN : non parkinsonien classé non parkinsonien

– FN : non parkinsonien classé parkinsonien.

6. Matrice de confusion

L'évaluation d'une classification est effectuée dans une matrice de confusion en fonction des classes de qualité initiales (Tableau II.6) La matrice contient des informations concernant le classement actuel dans la base de test ainsi que le classement prédit, La performance du système de classement est évaluée en utilisant les données de la matrice. Le (tableau II.6) présente la matrice de confusion pour deux valeurs de classe (positive, négative) [14].

•Vrai positif : exemple positif classé positif.

•Faux négatif: exemple négatif classé positif.

•Vrai négatif: exemple négatif classé négatif.

•Faux positif : exemple positif classé négatif.

Classe vrai → ↓ Classe retrouvé	Positif	Négatif
positif	VP	FN
Négatif	FP	VN
Total	P	N

Tableau II.6 - Matrice de confusion.

7. Conclusion

Dans ce chapitre, nous avons abordé les approches de classification plus précisément la supervisé, en donnant les principaux méthodes proposés dans la littérature. Il ne s'agit pas de faire une présentation exhaustive de toutes les méthodes mais seulement de préciser les plus classiques que nous utiliserons dans le cadre de notre travail en fonction de leurs propriétés particulières, cette classification est évaluer par une matrice de confusion pour présenté les bonnes décisions.

1. Introduction

Chaque observation est caractériser par un ensemble de variables, ces variables ne sont pas toute informatives. En effets certaines d'entres elles peuvent être peu significatives, ou non pertinentes.

La sélection des paramètres pertinentes présente un intérêt majeur qui permet non seulement de réduire la dimension des données a traité et par conséquent de réduire le temps de calcule et la complexité des algorithmes de classification mais aussi d'améliorer les performances de généralisation de ces derniers

Dans ce chapitre nous commencent donc par définir les différentes mesures de pertinence et de redondance rencontrées. Ensuite nous présenterons les différentes approches de sélection de variables et aussi les méthodes de sélection, plus particulièrement la sélection des variables de maladie de parkinson.

2. Sélection de variables

2.1 Principe

La sélection de variables est un domaine très actif depuis ces dernières années [9].
Généralement elle définie comme un processus de recherche permettant de trouver un sous-ensemble "pertinent" de variables parmi celles de l'ensemble de départ.
Lorsque le nombre de variables est de grande taille, l'algorithme d'apprentissage ne peut pas terminer l'exécution dans un temps convenable, alors la sélection réduit la dimension de l'espace des variables, Elle améliore la performance de la classification [23].

De ce fait la sélection des données consiste à choisir un sous-ensemble optimal de variables pertinentes, à partir d'un ensemble de variables.

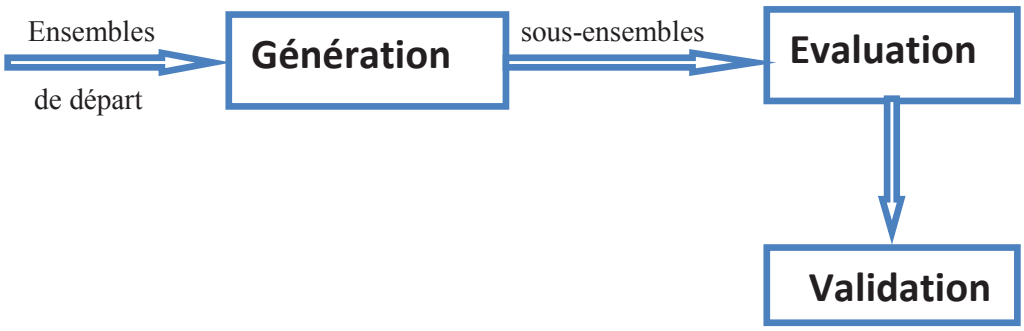

Figure III.2.1 -Procédure générale d'un algorithme de sélection de variables [15].

Il existe trois types de stratégies de sélection de variables :

Dans la première stratégie : la taille de sous-ensemble à sélectionner est prédéfinie et l'algorithme de sélection cherche à trouver le meilleur sous-ensemble de
Cette taille.

La deuxième stratégie : consiste à sélectionner le plus petit sous-ensemble dont la performance est plus grande.

La troisième stratégie : cherche à trouver un compromis entre l'amélioration de la performance et la réduction de la taille du sous ensemble. Le but est de sélectionner le sous-ensemble qui optimise les deux objectifs en même temps [10].

3. Pertinence et redondance de variables

La SV consiste à choisir parmi un ensemble de variables de grande taille, un sous-ensemble de variables intéressant pour le problème a étudier [15].

En présence de centaines, voire de milliers de variables, il y a beaucoup de chances pour que des variables soient corrélées et expriment des informations similaires, on dira alors qu'elles sont redondantes. D'un autre côté, les variables qui fournissent le plus d'information pour la classification sont appeler pertinentes. L'objectif de la sélection est donc de trouver un sous-ensemble optimal de variables qui ait les propriétés suivantes : il doit être composé de variables pertinentes et il doit chercher à éviter les variables redondantes. De plus cet ensemble doit permettre de satisfaire au mieux l'objectif fixé c'est-à-dire la précision et la rapidité de la classification ou bien encore l'explicabilité du classifieur.

3.1 Pertinence de variables

Ont peut classée les variables comme étant : très pertinente, peu pertinente et non pertinente.

Très pertinente : les variables fortement pertinentes sont donc indispensables et devraient figurer dans tout sous-ensemble optimal sélectionné, car leur absence entraine une détérioration significative de la performance du système de classification utilisée.
Peu pertinente : La faible pertinence suggère que la variable n'est pas toujours importante
Non pertinente : indique qu'une variable n'est pas du tout nécessaire dans un sous-ensemble optimal de variables, Ces variables seront en général supprimées de l'ensemble de variables de départ [10].

3.2 Redondance de variables

La notion de la redondance de variables se comprend intuitivement et elle est généralement exprimée en termes de corrélation entre variables. On peut dire que deux variables sont

redondantes (entre elles) si leurs valeurs sont complètement corrélées.
Cette définition ne se généralise pas directement pour un sous-ensemble de variables [15].

4. Approche de la sélection de variables

La sélection de variables est un dispositif crucial de l'apprentissage. Nous cherchons à évaluer un sous-ensemble qui permet de traiter efficacement les valeurs de la variable cible, tout en précisant le type d'approche utilisé. Dans la littérature de la sélection de variables trois approches nous citons deux :

- Approche enveloppe (wrapper).
- Approche filtre (filter).

4.1 Approche (wrapper)

Où encore nommée les méthodes enveloppes qui ont été introduites par Kohavi et John (John, et al, 1994) ; (Kohavi & John, 1997) [10]. Leurs principe est de sélectionner un sous-ensemble de variables en fonction des performances d'un classificateur construit sur ces variables, c'est-à-dire à chaque sélection d'une variable, nous calculons le taux de classification pour juger la pertinence d'une variable.
Elles ont souvent de meilleurs résultats par apport aux autres méthodes puisque la sélection est directement reliée à la performance d'un classifieur, mais au prix d'un temps de calcul plus important.

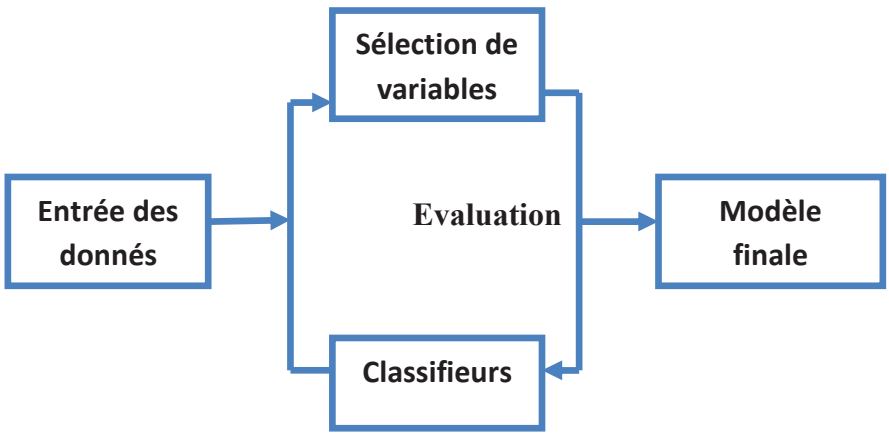

Figure III. 4.1 - Principe de l'approche wrapper [9].

4.2 Méthode filter

Le modèle filter a été le premier utilisé pour la sélection de variables. Son principe consistent à sélectionner des sous ensemble de variables indépendamment du classifieur en utilisant leurs variables générales c'est-à-dire La sélection se fait avant l'apprentissage. Il sert à rassembler les éléments d'une même classe et séparer les éléments étiquetés différemment. Cette méthode est considérée, davantage comme une étape de prétraitement (filtrage) avant la phase d'apprentissage [15].
D'après les résultats trouvés ils ont montrés que cette approche permet d'améliorer les performances en classification supervisé.

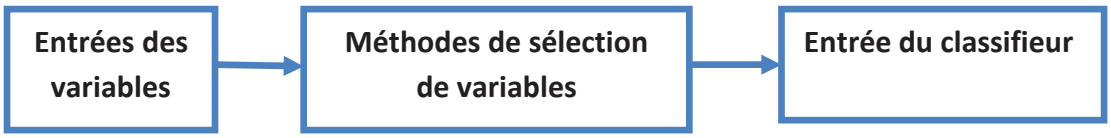

Figure III. 4.2 - Principe de l'approche filtre [9].

5. Méthodes de sélection des variables

5.1 Relieff

Une des méthodes de filtrage les plus connues pour la sélection de variables, qui a été proposée en 1992 par Kira et Rendell (Kira et Rendell[1992]) [9],cette méthode repose sur un algorithme qui donne une liste ordonnée de toutes les variables suivant leur pondération par un critère de distance [24]. Son principe est de calculer une mesure globale de la pertinence des variables[25] en accumulant la différence des distances entre des exemples d'apprentissage choisis aléatoirement et leurs plus proches voisins de la même classe et de l'autre classe.

5.1.1 Algorithme de sélection Relieff

Algorithme de sélection Relieff

Entrées : Une base d'apprentissage A = $\{X_1, X_2, \dots, X_M\}$ ou chaque exemple $X_i = \{x_{i1}, x_{i2}, \dots, x_{iN}\}$

Nombre d'itérations T

Sorties: W [N] : vecteur de poids des variables (f_i), -1 ≤ $w[i] \leq 1$

$\forall_i$, W [i] =0 ;

Pour t = 1 a T **Faire**

Choisir aléatoirement un exemple X_K

Chercher deux plus proches voisins (un dans sa classe (X_a) et un deuxième dans l'autre classe (X_b))

Pour i = 1 a N **Faire**

W[i] = W[i] + $\frac{x_{ki} - x_{bi}}{M * T}$ - $\frac{x_{ki} - x_{ai}}{M * T}$

Fin Pour

Fin Pour

Retourner W [10]

5.2 Fisher

Le test de Fisher est définit comme suit :

$$P = \frac{(x_1 - x_2)^2}{{s_1}^2 - {s_2}^2} \quad [24]$$

Où x_k et ${s_k}^2$ sont la moyenne et l'écart type de l'attribut pour la classe utilisé. Un score important indique donc que les moyennes des classes sont significativement différentes [10].

6. Conclusion

Dans ce chapitre, nous avons présenté le processus de sélection de variables et l'importance de cette dernière pour l'amélioration des performances des classifieurs Comme nous avons renforcé nos informations sur la sélection par des algorithmes qui ont été proposé dans la littérature qui vont nous servir dans le prochain chapitre de résultat.

1. Introduction

Dans le présent chapitre nous disposant notre contribution représentée dans la sélection de variables. Pour cela ce dernier est consacré à la présentation de nos résultats obtenu.

Dans cette optique, nous montrons les résultats acquis par l'application de notre Classificateur KNN et SVM linéaire sans sélection de variables, tout en soulevant l'apport des méthodes de sélection et nous terminons par une comparaison de nos résultats obtenus via les classifieurs (KNN, SVM linéaire) et ceux par les méthodes de sélection utilisée (Relieff, fisher).

2. Base de données

Cette base de données est créée par **Max Little** de l'Université d' Oxford, en collaboration avec le Centre national de la voix et la parole, elle concerne la maladie de parkinson et elle est composée d'une série de mesures vocales biomédicales de 31 personnes, dont 23 avec la maladie de Parkinson et 8 des personnes sains ou normaux, donc c'est une base déséquilibré. Les expériences ont était menées pour enregistré les signaux de parole de 195 voix donc les variables de notre base de donnée sous forme des fréquences, L'objectif principal des données est de discriminer les personnes en bonne santé de ceux avec la maladie de parkinson.

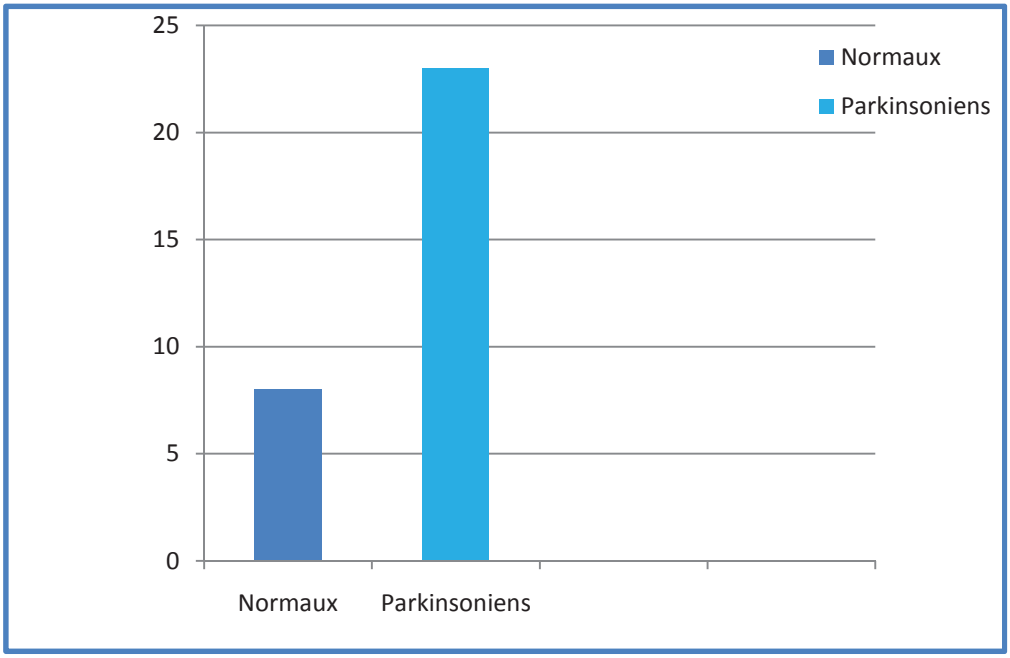

Figure IV. 2 - Répartition des différents cas de la base de la maladie de parkinson.

3. Les résultats obtenus et comparaisons

Notre travail consiste à séparer la base de données en deux parties: l'une sera utilisée pour l'apprentissage, et l'autre pour le teste.

Il faut juste définir la taille de chacun de ces deux échantillons. Généralement on utilise **2/3** des individus pour l'apprentissage, et **1/3** pour le teste.

Les objectifs des expérimentations effectuées sur notre base de donné sont d'une part destinés pour évaluer les performances des algorithmes de classification que nous avons utilisés (KNN et SVM linéaire) et d'autre part de tester l'efficacité de la sélection de variables par les deux méthodes de sélection pour l'amélioration de taux de classification.

Le classificateur KNN et SVM linéaire a été mis en œuvre dans MATLAB (environnement de programmation). Les performances du classifieurs mis en œuvre a été évaluée en calculant le pourcentages de sensibilité (SE) , la spécificité (SP) et le taux de classification (TC).

3.1 Résultats sans sélection de variables

3.1.1 classifieur KNN et SVM linéaire

L'utilisation d'un classifieur supervisé de type K plus proches voisins (KNN) et de type support a vaste marge (SVM linéaire) nous a permet d'obtenir les résultats suivant :

Classifieur utilisé					
KNN /K=5			SVM linéaire		
Sensibilité	Spécificité	Taux de classification	Sensibilité	Spécificité	Taux de classification
77,78%	81,82%	**78,46%**	86,49%	57,14%	73,85%

Tableau IV.3.1.1 - les performances d'un classifieur KNN et SVM linéaire sans sélection.

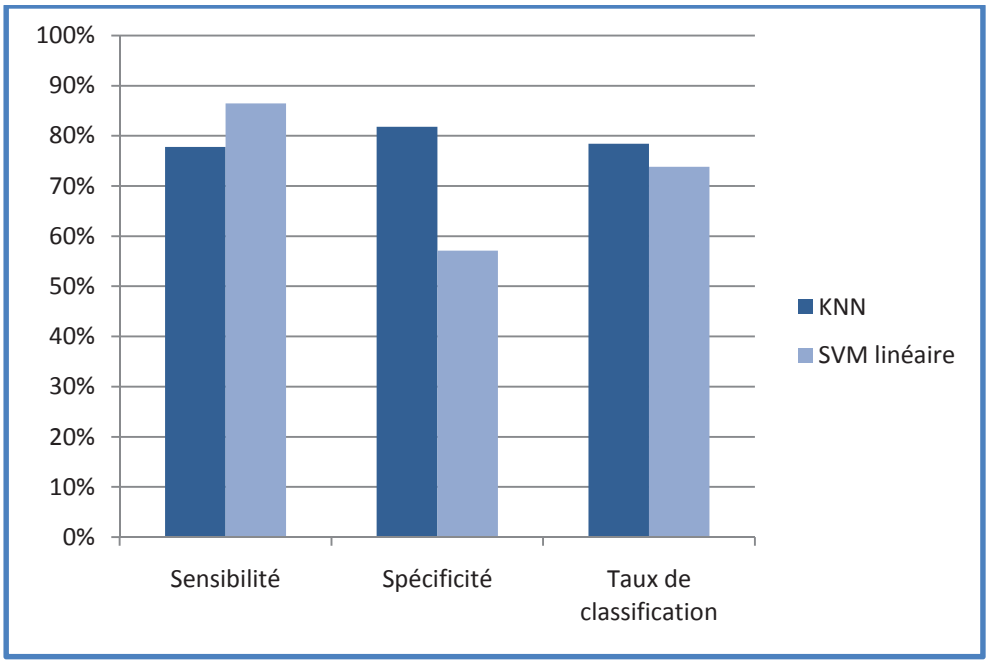

Figure IV.3.1.1 - les performances d'un classifieur KNN et SVM linéaire sans sélection.

Dans cette section, nous présentons les résultats de l'application de notre classifieur sur notre base de donné.

Le tableau **(Tableau IV. 3.1.1)** montre les différentes performances sans sélection de variables en utilisant un classifieur KNN Cela évite la difficulté d'une phase d'apprentissage et un classifieur SVM linéaire.

Plusieurs valeur de K ont été testé pour le KNN afin d'adopter un meilleur résultat donc nous avons choisit le k =5 qui fournit un meilleur taux de classification que ceux obtenus par le classifieur SVM linéaire (**Figure IV.3.1.1**).

4. Application du classifieur KNN

4.1 Résultats de sélection par la méthode de Relieff

La valeur K pris pour la méthode de relieff est fixée à 5 après avoir effectué plusieurs tests avec plusieurs valeurs de K, pour améliorer notre classificateur K plus proches voisins, ce qui nous a menés aux résultats suivants:

Sélection par relieff				
KNN	Nombre de variables sélectionnées	Sensibilité	Spécificité	Taux de classification
K=3	**3**	79,25%	83,33%	80%
	5	82,35%	85,71%	83,08%
	7	84%	86,67%	84,62%
	9	84%	86,67%	84,62%
K=5	**3**	76,36%	80%	76,92%
	5	**86%**	**93 ,33%**	**87,69%**
	7	77,78%	81,82%	78,46%
	9	77,78%	81,82%	78,46%

Tableau IV.4.1- Résultats obtenus selon le nombre de voisinage et le nombre de variables sélectionnés par la méthode de Relieff.

(Le tableau IV.4.1) montre les performances d'un classifieur KNN sur les variables les plus informatifs qui possédant un poids fort (3, 5 ,7et 9 variables) sélectionné par la méthode de Relieff , les résultats obtenus ici montrent que la valeur de K=5 du KNN donne un meilleur taux de classification avec 5 variables(les variables sont ordonné), puis il décroit pour les 7 variables et il reste stable lorsqu'il dépasse ces dernier jusqu'à la dernière variable sélectionné.

Nous pouvons remarquer que notre méthode est capable de sélectionner des sous-ensembles avec une meilleure performance à celle de la classification (KNN) du fait que la sélection des variables nous favorise d'avoir un meilleur taux .donc La méthode Relieff sélectionne les variables qui ont une grande pertinence, ce qui implique qu'elle permet effectivement de réduire les données inutiles.

4.2 Résultats de sélection par la méthode de fisher

On appliquant la méthode de sélection fisher sur notre base de données elle nous a permet d'obtenir les résultats suivant :

Sélection par fisher				
KNN	Nombre de variables sélectionné	Sensibilité	Spécificité	Taux de classification
K=3	**3**	84,44%	70%	80%
	5	84,44%	70%	80%
	7	76,36%	80%	76,92%
	9	82,35%	85,71%	83,08%
K=5	**3**	83,33%	76,47%	81,54%
	5	83,33%	76,47%	81,54%
	7	76,36%	80%	76,92%
	9	**84%**	**86,67%**	**84,62%**

Tableau IV. 4.2 - Résultats obtenus selon le nombre de voisinage et le nombre de variables sélectionnés par la méthode de Fisher.

(Le tableau IV.4.2) résume que la sélection des variables par la méthode de fisher , réalise un taux de classification remarquable comparé à notre classificateur KNN, pour 9 variables et une valeur de K=5 on a eu une amélioration du système.

4.3 Etude comparative entre Relieff et fisher

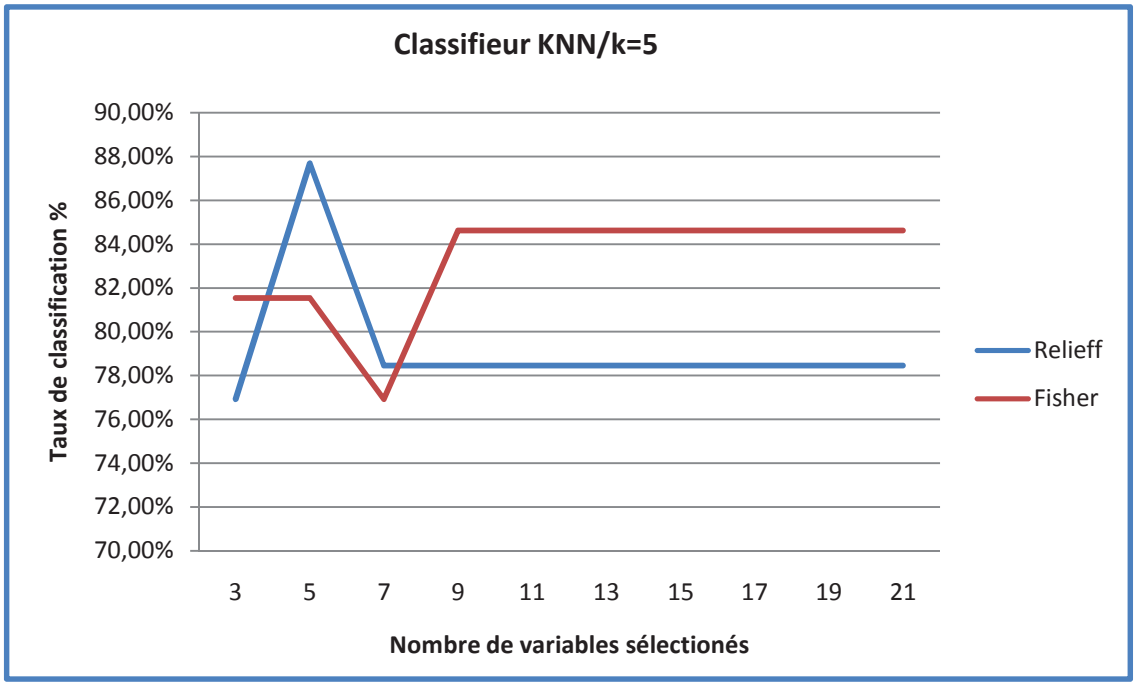

Figure IV. 4.3 - Taux de classification par les méthodes de sélection en utilisent un classifieur KNN.

Pour mieux comprendre et analyser les résultats obtenus dans les expériences précédentes, nous présentons dans la **(Figure IV.4.3)** les taux de la classification en fonction des de variable sélectionnées (relieff et fisher).

D'après notre figure on remarque deux distingues intervalles de nombres de variables sélectionnée :

De **[3-7]** on remarque que le taux de classification varie jusqu'à son maximum quand le nombre de variable égale a cinq. Contrairement à relief, Fisher ou on remarque que notre taux de classification varie jusqu'à son minimum lorsque le nombre sélectionnée est sept.

De **[7-9]** dans cette intervalle le taux de classification de relief chute à son minimum puis il se stabilise et pour Fisher il atteint son maximum pour un nombre de variable égale à nef et au de la de nef il se stabilise.

On résume notre comparaison, on peut dire que la sélection par relief nous à favorisé un meilleur taux de classification 87,82% comparer à Fisher qui nous a donner 81,54%, donc les résultats de cette technique est capable de fournir des sous-ensembles de petite taille avec une haute performance.

5. Application du classificateur SVM linéaire

Comme dans les expériences précédentes, on a sélectionné le même nombre de variables pour le cas Relief et Fisher, mais cette fois ci en utilisent un classificateur SVM linéaire.

5.1 Résultats de sélection par la méthode de Relieff

La méthode de Relief appliquée sur la maladie de parkinson a donné les résultats que nous présentons dans le tableau ci-dessous:

Sélection par Relieff			
Nombre de variables sélectionnés	Sensibilité	Spécificité	Taux de classification
3	83,33%	51,72%	69,23%
5	85,71%	53,33%	70,77%
7	83,78%	53,57%	70,77%
9	**86,84%**	**59,26%**	**75,38%**

Tableau IV.5.1 -les performances de la sélection par la méthode de relief en utilisent un classificateur SVM linéaire.

(Le tableau IV.5.1) représente les performances de classificateur SVM linéaire par la méthode de Relieff. On a obtenus un meilleur taux de classification avec 9 variables sélectionnés a celle des autres qui sont proches et comparables.

Nous pouvons constater que cette technique est capable de faire une sélection intéressante sur notre base de données grâce à de sa simplicité de calcule.

5.2 Résultats de sélection par la méthode de Fisher

La méthode de Fisher appliquée sur la maladie de parkinson a donné les résultats que nous présentons dans le tableau ci-dessous:

Sélection par Fisher			
Nombre de variables sélectionnés	Sensibilité	Spécificité	Taux de classification
3	82,93%	58,33%	73,85%
5	82,05%	53,85%	70,77%
7	84,21%	55,56%	72,31%
9	**86,84%**	**59,26%**	**75,38%**

Tableau IV.5.2 - Les performances de la sélection par la méthode de Fisher en utilisent un classificateur SVM linéaire.

(Le tableau IV.5.2) donne un aperçu sur les taux de classification de classificateur SVM linéaire en utilisant la méthode Fisher. On a remarqué que ce procédé réalise un meilleur taux de classification avec 9 variables sélectionné.

6. Comparaison entre les classificateurs KNN et SVM linéaire

	Classificateur utilisé					
	KNN			SVM linéaire		
Taux de classification %	Sans sélection	Relieff	Fisher	Sans sélection	Relieff	Fisher
	78,46%	**87,69%**	**84,62%**	73,85%	75,38%	75,38%

Tableau IV. 6 - Taux de classification(%) sans et avec la sélection pour un classificateur KNN et SVM linéaire.

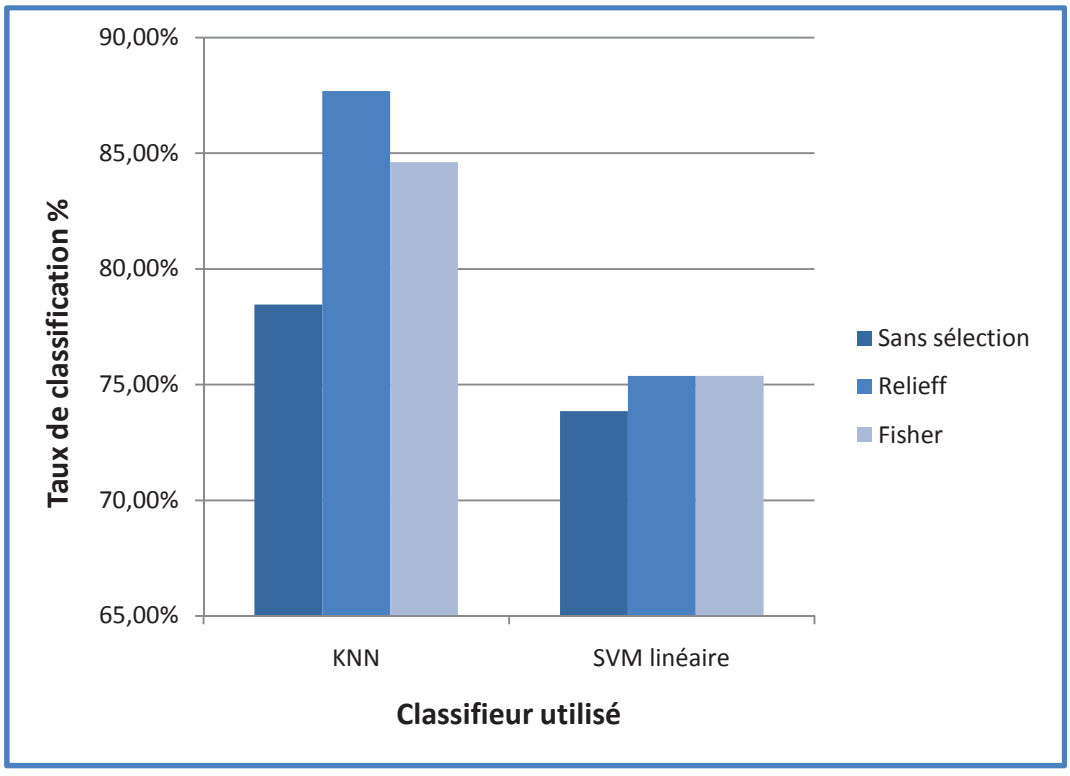

Figure IV. 6 - Les performances du classificateurs KNN et SVM linéaire avant et après la sélection.

(Le tableau IV.6 et la figure IV.6) montrent le taux de classification du classificateur KNN et SVM linéaire des sous ensembles de variables en appliquant la sélection par Relief et Fisher ainsi qu'une classification avec KNN et SVM linéaire sans sélection. On remarque que le taux de classification obtenue avec la sélection de variables est plus important que celui appliqué sans sélection. L'évaluation de nos résultats expérimentaux ont montré que le classificateur KNN offres les meilleurs résultats, sans et avec la sélection par rapport au classificateur SVM linéaire.

7. Conclusion

Dans ce chapitre nous avons présenté les résultats d'expérimentation de la sélection de variables par de différentes méthodes, ces dernières sont basées sur l'application du classificateurs supervisé, d'où on a étudié deux type le KNN et SVM linéaire. Une optimisation par les méthodes de sélection à fait l'objet de notre recherche pour l'amélioration de notre système d'aide au diagnostic.

Conclusion générale

Notre projet de fin d'Etude intitulé «Sélection de variables pour la *reconnaissance de la maladie de parkinson* » s'inscrit dans le cadre de l'élaboration d'un procédé automatisé dédié à l'aide au diagnostique. Nous nous sommes mis en œuvre à réaliser un prototype de reconnaissance de la maladie de parkinson qui repose essentiellement sur une classification supervisée.
Dans la littérature la sélection de variable est un domaine de recherche prometteux qui donne lieu à une étude poussé en classification destiné à répondre à plusieurs problèmes en matière de réduction des attributs.

Les divers travaux réalisés durant ce mémoire nous ont menés vers plusieurs contributions concernant la sélection de variables pour la détection de parkinson.
La première contribution développée dans cette étude traite les méthodes de classification de variables supervisés avec deux classificateurs KNN et SVM linéaire a fin d'évaluer les pertinences des sous –ensembles.les résultats obtenus lors de la comparaison entre ces classificateurs ont démontré que KNN est plus performant en fonction de taux classification .

La deuxième contribution de cette thèse correspond aux méthodes de sélection de variables qui sont moins redondantes avec les deux techniques Relieff et Fisher de l'approche filtre par le biais des deux classificateurs KNN et SVM linéaire. Les résultats expérimentaux ont démontré que les sous-ensembles sélectionnés par la méthode Relief présente des performances meilleures et plus efficace par KNN à celle de la méthode Fisher.

Ce travail montre l'intérêt d'utiliser les techniques de classification et de la sélection pour étudier une maladie rare et complexe comme la MP en utilisant une base de données existante.

Bibliographie

[1] Flavien Eger ANTOINE, Christophe GAUDET BLAVIGNAC et Arthur HAMMER. *La maladie de parkinson*. Université de Genève, Juin 2009.

[2] Valérie SOLAND, *la maladie de parkinson*, Avril 2011.

[3] Dominique BLACKBURN. *Hypothèse d'une étiologie environnementale de la maladie de parkinson : connaissances actuelles et données disponibles au Québec*. Mémoire de maître, Université de Sherbrooke, septembre 2007.

[4] CEN - Collège des Enseignants en Neurologie - http://www.cen-neurologie.asso.fr

[5] Line BEAUDET, Chantal BEAUVAIS, Sylvain CHOUINARD, Manon DESJARDINS, Michel PANISSET et Emmanuelle POURCHER, *La maladie de parkinson et ses traitements,* Avril 2009.

[6] Frédéric MOISAN. *Prévalence et facteurs de risque professionnels de la maladie de Parkinson parmi les affiliés à la Mutualité Sociale Agricole.* Université Paris Sud, Jun 2012.

https://tel.archives-ouvertes.fr

[7] Max Little, *Parkinsons Disease Data Set,* l'Université d' Oxford ,UCI machine repository (Parkinson).

[8] Hafida GAHDOUM. *Classification des données déséquilibrées médicale*. Mémoire en master, Université Abou Bakr Belkaïd, juin 2013.

[9] Amel HAFA. Sélection de Variables Biologiques par l'approche filtre. Mémoire en master, Université Abou Bakr Belkaïd, Juillet 2012.

[10] Hassan CHOUAIB. Sélection *de caractéristiques: méthodes et applications*. Doctorat en informatique, Université Paris Descartes, juillet 2011.

[11] Adrien BOUGOUIN. *État de l'art des méthodes d'extraction automatique de termes-clés*, May 2013.
[12] Jérôme AZE. *Cours Datamining K-plus proches voisins,* Université de Paris, mars 2007.

[13] Faicel CHAMROUKHI. *Classification supervisée : Les K-plus proches voisins*. Université du Sud Toulon, Mars 2013.

[14] Laurence BOUDET. Auto-*qualification de données géographiques 3D par appariement multi-image et classification supervisée. Application au bâti en milieu urbain dense*. Mémoire de Doctorat, Université Paris-Est Marne-la-Vallée, septembre 2007.

[15] Ali EL AKADI. *Contribution a la sélection de variables pertinentes en classification supervisée : application a la sélection des gènes pour les puces à ADN et des caractéristiques faciales. DOCTORAT* en Informatique et Télécommunications, UNIVERSITÉ MOHAMMED V – AGDAL FACULTÉ DES SCIENCES Rabat, mars 2012.

[16] J.DOSHAY, Lewis, Madison H.THOMAS. *Guide médical de la famille*, juin 1973.

[17] Olivier ZAMMIT, Xavier DESCOMBES, Josiane ZERUBIA, *Apprentissage non supervisé des SVM par un algorithme des K-moyennes entropique pour la détection de zones brulées,* juin 2004.

[18] Arnaud REVEL, *Support Vector Machines Séparateurs à vaste marge*, février 2010.

[19] Jérome AZE. *Classification supervisée Perceptron et SVM*, octobre 2010.
[20] Laurent CANDILLIER. *Contextualisation, visualisation et évaluation en apprentissage non supervisé*, doctorat en informatique, université Charles de Gaulle-lilles3, septembre 2006.
[21]SENOUCI Hafida. *Sélection de variable, Doctorat en informatique, université d'Oran, 2010.*
[22] Abdelhamid DJEFFAL, *utilisation des méthodes support vector machine (SVM) dans l'analyse des bases de données*. Doctorat en informatique, université Mohamed Khider Biskra, juin 2012.

[23] Alexandra DURAND ,*Méthodes de sélection de variables appliquées en spectroscopie proche infrarouge pour l'analyse et la classification de textiles*, Doctorat en instrumentation et analyses avancées, Novembre 2007.

[24] Marine CAMPEDEL, Eric moulines, *classification et sélection automatique de caractéristiques de textures*, Ecole nationale supérieur des télécommunications, juin 2004.

[25] Fan WENBING, Wang QUAUQUAU et Zhu HUIN*, Feature Selection Method Based on Adaptive Relief Algorithm*, université de Zhengzhou, China, octobre 2012.

[26] S.BOUCHIKHI, A .BOUBLENZA, A .BENOSSMAN, M .A.CHIKH, *parkinson's diseace detection with SVM classifier and relief-F features selection algorithm*, janvier 2014.

[27] Faouzi ZAIZ, *les supports vecteurs machines (SVM) pour la reconnaissance des caractères manuscrits arabes*, Magister en informatique, juillet 2010.

[28] Gérad RANSTEIN, *une méthode implicative pour l'analyse de données d'expression de gènes*, université de Nantes, septembre 2007.