Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abou Bekr Belkaid
Faculté de Technologie
Département de Génie Biomédical

MÉMOIRE DE FIN D'ÉTUDES

pour obtenir le grade de

Master en Génie Biomédical

Spécialité : **Informatique Biomédicale**

présenté et soutenu publiquement
par

Bendimerad El-batoul

Benikhlef Soumia

le 27 Mai 2015

Titre:

Adaptation de la forêt rotationnelle pour le dépistage de la maladie de Parkinson

Jury

Présidente du jury.	Mme BECHAIB Yasmine,	MAA UABB Tlemcen
Examinatrice.	Mme MEKKIOUI Nawel,	MAA UABB Tlemcen
Encadreur.	Melle. SETTOUTI Nesma,	MAA UABB Tlemcen
Co-encadreur.	Mr. EL HABIB Daho Mostafa,	Doctorant UABB Tlemcen
Invité d'honneur.	Pr. Chikh Mohamed Amine,	PROF UABB Tlemcen

Je dédié ce modeste travail

A mes très chers parents : *Aucune dédicace ne saurait être assez émouvante pour exprimer ce que vous méritez pour tous les sacrifices que vous n'avez cessé de me donner. Maman tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi.*
Papa rien au monde ne vaut le dévouement et les efforts fournis jour et nuit pour mon éducation et mon bien être merci d'être toujours pour moi le père, le frère et l'ami.

Je vous dédie ce travail en témoignage de mon profond amour, estime, reconnaissance et respect *puisse Dieu, le tout puissant, vous préserver et vous accorder la foi, santé, bonheur et longue vie. Votre présence seule me suffit, je vous aime énormément.*

A mes chères et adorables sœurs : *Wafaa, Imene et Sarah ; les mots ne suffisent guère pour exprimer l'attachement, l'amour que je porte pour vous. En témoignage de mon affection fraternelle, ma profonde tendresse et reconnaissance, je vous souhaite une vie pleine de bonheur et succès et que Dieu, le tout puissant, vous protège et vous garde.*

A mon cher neveu : *M.Oussama la prunelle de mes yeux rien que ta joie et ta gaieté me comblent de bonheur. Puisse Dieu te garder, éclairer ta route et t'aider à réaliser à ton tour tes vœux les plus chers.*

A mon cher beau-frère : *Kamel bien que tu aimes être discret et passer inaperçu, je sens toujours ta veille et ta protection pour moi, je te sens un frère et je suis sûre que je peux compter sur toi, Je te dédie ce travail en te souhaitant que du bonheur et de santé.*

A ma grande mère chérie *: Qui m'a accompagné par ses prières, sa douceur, puisse Dieu lui prêter longue vie et beaucoup de santé et de bonheur dans les deux vies.*

A mes chers oncles, tantes, leurs époux et épouses, a mes chers cousins et cousines : *Veuillez trouver dans ce travail l'expression de mon respect le plus profond et mon affection la plus sincère.*

A ma chère amie *: Soumia cela fait déjà un bon moment que nos chemins se suivent toujours en parallèle l'une à l'autre en témoignage de notre sincère et profonde amitié qui nous unit et des moments agréables, nos fou rires et nos pleurs que nous avons partagé ensemble, je ne peux trouver les mots justes et sincères pour t'exprimer mon affection, je te sens comme une sœur, je te dédie ce travail ainsi qu'à toute ta famille, je vous souhaite une vie pleine de santé et de bonheur et surtout de chance.*

A mon cher ami : *H.Mohamed mon conseiller qui m'a assisté dans les moments difficiles, je te suis très reconnaissante, et je ne te remercierai jamais assez pour ton amabilité, ta générosité et ton aide inestimable. Ton amitié m'est précieuse tu es pour moi un frère et un ami sur qui je peux compter. Puisse Dieu te bénir, bénir ta famille et vous combler de toute ses grâces.*

A tous mes ami(e)s ; A mes camarades de promotion : *En souvenir des bons moments passés ensemble, je regrette de ne plus vous voir aussi souvent.*

El batoul

Je dédié ce modeste travail

A mes chers parents symbole de sacrifice, de tendresse et d'amour ; sont les moindres sentiments que je puisse vous témoigner. Quoi que je fasse, je ne pourrais jamais vous récompenser pour les grands sacrifices que vous avez fait et continuez de faire pour mon éducation et mon bien être, c'est à vous que je dois cette réussite et je suis fière de vous l'offrir.

A toi mon père, école de mon enfance, qui a été mon ombre durant toutes les années des études, qui a veillé tout au long de ma vie à m'encourager, et à me protéger.

A la lumière de mes jours, la source de joie et de mes efforts, secrets de ma force, la flamme de mon cœur, ma vie et mon bonheur ; maman que j'adore. Puisse Dieu, le tout puissant, vous préserver et vous accorder santé, longue vie et bonheur.

A mes très chers frères et sœurs, Mustapha, Ismail, Hadjira, Chaimaa. En souvenir d'une enfance dont nous avons partagé les meilleurs et les plus agréables moments. Pour toute la complicité, l'ambiance dont vous m'avez entouré et l'entente qui nous unit, je dédie ce travail pour vous avec tous mes vœux de santé et de réussite.

A la mémoire de mes grands-parents, qui ont toujours été dans mon esprit et dans mon cœur, je vous dédie aujourd'hui ma réussite. Que Dieu, le miséricordieux, vous accueille dans son éternel paradis.

A toute ma famille, petits et grands, oncles et tantes, cousins et cousines.
Vous avez toujours été présents pour les bons conseils, votre affection et votre soutien m'ont été d'un grand secours au long de ma vie.

A ma chère amie El Batoul, qui était toujours à mes côtés, tu es pour moi une sœur, je ne peux trouver les mots justes et sincères pour vous exprimer mon affection et mes pensées, en témoignage de l'amitié qui nous unit et des souvenirs de tous les moments que nous avons passé ensemble, je te dédie à toi et à toute ta famille ce travail, je vous souhaite une vie pleine de santé et de bonheur.

A la personne qui m'a toujours aidé et encouragé, qui n'a cessé de me conseiller pour ma réussite ; M.Mohamed, je vous dédie ce travail ainsi qu'à toute votre famille avec tous mes vœux de santé de réussite et de prospérité.

A tous mes amis, mes camarades de promotion et tous ceux qui me sont chers.

Soumia

Remerciements

On dit souvent que le trajet est aussi important que la destination. Nos années de maîtrise nous ont permis de bien comprendre la signification de cette phrase toute simple. Ce parcours, en effet, ne s'est pas réalisé sans défis et sans soulever de nombreuses questions pour lesquelles les réponses nécessitent de longues heures de travail.

Nous tenons dans un premier temps à remercier ALLAH le tout puissant de nous avoir donné la foi, le courage et de nous avoir aidé tout au long de notre parcours éducatif.

Quoiqu'il en soit, nous avons conduit nos recherches en gardant de vue que, sans l'aide précieuse et les conseils avisés de certains passagers de cette aventure, nous n'aurions pu emprunter cette autoroute du savoir.

Nous tenons à remercier sincèrement Melle Settouti Nesma en tant que directrice de mémoire pour son attention aussi particulière, pour ses conseils avisés et son écoute qui ont été prépondérant pour la bonne réussite de ce projet ainsi pour son énergie et sa confiance qui ont été perçues pour des éléments moteurs. Nous avons vécu un plaisir incommensurable à travailler et à apprendre à ses côtés. Nous tenons vivement à lui adresser notre gratitude.

Nos vifs remerciements vont également à Mr El Habib Daho Mostafa, en tant que Co-encadreur, pour ses remarques et suggestions pour améliorer la qualité de ce mémoire.

Nos remerciements s'adressent à Mr Chikh Mohamed.Amine, qui nous a fait apprécier et aimer notre spécialité tout au long de notre formation. Nous avons beaucoup appris de son savoir, sa méthodologie et son expérience. Pour tout cela nous lui seront éternellement reconnaissantes. Avec tous nos respects et notre gratitude.

Nous remercions également Mme Benchaib Yasmine, de nous faire l'honneur de présider ce jury.

Nos remerciements à Mme Mekkioui Nawel qui a participé à l'examen de ce travail.

Nous remercions infiniment tous les enseignants qui nous ont aidé durant tout notre cycle d'études.

Sans omettre toute l'équipe CREDOM pour leurs gentillesses, leurs disponibilités et leurs soutiens. Merci d'avoir pris le temps de répondre à nos nombreuses questions.

Un grand merci à tous ceux qui ont contribué de près ou de loin pour l'élaboration de ce travail.

Résumé

Touchant près d'un million de personnes chaque année dans le monde, la maladie de Parkinson a atteint le second rang des maladies dégénératives. Ainsi, le champ de recherche s'est développé énormément pour conduire à un diagnostic médical précis. L'application des techniques d'apprentissage artificielle peut être une piste qui apparaît de plus en plus d'être très prometteuse. Ceci nous a amené à élaborer notre projet qui consiste à traiter la possibilité de l'implémentation d'un système d'aide aux médecins pour la reconnaissance de la maladie dans ces stades précoces.

Dans le cadre de notre projet de fin d'études, nous nous intéresserons à l'amélioration des performances de la classification par les forêts rotationnelles. Cette méthode combine la robustesse des arbres de décision, la puissance de l'extraction des caractéristiques tout en augmentant la précision et la diversité des arbres dans la forêt par l'utilisation de tous les paramètres. Les résultats expérimentaux appliqués sur la banque de données médicales réelles (Parkinson) montrent une efficacité dans la tâche de classification avec exactitude significative vérifiée statistiquement.

Mots clés

La maladie de Parkinson ; Forêt rotationnelle ; Méthode d'ensemble ; Forêt aléatoire ; Analyse en composantes principales ; Analyse en composantes indépendantes.

Abstract

Affecting nearly a million people each year in the world, Parkinson's disease has reached the second rank of degenerative diseases. Thus, the research field developed tremendously to lead to a precise medical diagnosis. The application of artificial learning techniques may be a track that appears increasingly to be very promising.This led us to develop our project, which comprises treating the possibility of the implementation of a system of assistance to doctors for the recognition of the disease in the early stages.

As part of our project of graduation, we will focus on improving the performances of classification by the rotational forests. This method combines the robustness of decision trees, the power of the feature extraction while increasing accuracy and tree diversity in the forest by using all parameters. The experimental results applied to the actual medical database (Parkinson) show effectiveness with significant accuracy in the classification verified statistically.

Keywords

Parkinson's disease ; Rotational forest ; Overall method ; Random forest ; Principal component analysis ; Independent component analysis.

Table des matières

Table des figures

Liste des tableaux

Glossaire

ACI : Analyse en Composantes Indépendantes
ACP : Analyse en Composantes Principales
ADN : Acide Désoxyribonucléique
Bagging : Bootstrap AGGregatING
BDD : Base de données
CART : Classification And Regression Tree
CFS : Correlation based feature selection
DFA : Detrended Fluctuation Analysis
FP : Faux positif
FN : Faux négatif
IRM : Imagerie par résonnance magnétique
LDA : Analyse discriminante linéaire
LSVT : Lee Silverman Voice Treatment
LTI : Linéair Time Invarient
MFCC : The mel-frequency cepstral coefficients
MP : Maladie de Parkinson
NDA : Analyse discriminante non paramétrique
OOB : Out-Of-Bag
PLM : Patients Like Me
PPE : Pitch Period Entropy
RF : Forêts Rotationnelles
RPDE : L'entropie de la densité de la période de retour
SE : Sensibilité
SP : Spécificité
SVM : Support Vector Machine
TC : Taux de classification
UCI : University California Irvine
UPDRS : Unified Parkinson Disease Rating Scale
VP : Vrai positif
VN : Vrai négatif

Introduction générale

La maladie de parkinson (MP) est l'une des principales affections dégénératives du système nerveux central ; décrite par James Parkinson en 1817 [1], cette affection est responsable de troubles essentiellement moteurs et caractérisée par la dégénérescence d'une population de cellules nerveuses situées dans la substance noire.

Un diagnostic médical est le résultat du raisonnement d'un médecin avec une décision très souvent prise à partir de plusieurs caractéristiques. Pour assurer un diagnostic exacte un système d'aide au diagnostic permet de guider les médecins en réduisant au maximum les erreurs possibles qui peuvent survenir pendant le diagnostic d'une maladie à l'aide d'un outil informatique performant dont toutes les données sont référentielles au médecin pour analyser la pertinence des hypothèses proposées après saisie de signes et/ou des examens complémentaires tout en sollicitant la médiation de l'intelligence artificielle qui a pour objectif de munir un caractère intelligent.

Dans un tel domaine, de nombreuses méthodes de classification ont été appliquées, cherchant toujours à améliorer et augmenter l'efficacité et l'interprétabilité. Aujourd'hui, les chercheurs s'intéressent de plus en plus aux méthodes d'ensemble qui ont continuellement un potentiel pour améliorer la précision de la classification grâce à leur robustesse et capacité de préserver l'information de variabilité des données.

Le principe de base des méthodes d'ensemble est de peser plusieurs classifieurs de motifs individuels, et de les agréger afin de parvenir à une classification qui est meilleure que celle obtenue par chacune d'entre elles séparément, l'intelligence collective ajoute au comportement individuel jugé insuffisant à l'influence du groupe, elle reflète l'émergence d'un comportement global en partant d'un groupe de classifieurs de base identique.

Dans ce projet de fin d'études, nous nous intéressons au domaine de la reconnaissance de la maladie de Parkinson qui est devenu l'un des sujets de recherche ressentie comme un ébranlement et va susciter de nombreuses interrogations en classification ; pour cela ce mémoire sera réparti comme suit :

- Chapitre 1 Contexte médicale, présente un aperçu général sur la maladie, sa reconnaissance par traitement à partir de la voix ainsi l'objectif de ce travail.

- Chapitre 2 Etat de l'art, fait le tour des différents travaux existants sur la maladie de Parkinson dans le traitement de signal, la télémédecine et l' intelligence artificielle. Egalement, la littérature de l'approche proposée, il retrace aussi notre contribution mise en œuvre et les motivations dans ce domaine.

- Chapitre 3 Principe des forêts rotationnelles, invoque les fondements théoriques des méthodes utilisées.

– Chapitre 4 Expérimentation et Résultats, décrit la population participante à notre étude et le protocole d'expérimentation, aussi bien que les résultats et leurs interprétations permettront de valider ou non notre hypothèse de départ.

– En dernier lieu, une conclusion générale et les perspectives à venir dans ce travail.

Chapitre 1

Présentation de la maladie de Parkinson

La maladie de Parkinson constitue un enjeu de santé publique mondial, son incidence augmente constamment en raison du vieillissement de la population. Avec d'autres maladies neuro-dégénératives comme la maladie d'Alzheimer, on s'attend à ce qu'elle dépasse le cancer au deuxième rang de la mortalité vers l'année 2040 [2].

Dans l'état actuel de la recherche, les causes précises ne sont pas connues, néanmoins, il est certain qu'elle n'est pas une maladie contagieuse. Elle semble n'être héréditaire que dans peu de cas. Les manifestations de la maladie sont variables : chaque personne atteinte aura son ou ses symptômes prédominants.

Actuellement, la maladie ne peut malheureusement être guérie, mais il existe tout un éventail de traitements permettant d'en atténuer les symptômes [3].

1 Contexte médicale

La maladie de Parkinson (MP) est l'une des affections neuro-dégénératives les plus fréquentes considérée comme la conséquence de la destruction relativement sélective du système dopaminergique nigrostriatal [1], elle affecte environ 1% des individus de plus de 60 ans dans le monde [2].

1.1 Épidémiologie

Incidence et Prévalence

l'incidence est basée sur des études de prévalences publiées [4], en utilisant deux méthodes différentes pour projeter le nombre de personnes ayant la maladie de Parkinson dans cinq nations en Europe de l'ouest et dix nations les plus peuplées du monde. Le nombre des individus avec la MP qui ont plus de 50 ans dans ces pays se situait entre 4,1 et 4,6 millions en 2005 et va croitre considérablement entre 8,7 et 9,3 millions en 2030 comme il est illustré dans la figure 1.1) ; le tableau 4.1 englobe en chiffres l'étude réalisée :

1. Nigrostriatal : Système moteur extrapyramidal dans le cerveau qui régule le tonus musculaire.

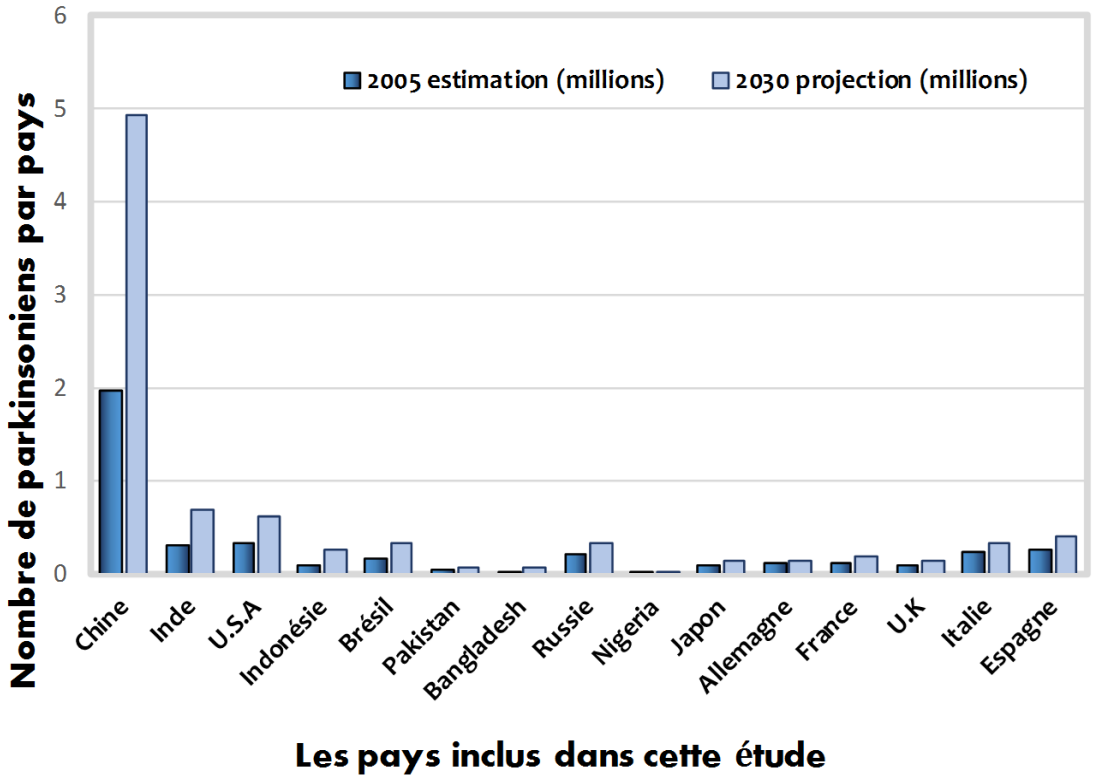

FIGURE 1.1 – Projection de la maladie de Parkinson dans les pays les plus peuplés 2005-2030.

	la prévalence par strates d'âge/100.000						
Pays	50-54	55-59	60-64	65-69	70-74	75-79	>80
Chine	NA	320	320	1.130	1.130	2.740	2.740
Inde	128	128	260	260	260	260	260
U.S.A	128	128	128	550	550	958	958
Indonésie	50	50	280	280	510	510	1.250
Brésil	371	371	443	443	443	443	443
Pakistan	128	128	260	260	260	260	260
Bangladesh	128	128	260	260	260	260	260
Russie	127	127	493	493	1.232	1.232	1.109
Nigeria	58	58	58	58	58	58	58
Japon	20	64	97	196	322	525	341
Allemagne	NA	NA	NA	0	700	1.800	700
France	NA	NA	NA	500	400	1.800	2.200
U.K	76	111	159	343	664	859	1.044
Italie	116	116	621	621	1978	1.978	3.055
Espagne	0	0	630	630	1300	1.300	10.400

TABLE 1.1 – Prévalence et incidence de la maladie de Parkinson dans les pays les plus peuplés 2005-2030.

Origine multi-factorielle

Les causes de la maladie de Parkinson ne sont pas encore déterminées mais sont probablement multi-factorielles [5]. Même si le mécanisme du déficit en dopamine est bien compris, les causes restent encore inconnues, ce qui provoque souvent chez le malade des difficultés à accepter sa maladie.

La survenue de la maladie de Parkinson apparaîtrait alors lors de la conjonction des facteurs génétiques puisque dans 15 % des cas, des antécédents familiaux de MP sont retrouvés. On parle plutôt de prédisposition génétique que d'hérédité, et de facteurs environnementaux tels que l'exposition à des produits chimiques (herbicides, insecticides, pesticides) ou métaux lourds (plomb, manganèse, cuivre) ont été mis en évidence.

1.2 Physiopathologies

La maladie de Parkinson est une maladie neurologique chronique due à un déficit en dopamine, un neurotransmetteur indispensable au contrôle des mouvements du corps, en particulier des mouvements automatiques. Elle se manifeste par un ensemble de symptômes et une évolution variable d'un individu à l'autre.

Symptômes

- *Troubles moteurs* : le syndrome parkinsonien comporte un syndrome moteur défini par la présence d'une akinésie [2], associée à au moins l'un des symptômes suivants : tremblement de repos, rigidité plastique, instabilité posturale [3].

- *Troubles axiaux* : les troubles axiaux sont les troubles de la marche, de la posture, de la parole et de la déglutition qui apparaissent plus tardivement dans l'évolution de la maladie et ont un retentissement psychologique important [6].

- *Troubles cognitifs* : les troubles cognitifs sans démence sont présents dans la MP, de manière discrète au stade précoce, et s'intensifient au fur et à mesure de l'évolution de la maladie [6]. Dans la plupart des cas, il n'y a pas de déclin cognitif global mais plutôt « une perte de rendement intellectuel accompagnée d'un déclin amnésique », un ralentissement de la pensée. Ces troubles touchent généralement la vitesse de traitement de l'information, l'attention, la mémoire et les fonctions exécutives, et prennent la forme d'un « syndrôme sous-corticofrontal » qui réduit les capacités de supervision de l'action.

Diagnostic

Le diagnostic n'est pas évident, surtout au début de la maladie car la maladie de Parkinson se développe généralement de manière progressive, et il peut passer un grand nombre de mois, voire d'années, avant que les symptômes ne deviennent assez gênants pour en faire part à son médecin, mais malheureusement il n'existe pas de test définitif pour confirmer la maladie de Parkinson et le diagnostic est fondé uniquement sur les symptômes et un examen clinique qui sont habituellement insuffisants. Cependant, il peut être nécessaire de réaliser d'autres tests (analyses sanguines) et examens radiologique (scanner, IRM...) pour éliminer la possibilité d'autres causes médicales des symptômes présents.

2. Akinésie : La rareté et/ou la lenteur à l'initiation d'un mouvement.

Traitement

À l'heure actuelle, aucun remède pour la maladie de Parkinson n'existe. Cependant, plusieurs symptômes moteurs sont traités au moyen de médicaments, ceux-ci utilisés, essentiellement, afin de compenser le déficit de dopamine ou imiter les effets de la dopamine dans le cerveau. Les médicaments permettent d'atténuer les symptômes, mais sans freiner la progression de la maladie d'où le dosage des médicaments est modifié en fonction de l'évolution des symptômes.

Pour traiter les symptômes de la maladie de Parkinson, on a parfois recours à la chirurgie du cerveau par une intervention chirurgicale appelée « stimulation profonde du cerveau » ; la procédure comprend l'insertion d'une sonde dans le cerveau qui cible les régions spécifiques pouvant contrôler les tremblements ou les mouvements involontaires.

Prise en charge

Pour parler de la prise en charge d'un patient atteint de la maladie de Parkinson, il faut tout d'abord mentionner que cette dernière est à caractère majoritairement ambulatoire [7]. En effet, le traitement d'un patient atteint de cette maladie ne nécessite pas, sauf exception d'hospitalisation tout comme d'autres pathologies dégénératives du système nerveux telles que la maladie d'Alzheimer ou la sclérose latérale amyotrophique.

La prise en charge d'une affection comme la maladie de Parkinson nécessite un éventail très large de compétences, qui s'étend du neurologue à l'aide à domicile, en passant par l'ergothérapeute, le physiothérapeute, le médecin de premier recours, et tant d'autres.

2 Parkinson et la dysarthrie

Les troubles de la parole sont significatifs dans la maladie de parkinson : la dysarthrie fait partie des signes axiaux de la maladie qui sont réputés d'être peu sensibles aux traitements médicamenteux et chirurgicaux d'où la voix est le résultat d'une coordination du larynx, du diaphragme, des cordes vocales, de la langue et des lèvres ; chez un malade de Parkinson, cette coordination est altérée.

La voix a des rigidités, des faiblesses et des tremblements ; elle est un bon marqueur de la maladie, il est possible qu'elle soit même l'une des premières fonctionnalités affectées, La dysarthrie hypokinétique [8] reste considérée comme caractéristique des troubles de la production vocale observés au cours de la maladie de Parkinson.

2.1 La dysarthrie hypokinétique

La dysarthrie fait partie des symptômes axiaux au même titre que les troubles de l'équilibre caractérisée par des troubles de l'exécution motrice de la parole, dont l'origine est une lésion du système nerveux central et/ou périphérique. Cette dysarthrie parkinsonienne, ou hypokinétique [9] est provoquée par la bradykinésie [3] et la rigidité des muscles de la respiration et de la phonation, elle est d'apparition progressive et d'aggravation lente.

3. Bradykinésie : Ralentissement de l'excusion des mouvements et la perte de la finesse du mouvement, comme l'écriture.

Le terme « hypokinétique » fait référence aux mouvements articulatoires réduits et à la diminution de la modulation prosodique qualifiée de monotone aussi distingué par trois principales caractéristiques tels que : une dysphonie parkinsonienne correspond aux anomalies de fonctionnement du vibrateur laryngé, la prosodie a un rôle important dans la communication, elle permet d'exprimer les attitudes, les sentiments, les émotions et des troubles d'articulation. Ces derniers, sont moins fréquents que les troubles phonatoires et prosodiques, et ils sont corrélés au degré de dysarthrie.

La dysarthrie parkinsonienne existe à différents degrés [8], selon l'évolution de la maladie. Elle est « légère » lorsqu'il y a uniquement des troubles de la prosodie et des modifications de la qualité vocale. On la qualifie de « modérée » quand vient s'associer aux troubles précédents une intelligibilité gênée par des troubles d'articulation.Enfin, la dysarthrie est « sévère » dès que la parole n'est plus fonctionnelle, avec des troubles de l'initiation de la parole, une aphonie [4], ou des bribes courtes mal articulées.

3 Reconnaissance de Parkinson par traitement à partir de la voix

La production de la parole met particulièrement en évidence les notions d'automatisation (après acquisition et apprentissage) et d'organisation séquentielle ; d'où la production est un système dynamique dont le comportement à un moment donné dépend de ses états antérieurs.

Plusieurs études s'accentuaient de manière à démontrer l'intérêt de l'analyse de la voix afin de prouver sa relation avec la maladie de Parkinson tel que cette dernière sera affectée au fil de son évolution. Souvent, elle devient plus ténue et indistincte. Beaucoup de patients ne remarquent pSr ce changement. Il est possible d'améliorer le dépistage en travaillant de manière ciblée la voix.

Parmi les différents travaux pour la reconnaissance de la maladie de Parkinson, émerge un travail bien à part du mathématicien Max Little qui est connu pour son projet sur l'initiative voix de la maladie de Parkinson [10], dans lequel lui et son équipe de l'université britannique d'Oxford ont développé un outil simple qui utilise un logiciel d'analyse précise de la voix d'une manière scandaleusement facile à dépister la maladie de Parkinson. Il serait aussi simple que de décrocher le téléphone et dire «aaah» [11].

L'idée que propose Little et al. [12] est de diagnostiquer la maladie de Parkinson via des enregistrements vocaux collectés qui contiennent suffisamment d'informations pour détecter la maladie dans des stades précoces, puisque les perturbations de la voix pourraient bien être l'un des premiers indicateurs de la maladie.

En revanche aux tests cliniques couteux et qui prennent du temps, cette technique pourrait être le moyen le plus simple et non invasive pour diagnostiquer la maladie de Parkinson ainsi qu'une façon et une possibilité très tentante pour aider les patients de bénéficier d'un traitement antérieur.

4. Aphonie : L'incapacité d'émettre de la voix au-delà du chuchotement.

4 Conclusion

La maladie de Parkinson bénéficie principalement des efforts de la recherche fondamentale, cela signifie que le processus d'investigation se déroule continuellement dans ce domaine tel que les progrès qui sont liés à la détection de signes cliniques qui pourraient être annonciateurs de la maladie de Parkinson.

Un grand appui aujourd'hui consiste à trouver des moyens pour détecter précocement la maladie d'où il est possible à présent de réaliser un diagnostic plus préalable qui permettrait sans doute à terme de ralentir voire d'enrayer l'évolution de la maladie bien avant le stade où elle est visible.

Afin de répondre à ce besoin, plusieurs recherches dans cette possession ont mis au point des méthodologies et des outils pour développer des systèmes intelligents de manière à résoudre des problèmes complexes en faisant appel à diverses disciplines telles que la médecine, l'informatique et l'intelligence artificielle. . .

Dans ce cadre, notre travail donne un intérêt plus particulier à l'apprentissage automatique afin d'établir un diagnostic médical à partir d'un ensemble de descripteurs cliniques d'un patient parkinsonien ; en d'autres termes la procédure générée devra classifier correctement les paramètres du patient mais surtout avoir un bon pouvoir prédictif pour classifier correctement de nouvelles descriptions.

Chapitre 2

État de l'art

1 Introduction

Le diagnostic assisté par ordinateur en médecine est un nouveau domaine ayant une grande importance pour fournir le pronostic des maladies, à travers des systèmes de classification qui permettent de rechercher une éventuelle solution pour des problèmes médicaux difficiles et non résolus. Ces systèmes pourraient offrir d'avantages de facilités aussi bien pour les patients que pour les médecins, cependant la conception d'un tel système nécessite l'intervention de l'apprentissage automatique qui a pour but de doter ce dernier d'un comportement intelligent.

En raison des difficultés que pose le cerveau à la recherche, la maladie de Parkinson bénéficie principalement des efforts de la recherche fondamentale ; c'est la diversité des symptômes dans la constitution des syndromes individuels, qui rend difficile une recherche appliquée directe.

Dans ce chapitre, un état de l'art des différentes études existantes sur la maladie de Parkinson (MP) est exposé. Nous présenterons par la suite, nos motivations et contribution dans ce domaine à des fins d'aide au diagnostic de façon complémentaire pour la reconnaissance de Parkinson.

2 L'état de l'art de la maladie de parkinson

Dans la littérature, plusieurs travaux traitent le sujet d'aide au diagnostic pour la maladie de Parkinson. Nous constatons que les différents travaux de recherches ciblant ce sujet sont répartis sur trois grandes disciplines portant respectivement sur le traitement de signal, la télé-médecine et l'intelligence artificielle.

2.1 Parkinson et le traitement de signal

La première catégorie du traitement de signal porte sur les changements possibles affectant les signaux physiologiques durant la maladie de Parkinson :

Harel et al. [13] ont présenté une première tentative pour identifier les changements possibles de la parole lors d'une phase prodromique [1] de la MP. Pour mieux comprendre

1. La phase prodromique désigne la phase débutante d'une pathologie.

la sensibilité au changement de la fréquence fondamentale de la voix, car une diminution de cette dernière peut être détectée même avant cinq ans du diagnostic clinique.

Dans le même contexte, Little et al. [14] ont suggéré une détection automatique de la pathologie via des enregistrements de l'acoustique de la parole en temps réel à l'aide de deux approches non linéaires RPDE : (*L'entropie de la densité de la période de retour*), DFA : (*Detrended Fluctuation Analysis*) et linéaire LDA (*Linear Discriminenet Analysis*), cette étude indique une performance globale de 98,2% en combinant ces approches.

Dans une continuité de ces travaux, Little et al. [15] appliquent l'analyse numérique de la parole à l'aide des méthodes de traitement de signal non linéaires qui ont prouvé leurs efficacité spécialement dans le cas où les hypothèses linéaires ne sont pas appropriées pour tous les signaux de parole. Cette étude donne plus d'importance aux systèmes LTI (*Linear Time Invarient*) dont la sortie ne dépend pas explicitement du temps.

Par la suite, Little [16] présente deux nouveaux outils pour l'analyse de la parole : la méthode de récurrence et celle de mise à l'échelle fractale, il démontre également que ces nouvelles mesures permettent d'atteindre une performance globale de classification supérieure aux mesures de perturbation classiques existantes. Les résultats montrent que les nouvelles mesures non linéaires sont plus précises que les mesures traditionnelles et les composantes du bruit .

Afin de confirmer que les troubles moteurs dans la MP entraînent des changements acoustiques indésirables et variés affectant un certain nombre de contrastes prosodiques [2] dans la parole et que ces modifications semblent se produire dans les premiers stades de progression de la maladie. Henry et al. [17] ont évalué globalement des caractéristiques acoustiques des énoncés produites par haut-parleurs dans les premiers stades de la MP idiopathique [3]. Ils ont comparé ces caractéristiques aux participants appariés sans la maladie. Henry et al. se sont focalisés sur les changements acoustiques possibles dans la production de la parole dont ces derniers produisent différentes distinctions prosodiques qui sont communs aux interactions quotidiennes, comme la capacité de produire le stress, l'accent, et les émotions grâce à la modulation des paramètres appropriés de la voix.

Dans [18], Max Little propose une nouvelle méthode de mesure des dysphonies appelée PPE (*Pitch Period Entropy*), pour détecter une certaine forme de déficience vocale qui peut également être l'un des premiers indicateurs pour le début de la maladie de Parkinson. La combinaison de nouvelles mesures non-standard (HNR, RPDE, DFA, PPE) a conduit aux meilleures performances par le classifieur SVM (*Support Vector Machine*) avec un taux de reconnaissance de 91,4%.

Pareillement, Jan Rusz [19] a cherché à trouver la meilleure compréhension du rôle de troubles de la parole dans la MP à l'aide des méthodes innovantes d'analyse acoustique, des techniques de traitement du signal de parole, et statistiques avancées. Les résultats montrent que 80 à 90% des premiers sujets de MP non traités montrent une forme de déficience vocale, ce qui confirme que la parole peut être un marqueur important de la progression de la maladie et l'efficacité du traitement dans la MP.

2. La prosodie est généralement associée au rythme (accentuation) et à l'intonation (mélodie) produits par un locuteur.

3. Idiopathique désigne un symptôme ou une maladie présentant une origine inconnue.

Gillivan-Murphy [20] a,quant à lui, donné une vue générale sur MP en s'intéressant au tremblement de la voix dans cette dernière qui se caractérise comme un trouble de l'exécution motrice de la parole, tout en essayant de répondre à un ensemble de questions qui peuvent se résumer en comment déterminer la voix comme caractéristique de MP pour éclaircir au mieux cette problématique en comparant deux groupes de sujets parkinsoniens et sains, aussi voir si la différence est statistiquement significative et de déterminer les moyens optimaux dont elle devrait être mesurée.

Dans une autre étude, Taha Khan [21] vise à utiliser MFCC (*The mel-frequency cepstral coefficients*) pour la classification de la sévérité des symptômes de la parole selon l'échelle de notation de la maladie de Parkinson UPDRS(*Unified Parkinson Disease Rating Scale*). Il effectue une analyse comparative entre la performance de classification des MFCC, calculée à partir des tests enregistrés TRS et MFCC calculé à partir des essais enregistrés SVP et DDK en utilisant la méthode SVM. L'étude revèle que MFCC des TRS sont meilleurs classifieurs de symptômes de la parole et donne les meilleures performances (84% et 85%).

Little et al. [22] ont amélioré l'efficacité du traitement de la parole de réadaptation en développant un algorithme approprié pour le système Companion LSVT (*Lee Silverman Voice Treatment*) qui est capable de détecter des caractéristiques vocales inacceptables lors de l'utilisation des logiciels à partir des conseils d'experts cliniques. Leur approche permet d'aider le patient a ne pas utiliser la voix d'une manière inacceptable, et par la suite améliorer ces caractéristiques de la voix grâce à la rétroaction.

2.2 Parkinson et la télé-médecine

La deuxième catégorie est la télé-médecine, repose sur des tests auto- administrés sur des appareils mobiles pour effectuer un diagnostic et un suivi de la maladie à domicile même.

Dans un premier temps, Tsanas et al. [23] ont exploité l'idée que UPDRS peut être objectivement évalué à l'exactitude clinique en utilisant des tests vocaux auto-administrés. Ils appliquent une large gamme d'algorithmes de traitement du signal de parole à une grande banque de données, les résultats obtenus ont fourni une bonne preuve statistique démontrant que les troubles de la parole sont inhérents, et intuitivement justifiant la condition qu'UPDRS peut être prédite par l'analyse de signaux de parole.

Little et al. [24] ont exploré la possibilité d'utiliser la norme de la voix sur GSM (2G) ou UMTS (3G) de téléphonie mobile cellulaire pour la télésurveillance des parkinsoniens. Ils ont testé la robustesse de cette approche en utilisant une communication mobile bruyante simulé réseau. Ils améliorent leur approche dans [25], en réalisant une quantification de la progression de la maladie de Parkinson par traitement depuis des sites internet en évaluant la valeur clinique de ces données à l'aide de l'échelle UPDRS. Ce dernier est un outil d'évaluation global du retentissement de la maladie en faisant appel a deux sources de données PLM (*Patients Like Me*), (PD-DOC).

Récemment, Graça [26] était conscient qu'à ce jour tout type de diagnostic pour la MP se fait par l'observation d'un professionnel de la santé spécialisé dans ce domaine. De ce fait, il est nécessaire de développer une méthode qui est simple et efficace pour les professionnels de la clinique générale de soins de santé afin qu'ils puissent avoir une sauvegarde pour décider de transmettre un patient possible à un spécialiste. Dans ce contexte, une application mobile où un professionnel des soins de santé peut intégrer les symptômes de

patients possibles ou le patient lui-même peut réaliser un test. L'étude réalisée dans [26] sur 35 sujets testés avec différents algorithmes d'apprentissage automatique a montré des résultats prometteurs dans l'ensemble, mais pas encore suffisants pour justifier l'utilisation dans le cadre pratique réel.

2.3 Parkinson et les techniques d'intelligence artificielle

La dernière catégorie concerne les techniques d'intelligence artificielle qui vise continuellement a optimisé la classification automatique de la maladie de parkinson en faisant appel aux processus d'induction et de prédiction ...

Jan rusz et al. [27] fournissent une classification basée sur le théorème de Bayes pour séparer les patients sains, des personnes atteintes de la MP en évaluant leurs dépréciations vocales dégradables dans le cadre de la dysarthrie lier à MP. L'étude a pu atteindre un taux de 91.30% en classification.

Dans une autre étude, Jan rusz et al. [28] mettent l'accent sur les différentes mesures de la parole pour évaluer la dépréciation vocale en identifiant les signatures acoustiques liées à MP à l'aide d'une tâche [29] d'évaluation de pertinence des mesures individuelles qui a été confirmée fiable dans cette étude .

Akin Ozcift et al. [30] ont effectué une classification en utilisant les forêts rotationnelles sur la banque de données de Parkinson du Max Littel ainsi en avant ils ont utilisé une étape de sélection de variables par CFS (*correlation based feature selection*)où ils ont obtenu que dix caractéristiques comme des variables les plus pertinentes. Avec 30 ensembles de classifieurs et 10 validations croisées ils ont atteint une moyenne du taux de classification de 84,4% sans sélection et 87,1% avec sélection, c'est une augmentation moyenne de 2,7% en précision globale.

Hazan et al. [31] montrent que des outils d'apprentissage artificiel peuvent être utilisés pour fournir un diagnostic précoce de la maladie de Parkinson à partir de deux ensembles de données distincts de la parole (des Etats-Unis et d'Allemagne) en utilisant un classifieur classique SVM. Ces données de la parole peuvent fournir souvent une détection précoce de la maladie de Parkinson avec un taux de 90% .

Récemment, Boublenza et al. [32] ont pu atteindre une meilleure performance avec un système de reconnaissance automatique des caractéristiques de la MP. Une sélection de variables par Relief-F est mise en œuvre, pour réduire le nombre de variables et augmenter les performances du système. Les résultats expérimentaux montrent que la performance de l'approche proposée réalise un taux de 96, 88% à l'aide du classifieur SVM pour la reconnaissance précoce de la maladie de Parkinson.

Little et al. [33] ont effectué une évaluation de l'efficacité des Smartphones utilisés comme un outil d'aide au diagnostic précis pour la discrimination des participants Parkinsoniens et des participants témoins sains via des tests auto-administrés de la démarche et balancement postural en appliquant le classifieur « forêts aléatoires » [34]. Ils ont obtenu un taux de sensibilité de 98.5 % et un taux de spécificité de 97.6 %.

Une continuité de ce travail a été réalisé en [35] pour vérifier statistiquement les résultats obtenus par le classifieur forêts aléatoires en le comparant avec un autre classifieur aléatoire et un classifieur aléatoire conditionnel dont les forêts aléatoires ont montées encore une fois leur efficacité dans les tâches de classifications.

Dans une autre étude Parkinson's voice initiative [36], l'équipe de Max Little visent à prédire avec précision la sévérité des symptômes de la maladie de Parkinson selon une échelle clinique (UPDRS), elle permet de dépister la maladie grâce à des enregistrements de phonations prolongés (les sons aaah) à travers des lignes audio numériques avec une précision maximum de 98.6% et une erreur de prédiction moyenne de 3,5 points sur l'échelle.

3 Motivations

À la lumière des récentes découvertes du mathématicien Max Little concernant la reconnaissance automatique de MP par la voix, jugées de plus en plus prometteuses. Ce travail poursuit la même piste, en continuité de notre projet de fin d'études de Licence [37] qui constituait un premier pas dans le développement d'un système intelligent d'aide au diagnostic médical afin d'aider les experts pour la détection précoce de Parkinson. Ce travail consistait à mettre en œuvre une classification par réseaux de neurones pour fournir une détection préalable de la maladie où la capacité de réseau neuronal a mis en évidence son efficacité pour un apprentissage plus rapide avec une structure réduite. Grâce à la méthode de réduction : "ACI" Analyse en Composantes Indépendantes les résultats atteignent le taux de classification de 93.75 % ; avec trois composantes indépendantes, ainsi que les variables pertinentes ont pu être extraites.

Ce projet de fin d'études de Master suit cette même vision de recherche, de ce fait, nous proposons de réaliser une classification de la maladie de Parkinson par les méthodes d'ensemble. Ces dernières ont reçu une attention croissante dans le domaine d'apprentissage de la machine grâce à leur potentiel d'améliorer considérablement la performance de prédiction d'un système intelligent en traitant plusieurs ensembles d'apprentissage en parallèle, permettant d'améliorer l'ajustement par une combinaison ou agrégation d'un grand nombre de modèles qui s'apparente du fameux dicton de la vie réelle ˝l'union fait la force˝.

Plus particulièrement, la méthode des forêts rotationnelles vise à mettre en œuvre un ensemble de classifieurs à la fois précis et variés. Sa principale heuristique consiste dans l'application de l'extraction de caractéristiques de sous-ensembles et la reconstruction d'un ensemble complet de variables pour chaque classifieur de base dans l'ensemble, les arbres de décision ont été choisis ici car ils sont sensibles à la rotation des longs axes.

4 L'état de l'art des forêts rotationnelles

Au cours des deux dernières décennies plusieurs travaux se sont intéressés aux potentiels de cette approche considérée comme une méthode d'ensemble nouvellement proposée en comparaison avec d'autres systèmes d'ensemble, elle est plus robuste, car elle conduit généralement à améliorer la précision des classifieurs individuels et la diversité dans l'ensemble en même temps [38].

Rodriguez et al. [39] sont les initiateurs de cette nouvelle méthode d'ensemble, nommée forêt rotationnelle, qui vise à mettre en œuvre un ensemble de classifieurs précis et variés, elle transforme les données avec différents axes tout en préservant les informations complètes dans laquelle l'ensemble d'apprentissage pour chaque classifieur de base est formé en appliquant ACP (*Analyse en Composantes Principales*). Dans [39], les auteurs testent les capacités de cette approche en comparaison avec plusieurs méthodes d'ensemble : Bagging (*Bootstrap AGGregatING*) [40], AdaBoost [41], et forêts aléatoires [34] sur un jeu de données de 33 banques de référence à partir de l'UCI (*University California Irvine*). Les résultats ont donné l'avantage aux forêts rotationnelles avec 23 victoires sur 33 ; ces performances sont jugées statistiquement significatives en faveur de la forêt rotationnelle.

Aussi Ludmila et al. [42] portent une étude de la lésion sur les forêts rotationnelles, cette approche permet de savoir lequel des paramètres et heuristiques de randomisation sont responsables de la bonne performance et explore l'effet des choix de conception et les valeurs des paramètres sur les performances des ensembles de classifieurs dans la forêt rotationnelle. la conclusion de ce travail est que les caractéristiques extraites par ACP sont meilleures que les alternatives d'extraction de caractéristiques comme NDA (*analyse discriminante non paramétrique*) ou projections aléatoires.

Dans une autre étude, Kun-Hong et al. [43] suggèrent la méthode des forêts rotationnelles comme une nouvelle théchnique pour lutter contre le problème de la classification des données de puces à ADN (*Acide Désoxyribonucléique*) en l'appliquant sur deux ensembles bien connus : ensemble de données du cancer du sein et l'ensemble de données du cancer de la prostate ; où ils ont pu constater que l'ACI est plus sensible que l'ACP et RF (*Forêts Rotationnelles*) concernant la rotation des forêts lorsqu'il s'agit de jeux de données de puces à ADN, et pour évaluer la performance des forêts rotationnelles, Bagging et Boosting [44] ont également été appliquées à titre de comparaison. Les résultats expérimentaux ont indiqué que la méthode des forêts rotationnelles est robuste en classification de puces à ADN, qui conduit généralement à la plus grande précision puisque l'idée de l'approche de rotation est d'encourager la précision à la fois individuelle et la diversité dans l'ensemble. La diversité est favorisée par l'extraction de caractéristiques pour chaque classifieur de base en particulier pour les ensembles de petites tailles, en outre, la méthode des forêts rotationnelles basées sur ACI est une approche nouvelle et plus robuste.

Par la suite Zhang et al. [45] présentent une nouvelle technique de génération des méthodes d'ensemble RotBoost, qui est définie simplement comme une combinaison de AdaBoost et les forêts rotationnelles ainsi faire preuve que RotBoost est considérée comme la meilleure parmi les procédures de classification d'ensemble puisqu'elle a démontré une amélioration significative que ce soit contre les forêts rotationnelles, Boosting, MultiBoost ou AdaBoost à travers les 36 ensembles de données UCI utilisés ; cette amélioration de la précision de la prédiction réalisée par RotBoost est obtenue avec une augmentation négligeable dans les coûts de calcul. En fait, Rot- Boost offre un avantage potentiel de calcul sur AdaBoost dès qu'il se prête à l'exécution en parallèle de plus, l'étude a pu révéler que ce dernier est rapide, simple et ne nécessite aucune connaissance préalable puisqu'il n'a pas de paramètres à régler.

Également, Zhang et al. [46] présentent une nouvelle méthode de génération de classifieur d'ensemble qui peut être considérée comme une combinaison de forêt rotationnelle et du Bagging par conséqent la principale différence entre la forêt rotationnelle et la méthode proposée réside dans la construction des classifieurs, en ce qui concerne l'approche clas-

sique, elle utilise directement la formation transformée, tandis que leur approche fournit un sous-échantillon tiré au hasard à partir de la transformée pour former les classifieurs puis comparé cette dernière avec plusieurs autres générations de classificateur d'ensemble (Bagging, AdaBoost et Forêts rotationnelles) en faisant appel aux 33 ensembles de données à partir du référentiel de l'UCI ; alors les résultats ont montré que le nouvel algorithme est plus robuste au bruit que les autres méthodes, les forêts construites par l'approche proposée, ont réalisé des erreurs de prédiction beaucoup plus faibles que celles construites par la forêt rotationnelle classique sur tous les ensembles de données.

En effet, dans un autre travail, Kotsiantis [47] réalise une combinaison des méthodes d'ensemble Bagging, le Boosting, les forêts rotationnelles et l'ensemble de sous espaces aléatoires et évalue sa performance individuelle avec 3 classifieurs de base «l'arbre C4.5 » , « réseau bayésien» et « part» , où en déduisant que leur combinaison a donné de meilleurs résultats dans la plupart des ensemble de données et avec les trois classifieurs de base et que la clé de la réussite des méthodes d'ensembles est la diversité au sein des classifieurs, autrement dit un ensemble de classifieurs tente d'exploiter le comportement local des différents classifieurs de base pour améliorer la précision et la fiabilité de l'ensemble du système d'apprentissage inductif.

Et finalement Chun-Xia Zhang et al. [48] examinent la performance de la méthode d'ensemble forêt rotationnelle à l'amélioration de la capacité de généralisation d'un prédicteur de base pour résoudre les problèmes de régression à travers des expériences menant à plusieurs ensembles de données et de la comparer à celle de Bagging, forêt aléatoire, Adaboost.R2, et un arbre de régression simple comme algorithme d'apprentissage de base. La principale déduction c'est que les forêts rotationnelles effectuent généralement des résultats meilleurs que les autres méthodes d'ensemble. Cependant, on ne sait jamais si la rotation des forêts se comporte bien dans la résolution des problèmes de régression puisque la sensibilité de la rotation est liée au choix des paramètres inclus dans les différentes études.

5 Contribution

De l'avis des experts, il est important de noter que la méthode dite « forêts rotationnelles » est une méthode d'ensemble de classification très réussie, fournit généralement des résultats meilleurs d'autres classifieurs d'ensemble comme le Bagging, Adaboost et les forêts aléatoires lorsque la taille de l'ensemble est relativement petite.

La taille d'arbre semble avoir un certain effet sur la performance des méthodes d'ensemble parce que les axes verticaux des bonnes parcelles sont plus grands que ceux d'emplacements laissés. En outre, l'élagage d'arbre change aussi, légèrement le classement des méthodes d'ensemble sur chaque ensemble de données.

De même, une autre méthode prometteuse de transformation linéaire : l'ACI souvent appliquée pour éliminer une grande variété d'artefacts des signaux. Il s'agit d'une méthode traitant des observations vectorielles afin d'en extraire des composantes linéaires qui soient aussi indépendantes que possible.

Cette simple idée s'est révélée très fructueuse pour le traitement des signaux dans de nombreux domaines en revanche le cas monophonique, et mono capteur en général, est particulièrement compliqué puisqu'il requiert des connaissances a priori sur les sources dont la pertinence est fondamentale pour obtenir une bonne séparation.

Une mise en œuvre particulièrement utile et largement utilisée de l'ACI est l'algorithme FastICA [49] jugé très rapide et robuste. Il est adapté aux grands ensembles de données, et même en présence de données bruyantes.

FastICA a un plus grand avantage sur ACP dans de nombreux aspects qui peut être aussi efficace en l'appliquant avec les forêts rotationnelles afin de les optimiser avec un meilleur paramétrage ; c'est ce qu'on va tenter de démontrer avec le jeu de données de Parkinson pour une meilleure reconnaissance.

6 Conclusion

Plusieurs initiatives de recherche sur la maladie de Parkinson sont mises en œuvre actuellement, en particulier persiste une voie qui s'intéresse au diagnostic par la voix ; elle permet d'effectuer un dépistage préalable de la maladie qui reste souvent cruciale afin de mettre en place des stratégies thérapeutique avant la destruction massive des neurones dopaminergiques.

Les techniques d'apprentissage automatique et surtout les méthodes de classification sont constamment favorisées dans ce genre d'étude car elles ont montré à maintes fois leurs efficacité dans cette tâche dont la fiabilité est vérifiée statistiquement d'où l'exactitude de leurs résultats.

Chapitre 3

Principe des forêts rotationnelles

1 Introduction

Les méthodes de classification ont pour but d'identifier les classes auxquelles appartiennent des objets à partir de certains traits descriptifs. Elles 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, ce dernier consiste à 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 tâche générée tentera à identifier les classes aux quelles appartiennent des objets à partir de certains traits descriptifs et avoir une meilleure prédiction pour classifier les nouveaux exemples.

Les méthodes d'ensemble constituent l'une des principales orientations actuelles de la recherche sur l'apprentissage machine, elles ont été appliquées à un large éventail de problèmes réels. Malgré l'absence d'une théorie unifiée sur des ensembles, il y a beaucoup de raisons théoriques pour combiner plusieurs apprenants, et une preuve empirique de l'efficacité de cette approche.

Dans ce chapitre, nous nous appliquons à la présentation des méthodes d'ensemble et plus particulièrement l'approche des forêts rotationnelles où l'hypothèse sous-jacente imite la nature humaine telle que la combinaison des opinions produira une décision qui est mieux que chaque opinion individuelle ; ici plusieurs classifieurs (des membres de l'ensemble) sont construits et leurs sorties sont combinées en général par vote ou un système de pondération moyenne pour obtenir le classement final ; cependant pour que cette approche soit efficace, deux critères doivent être respectés : la précision et la diversité.

2 Les méthodes d'ensemble

Le principe des méthodes d'ensemble est de construire une collection de prédicteurs ensuite d'agréger l'ensemble de leurs décisions ; son objectif est d'être en mesure de trouver un ensemble d'hypothèses qui sont différentes dans leurs prises de décision afin qu'elles puissent se compléter mutuellement.

Dans un cadre de classification, l'agrégation revient par exemple à faire un vote majoritaire parmi les classes fournies par les classifieurs c'est à dire au lieu d'essayer d'optimiser un modèle qui contient "une seule hypothèse", les méthodes d'ensemble génèrent plusieurs règles de prédiction et ensuite, mettent en commun leurs différentes réponses tout en explorant grandement l'espace des solutions et qu'en agrégant toutes les prédictions, on récupère un prédicteur qui prend en considération toute cette exploration.

L'efficacité d'un ensemble de classifieurs repose sur la combinaison de classifieurs de nature complémentaires ou divers. Chaque classifieur doit être relativement bon et différent des autres classifieurs, autrement dit même si les classifieurs individuels commettent des erreurs, il est peu probable qu'ils commettent les mêmes erreurs pour les mêmes entrées. Ici, surgit l'idée que où un prédicteur se trompe, les autres doivent prendre le relais en ne se trompant pas ; tout cela peut conduire a garantir deux points importants [38] :

- La précision : exige que chaque classifieur doit être aussi précis que possible afin de réduire individuellement l'erreur de l'ensemble.
- La diversité : nécessite de minimiser la corrélation entre les classifieurs, car agréger des prédicteurs qui sont quasiment pareils donnera encore un prédicteur semblable et n'améliorera pas les prédictions.

En revanche il faudra noter que l'agrégation de prédicteurs mauvais ne pourra vraisemblablement pas donner un bon prédicteur final.

La plupart des méthodes d'ensemble peuvent être utilisées avec n'importe quelle méthode de classification, mais les arbres de décision sont les plus couramment utilisés.

3 L'approche Bagging

Le Bagging ("Bootstrap AGGregatING") est une méthode d'ensemble proposée par Breiman en 1996 [40] qui consiste a construire un ensemble de classifieurs a partir de différents échantillons d'un même ensemble de données d'apprentissage (Algorithm 1). Ainsi chaque classifieur élémentaire de l'ensemble sera entrainé sur un des échantillons bootstrap de sorte qu'ils soient tous entrainés sur un ensemble d'apprentissage différent. L'agrégation de ces classifieurs combinés est réalisée par un vote majoritaire ou une méthode de fusion permet d'obtenir un prédicteur plus performant.

Algorithm 1 Pseudo code de l'algorithme Bagging

Entrées :
S : l'ensemble d'apprentissage($N * n$)
T : le nombre d'itérations
Pour i=1→T **faire**
$S_T \leftarrow \emptyset$
Pour i=1→N **faire**
Tirer un exemple au hasard avec remise dans S et l'ajouter dans S_t.
Fin Pour
$h_t = A(S_t)$, A un algorithme d'apprentissage produisant les hypothèses $h_1....h_t \in$ H
Fin Pour
Sortie : H

Cet algorithme utilise une méthode nommée "bootstrapping" (Efron et Tibshirani [50]) est une méthode d'échantillonnage pour générer différents ensembles de même taille de celui de l'ensemble original à l'aide d'un tirage aléatoire avec remise (Figure3.1).

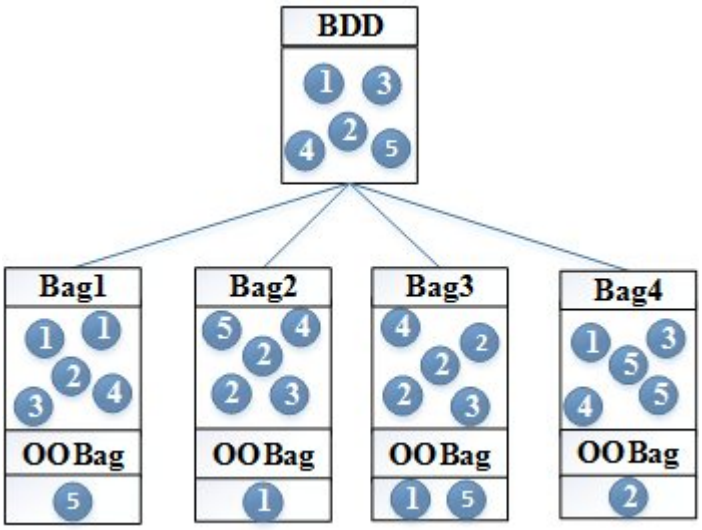

FIGURE 3.1 – Exemple d'échantillons bootstrap d'une banque de données.

Les Out-Of-Bag (OOB) : est l'ensemble des exemples qui ne sont pas sélectionnés dans les échantillons bootstrap. Ce paramètre introduit par la méthode bootstrap permet l'évaluation interne du classifieur.

4 Les forêts aléatoires

Parmi les méthodes issues du Bagging, les forêts aléatoires [34] est l'une des derniers aboutissements de la recherche les plus efficaces pour l'apprentissage d'arbres de décision consacrée a l'agrégation d'arbres randomisés en synthétisant les approches développées respectivement par Breiman 1996 [40], Amit et Geman 1997 [51] , elle génère un jeu d'arbres doublement perturbés au moyen d'une randomisation opérée a la fois au niveau de l'échantillon d'apprentissage et des partitions internes (Algorithm 2).

La méthode des forêts aléatoires consiste à créer une famille d'arbres non élagués (où à chaque nœud la séparation des individus est réalisée à partir d'un sous-ensemble des attributs prédictifs choisi aléatoirement dans l'ensemble des attributs parlant ici du ˜Mtry˝généralement égale à la racine carrée du nombre total d'attributs) à partir d'échantillon bootstrap (échantillon construit par tirage aléatoire avec remise de l'ensemble d'apprentissage).

Les performances d'une forêt d'arbres dépendent de la qualité des individus la composant et leur indépendance ; aussi les forêts aléatoires sont fondées sur des arbres non élagués afin de réduire l'erreur de biais ; en outre, le processus aléatoire permet d'assurer une faible corrélation entre les arbres afin d'assurer leur diversité.

Pour prédire l'étiquette d'un nouvel individu, on utilise un vote majoritaire des arbres de la forêt dans le cas de la classification ou la moyenne des prédictions des arbres dans le cas d'une régression.

L'algorithme de forêts aléatoires est reconnu pour sa précision ; en plus d'être rapide et robuste face aux données bruitées.

Algorithm 2 Pseudo code de l'algorithme forêts aléatoires

Entrées :
T l'ensemble d'apprentissage,
L le nombre d'arbres dans la forêt

Pour i=1→L **faire**
$T_i \leftarrow$ ensemble bootstrap, dont les données sont tirées aléatoirement(avec remise)de T
$C_i \leftarrow$ Construire l'arbre(T_i) où à chaque noeud :
Sélection aléatoire de $K = \sqrt{(M)}$Variables à partir de l'ensemble d'attributs M.
Sélection de la variable la plus informative K en utilisant l'index de Gini.
Création d'un nœud fils en utilisant cette variable.
$E \leftarrow$E UC_i
Fin Pour
Retourner$E$
Sortie : Ensemble des arbres E qui composent la forêt

5 Les forêts rotationnelles

Les forêts rotationnelles développées par Rodriguez et al. [52] basées sur la sensibilité des arbres de décision à des rotations de l'axe ; Celle-ci est une méthode pour générer des ensembles de classifieurs basés sur l'extraction de caractéristiques.

Pour créer les données d'apprentissage d'un classifieur, l'ensemble des caractéristiques est divisé aléatoirement en K sous-ensembles (K est un paramètre de l'algorithme) et l'Analyse en Composantes Principales (ACP) est appliquée à chaque sous-ensemble (Algorithme 3).

Toutes les composantes principales sont retenues dans le but de préserver l'information de variabilité des données. Néanmoins, l'ACP n'est pas utilisée pour réduire la dimension, mais seulement pour faire tourner l'ensemble de données pour but d'encourager simultanément la précision individuelle et la diversité dans l'ensemble.

Algorithm 3 Pseudo code de l'algorithme forêts rotationnelles

Entrées :

X : l'ensemble de données d'apprentissage $(N * n)$.

Y : la classe de la banque d'apprentissage$(N * 1)$.

F : l'ensemble de caractéristiques de la banque.

L : le nombre de classifieurs.

K : le nombre de sous ensemble.

Pour i=1→L **faire**

Préparation de la matrice de Rotation ${R_i}^a$.

Diviser F en K sous ensembles : $F_{i,j}$ $(j = 1.....K)$.

Pour j=1→ K **faire**.

- Soit $X_{i,j}$ l'ensemble de données associer à chaque sous ensemble.
- Suppression aléatoire d'un sous ensemble de classe.
- Bootstrap des sous ensemble sélectionnés (75% de $X_{i,j}$) noté $X^{'}_{i,j}$.
- Application de l'ACP sur $X^{'}_{i,j}$ pour obtention des coefficients $C_{i,j}$.
- Construction des ${R_i}^a$ en arrangeant les colonnes de R_i

Fin Pour

Construire les classifieurs D_i en utilisant $(X{R_i}^a, Y)$.

Fin Pour

Sortie : D

Les points suivants démontrent la rotation des arbres de décision nécessaires pour construire une forêt rotationnelle :

1. Diviser F de façon aléatoire en K sous-ensembles (K est un paramètre de l'algorithme), afin de maximiser les chances d'avoir une grande diversité. Il est recommandé de choisir des sous-ensembles disjoints. Par souci de simplicité, nous supposons que K est un facteur de n, de sorte que chaque sous-ensemble contient M caractéristiques telles que $M = n/K$.
2. Notons par $F_{i;j}$ le sous-ensemble des caractéristiques j pour l'ensemble de classifieurs D_i. Pour chaque tel sous-ensemble, choisir au hasard un sous-ensemble non vide de classes, puis prélever un échantillon bootstrap d'objets $X^{'}_{i,j}$, de la taille 75 % de la banque d'apprentissage, et une analyse en composante principale (ACP) est appliquée à chaque sous-ensemble $X^{'}_{i,j}$. Tous les composantes principales sont retenues dans le but de préserver l'information de variabilité des données.
3. Organiser les vecteurs obtenus à coefficients des composantes principales sous forme d'une matrice de rotation R_i. Les colonnes de R_i doivent être réarrangées en fonction de la séquence d'originalité, et la matrice de rotation réarrangée est indiquée par R^a_i.
4. Apprentissage des classifieurs D_i sur les données X*R^a_i, enfin les sorties de tous les arbres sont fusionnées par la règle de la moyenne (vote majoritaire).

5.1 L'Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) fait partie du groupe des méthodes descriptives multidimensionnelles appelées méthodes factorielles, c'est une technique qui permet de trouver des axes décorrélés et orthogonaux en espaces de dimensions plus petits dans lesquels il est possible d'observer au mieux les individus.

Sa démarche essentielle consiste à transformer les variables quantitatives initiales, plus ou moins corrélées entre elles, en des variables quantitatives, non corrélées, combinaisons linéaires des variables initiales et appelées composantes principales. Les composantes principales sont donc de nouvelles variables.

Globalement le but de l'ACP réside à fournir des représentations synthétiques de vastes ensembles de données numériques, essentiellement sous forme de visualisations graphiques planes en cherchant donc des espaces de dimensions réduites qui ajustent au mieux les espaces initiaux des individus et des variables étant de trop grandes dimensions ,tout en limitant au maximum la perte d'information.

L'algorithme ACP en quelques points

Supposons que nous sommes en présence d'un ensemble de données $X = (x_1 ; x_2 ; \ldots ; x_M)$ composé de M observations où chaque observation $x_i = (x_{i1} ; x_{i2} ; \ldots ; x_{iN})$ est composée de N caractéristiques.

X est associé a une matrice de données A de taille $N * M$ où chaque colonne représente une caractéristique [53]. En pratique, le calcul de l'ACP pour la matrice X revient à réaliser les opérations ci-dessous afin de trouver les composantes principales :

1. Calculer le vecteur $\mu = (\mu_1; \mu_2; ...; \mu_m)^T$ qui représente le vecteur moyen où μ est la moyenne de la $i$ème composante des données.
2. Calculer la matrice χ en soustrayant le vecteur moyen a toutes les colonnes de A dans le but d'obtenir des données centrées.
3. Calculer la matrice S (de taille $N * N$) de covariance de χ avec ($S = \chi . \chi^T$) .
4. Calculer la matrice U (de taille $N * N$) qui est composée des coordonnées des vecteurs propres $\overrightarrow{v_j}$ de S triées par ordre décroissant des modules des valeurs propres λ_j (la première colonne de U est le vecteur propre qui correspond à la plus grande valeur propre).
5. Garder les R premières colonnes de U pour former la matrice $\tilde{U} : N * R$ qui représente les R premières composantes principales.

5.2 L'Analyse en Composantes Indépendantes

Depuis le début des années 1980, de plus en plus de chercheurs se sont penchés sur le problème dit de séparation aveugle de sources, notamment le célèbre problème de la soirée cocktail (cockail party problem) [54], où il s'agit d'extraire les conversations individuelles des convives à partir d'enregistrements sonores contenant un mélange de ces conversations.

Plusieurs méthodes permettant de résoudre ce problème ont fait surface telles que l'analyse en composantes indépendantes (ACI) : une technique statistique dont l'objectif est de décomposer des données aléatoires multi-variables X (les données mesurées) en une combinaison linéaire de composantes mutuellement indépendantes (les données sources), pour faire ressortir des composantes aussi indépendantes que possible à partir des données mesurées autrement dit rechercher les axes (décorrélés mais pas forcément orthogonaux) qui représentent le mieux les données.

Nous pouvons dire que l'ACI est une extension de l'analyse en composantes principales et de l'analyse de facteurs dans le sens ou elle trouve des composantes qui sont mutuellement indépendantes au lieu d'être seulement non-corrélées.

L'algorithme ACI en quelques points

Le modèle de l'ACI s'écrit sous la forme suivante [54] :

$$X = AS$$

où :

- X =$(X_1, ..., X_p)^T$ est un vecteur aléatoire observable $p1$ à valeurs continues des p mélanges de sources,
- A=(a_{lk}) est une matrice de mélange inconnue non aléatoire de dimension pp,
- S = $(S_1, ..., S_p)^T$ est un vecteur aléatoire non observable $p1$, à valeurs continues, des p sources que l'on souhaite retrouver.

A première vue, le problème semble impossible a résoudre puisque A et S sont tous les deux inconnus. Afin de pouvoir identifier les sources S_j, l'ACI repose sur trois hypothèses fondamentales :

1. Les sources S_j sont supposées statistiquement indépendantes. C'est grâce à cette hypothèse que l'on obtient les estimations des sources, appelées composantes « Indépendantes » et notées Y_j.
2. Parmi les p sources, une au plus peut avoir une distribution gaussienne. Intuitivement, la distribution gaussienne est trop simple, car ses cumulants d'ordre supérieur à deux sont nuls.
3. La matrice de mélange A est supposée carrée et inversible, W est notée son inverse. Les sources en fonction peuvent être exprimées par des mélanges de sources selon le modèle d'ACI inverse tel que $S = WX$.

Nous remarquons que les sources sont une combinaison linéaire des mélanges de sources. Il suffit de trouver la bonne matrice de séparation W qui va nous permettre de retrouver les sources.

Méthode Fastica L'algorithme FastICA [49] est un algorithme très performant de maximisation de la fonction de contraste pour les sources non gaussiennes (Hyvärinen, 1999).

FastICA utilise un schéma d'optimisation de point fixe sur la base de Newton-itération et une fonction d'objective liée à la néguentropie ; l'algorithme FastICA peut rechercher les composants indépendants un à la fois ou tous à la fois.

Plusieurs propriétés intéressantes sont associées à FastICA tel que, La convergence de l'algorithme est cubique, donc très rapide comparé aux algorithmes ACI standard, aucune grandeur de pas d'adaptation à choisir, donc facile à utiliser contrairement aux algorithmes basés sur la descente du gradient où ce paramètre doit toujours être ajusté, de même grâce

au choix de la non-linéarité gaussienne, l'algorithme peut être optimisé aux besoins de l'utilisateur.

6 Conclusion

Le recours aux méthodes d'ensemble est une approche qui consiste à produire un modèle de classification plus précis en construisant plusieurs classifieurs différents pour résoudre le problème initial. L'idée principale est de générer, à partir d'un unique algorithme d'apprentissage, un ensemble de classifieurs capables de donner des prédictions différentes sur un même ensemble de données à classer.

La clé de la réussite des méthodes d'ensemble est de savoir si les classifieurs dans un système sont suffisamment diverses , ou en d'autres termes, que les classifieurs individuels ont un minimum d'erreurs en commun. Si un classifieur fait une erreur alors les autres ne doivent pas être susceptibles de faire la même erreur.

La technique des forêts rotationnelles permet la génération d'un ensemble d'arbres de décision indépendamment avec succès. Dans ce cadre, une méthode de transformation linéaire est nécessaire pour projeter les données dans un nouvel espace de fonction pour chaque classifieur pour but de faire tourner les axes, puis les classifieurs de base sont formés dans différents nouveaux espaces afin d'améliorer simultanément l'exactitude individuelle et la diversité au sein de l'ensemble où la diversité est obtenue par l'application d'extraction de caractéristiques et la précision est remportée en gardant toutes les composantes.

Chapitre 4

Expérimentations et Résultats

1 Introduction

Dans le chapitre 2, nous avons parlé les différentes études et travaux qui mettent le point sur l'algorithme des forêts rotationnelles ainsi que notre contribution par rapport à la maladie de Parkinson. Pour cela nous allons présenter dans un premier temps la banque de données élaborée dans ce mémoire par la suite nous allons exposer nos expérimentations réalisées tous en discutant les résultats obtenus avec leurs interprétations.

2 Banque de données

Le jeu de données collectées par Max Little, de l'Université d'Oxford, en collaboration avec le centre national pour la voix et de la parole est appliqué dans ce projet de fin d'études.Elles présentent des enregistrements de signaux de parole. L'étude originale a publié les méthodes d'extraction de caractéristiques pour les troubles de la voix en général.

Cette banque est composée d'une série de mesures vocales biomédicales de 31 personnes, 23 atteintes de la maladie de Parkinson [55].

Les données sont au format CSV ASCII [1] ; chaque colonne de la table CSV est une mesure de voix particulière, et chaque ligne correspond à un enregistrement vocal de 195 de ces personnes (colonne «name»), il y a environ six enregistrements par patient.

L'objectif principal de ces données est de discriminer les personnes en bonne santé de ceux qui sont malades, selon colonne "statuts" qui sont mis à 1 pour la santé et 2 pour parkinsoniens.

1. Comma-separated values,American Standard Code for Information Interchange, est un format informatique ouvert représentant des données tabulaires sous forme de valeurs séparées par des virgules avec codage américain normalisé pour l'échange d'information.

3 Description des paramètres

Le tableau 4.1 décrit les différents variables de la banque de données utilisée :

Paramètres	**Définition**
MDVP : F0(Hz)	Moyenne vocal fréquence fondamentale
MDVP : FHI(Hz)	Maximum vocal fréquence fondamentale
MDVP : Fl0(Hz)	Minimum vocal fréquence fondamentale
MDVP : Jitter(%), MDVP :Jitter(Abs), RAP, MDVP :PPQ, Jitter :DDP	Plusieurs mesures de variation de fréquence fondamentale
MDVP :Shimmer, MDVP :Shimmer(dB) ,Shimmer :app3, Shimmer :APQ5, MDVP :APQ, Shimmer :DDA	Plusieurs mesures de variation d'amplitude
NHR, HNR	Deux mesures de rapport de bruit à composantes tonales de la voix
Statut	État de santé du sujet (deux) de parkinson,(un) de sain
RPDE, D2	Deux dynamiques non linéaires de mesures de complexité
DFA	Signal d'échelle fractale exposant
Spread1, Spread2, PPE	Trois mesures non linéaires de variation de fréquence fondamentale

TABLE 4.1 – Définition des paramètres.

4 Analyse des données de la banque

Attributs	Min/max	Moyenne	Ecart type
MDVP :F0(Hz)	88.3330/260.105	154.2268	41.3901
MDVP :FHI(Hz)	102.145/592.03	197.1049	91.4915
MDVP :Fl0(Hz)	65.476/239.17	116.3246	43.5214
MDVP :Jitter(%)	0.0017/0.0332	0.0062	0.0048
MDVP :Jitter(Abs)	$7.10^{-6}/2.6.10^{-4}$	$4.3959.10^{-5}$	$3.4822.10^{-5}$
MDVP :RAP,	$6.8.10^{-4}/0.0214$	0.0033	0.0030
MDVP :PPQ	$9.2.10^{-4}/0.0196$	0.0034	0.0028
Jitter :DDP	0.002/0.0643	0.0099	0.0089
MDVP :Shimmer	0.0095/0.1191	0.0297	0.0189
MDVP :Shimmer(dB)	0.0850/1.302	0.2823	0.1949
Shimmer :apq3	0.0046/0.565	0.0157	0.0102
Shimmer :APQ5	0.0057/0.0749	0.0179	0.0120
MDVP :APQ	0.0072/0.1378	0.0241	0.0169
Shimmer :DDA	0.0136/0.1694	0.0470	0.0305
NHR	$6.5.10^{-4}/0.3148$	0.0248	0.0404
HNR	8.4410/33.047	21.886	4.4258
RPDE	0.2566/0.6852	0.4985	0.1039
D2	0.5743/0.8253	0.7181	0.0553
DFA	-7.965/-2.434	-5.6844	1.0902
Spread 1	0.0063/0.4505	0.2265	0.0834
Spread 2	1.4233/3.6712	2.3818	0.3828
PPE	0.445/0.5274	0.2066	0.0901

TABLE 4.2 – Informations sur les descripteurs de la banque.

Avant d'aborder les expérimentations et les tests, nous nous sommes attachés à traiter et à structurer les données pour les rendre plus parlantes, plus explicites à travers l'usage de représentation graphique dite histogramme qui représente dans ce cas une répartition des différentes valeurs de chaque variables de la banque à des intervalles selon l'effectif (fréquence) des enregistrements et en projection avec l'aspect médicale de ces derniers.

Par conséquent dans le contexte médical, la dysarthrie parkinsonienne se caractérise par un débit de parole lent avec des accélérations brutales, une diminution de l'intensité vocale, une tendance vers une perturbation de la fréqence fondamentale moyenne, une monotonie de la voix et de son intensité (écart-type de F0 en phonation, jitter et shimmer) et une diminution de la qualité vocale (NHR) de même la norme de la fréquence vocale lors d'un /a/ tenu est de 120-154 Hz chez les hommes et de 181-212 Hz chez les femmes [56].

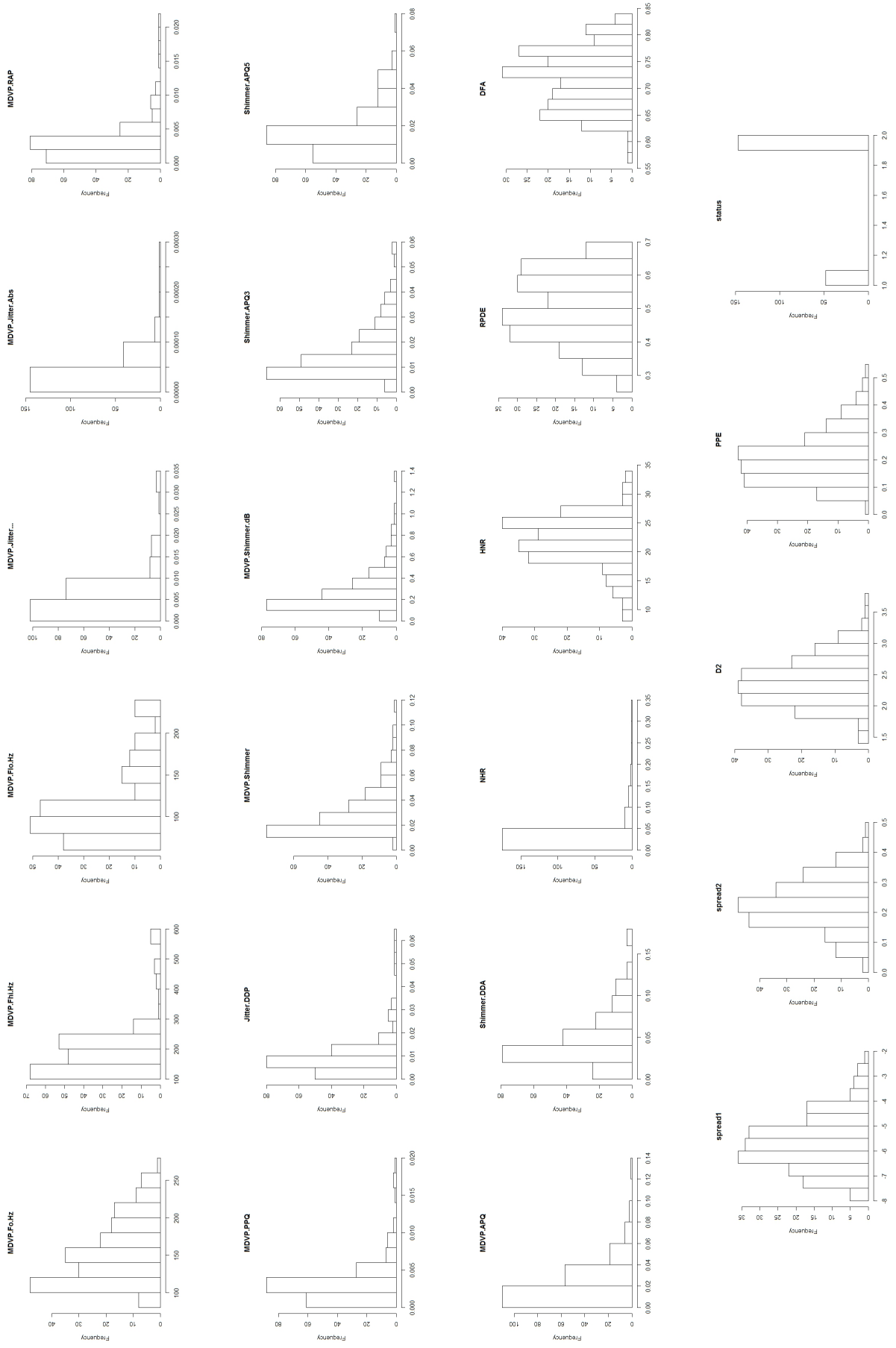

FIGURE 4.1 – Histogrammes des paramètres de la banque de données.

De ce fait, l'interprétation des histogrammes des paramètres caractérisants la maladie de Parkinson à travers la voix (Figure4.1) est résumée comme suit :

- L'histogramme de la qualité vocale (NHR) a bien montré la diminution puisque la majorité des valeurs sont dans l'intervalle [0.00 ; 0.15] qui est inférieure à la normale 0.19.
- Nous constatons que pour les données de shimmer et jitter chez les parkinsoniens, il y'a une tendance d'une monotonie ; où on peut lire d'après les différents histogrammes qu'il y'a une diminution.
- Ainsi l'histogramme du "statut" montre clairement qu'on a plus d'individus malades que d'individus sains (23 malades et 8 sains).
- Seule la donnée de la fréquence fondamentale et ces variations peut fournir une idée préalable sur l'état du sujet suivant les intervalles de la norme.

5 Protocole d'expérimentation

L'objectif principal de ce travail est de dépister la maladie de Parkinson en effectuant une classification supervisée où une étape d'échantillonnage est primordiale. De ce fait, nous avons effectué le protocole de répartition de la banque en deux parties 2/3 pour une phase d'apprentissage et 1/3 pour une phase de test.

Initialement, il nous semble qu'une application d'une classification par les forêts aléatoires donne plus de sens à notre étude et surtout plus d'avantages aux méthodes d'ensemble par rapport aux techniques classiques en confirmant que l'utilisation des ensembles d'arbres mène généralement à une amélioration significative de la prévision. Autrement dit ; une meilleure tendance à prévoir les classes des nouvelles données où la réponse de chaque arbre dépend du sous-ensemble de prédicteurs choisis indépendamment et avec la même distribution pour tous les arbres de la forêt.

Le principe du Bagging : Boostrap et Agrégation a été appliqué pour construire des sous banques qui sont égales aux nombres des arbres, sachant que nous utilisons des arbres de type CART selon le modèle forêts aléatoires en utilisant une séléction aléatoire d'entrée [34]. Ces derniers utilisent un processus récursif de division de l'espace des données en sous régions de plus en plus pures en termes de classes, ainsi leurs lisibilité, leurs rapidité d'exécution et le peu d'hypothèses nécessaires a priori expliquent leurs popularité.

Les données OOB (Out Of Bag) sont utilisées pour l'étape de validation, afin d'estimer les attributs importants et l'erreur de la forêt aléatoire une fois l'arbre courant construit.

Dans la phase de test, l'erreur est estimée via un vote majoritaire dont le principe est de prendre les résultats de deux classifieurs ou plus pour chaque test évalué, à la fin retenir la réponse qui emporte la majorité.

5.1 Choix des paramètres d'algorithmes

Par rapport au choix des paramètres de l'algorithme on peut dire qu'il existe deux principaux paramètres :

– Le paramètre le plus important est le nombre de variables choisi aléatoirement à chacun des nœuds des arbres. Il est nommé *mtry* dans ce contexte ; suivant le modèle forêts aléatoires en utilisant une séléction aléatoire d'entrée [34] la valeur de *mtry* est égale a la racine carrée du nombre de toutes les variables de la banque.

– Le deuxième paramètre est le nombre d'arbres de la forêt. Il est un choix judicieux fait par l'utilisateur. En effet, il sera toujours nécessaire de le justifier après plusieurs expérimentations.

Dans un même volet, nous avons procédé une classification du même jeu de données par les forêts rotationnelles. L'état de l'art a révélé que ces derniers sont plus robustes en les comparant avec les forêts aléatoires dans la plupart des études, puisque le succès des forêts rotationnelles réside dans l'application de la matrice de rotation construite par transformation des sous-ensembles linéaires, nous avons remplacé l'ACP par une méthode prometteuse l'ACI particulièrement Fastica, qui a un plus grand avantage sur ACP dans de nombreux aspects.

– L'ACI fournit un meilleur modèle probabiliste de données, qui peuvent mieux identifier où les données se concentrent dans un espace à n dimensions.
– Elle peut estimer une banque de données non nécessairement orthogonale, qui peut reconstruire mieux les données de l'ACP en présence du bruit.
– L'approche est sensible à des statistiques d'ordre élevé dans les données, et pas seulement la matrice de covariance.

Par rapport au paramétrage de l'algorithme il existe deux principaux paramètres :

– Le premier paramètre est le nombre de sous ensembles : un choix effectué par l'utilisateur, en revanche il sera important de le fixer en tenant compte du nombre de variables de la banque utilisée dans l'étude.

– Le deuxième paramètre est le nombre d'arbres ; un choix approprié par l'utilisateur après différents tests.

En ce qui concerne le choix du nombre d'arbres des deux algorithmes, nous proposons de faire des tests avec un nombre d'arbres variant de 5 à 100 en utilisant 5 validations croisées pour l'ensemble d'apprentissage.

5.2 Choix des paramètres d'évaluations

Les performances des classifications implémentées ont été évaluées par le calcul des critères classiques : pourcentage de sensibilité (SE), la spécificité (SP) et taux de classification (TC), ces derniers sont respectivement définis comme suit :

– **Sensibilité (SE%) :** [SE = 100 * VP / (VP + FN)] la sensibilité (Se) du test est sa capacité de donner un résultat positif quand la maladie est présente. Représente ceux qui sont correctement détectés parmi tous les événements réels.
– **Spécificité (SP %) :** [SP = 100* VN / (VN + FP)] la spécificité du test est cette capacité de donner un résultat négatif quand la maladie est absente. Elle est représentée pour détecter les patients non parkinsoniens.
– **Taux de classification (TC%) :** [TC = 100 * (VP + VN) / (VN + VP + FN +FP)] est le taux de reconnaissance.

Tel que :
- VP : parkinsonien classé parkinsonien ;
- FP : non parkinsonien classé parkinsonien ;
- VN : non parkinsonien classé non parkinsonien ;
- FN : parkinsonien classé non parkinsonien.

6 Résultats et Discussion

Les résultats de nos expérimentations se résument dans la figure 4.2, nous remarquons clairement qu'à partir du nombre d'arbre égal à 50, les résultats des forêts aléatoires s'améliorent ce qui est conforme à son principe énoncé par Breiman [34] où l'une des conditions d'atteindre une bonne classification est un nombre d'arbre élevé.

Par contre nous distinguons très clairement que les forêts rotationnelles ont atteint une valeur maximal de reconnaissance égale à 95.38% en appliquant FastICA comme méthode de transformation et à 93.53% avec ACP en n'utilisant que 20 arbres. Cela s'explique par le fait que dans le cas des forêts rotationnelles l'ensemble de données transformé a autant d'exemples que l'ensemble de données d'origine et la diversité est obtenue par l'application d'extraction de caractéristiques.

En comparant les résultats obtenus dans le tableau 4.3, nous pouvons constater plusieurs points tel que :

– Un des points remarquables de cette étude est le mérite des forêts rotationnelles en créant l'ensemble de classifieurs de base avec l'amélioration possible des performances de classification par rapport à la classification effectuée par les forêts aléatoires.

– L'explication heuristique de ces améliorations revient à la rotation des arbres, ce qui rend ces derniers encore plus différents les uns des autres. Par conséquent, une question naturelle est de savoir s'il est possible d'obtenir divers classifieurs sans écarter aucune information dans l'ensemble de données.

– La sensibilité des arbres de décision à l'axe de rotation est généralement considérée comme un inconvénient, mais là elle peut être très bénéfique car les arbres de décision obtenus à partir d'un ensemble de données en rotation peuvent encore être précis, car ils utilisent toutes les informations disponibles dans le jeu de données, tout en étant très diverses.

– De même les forêts rotationnelles avec FastICA ont fourni de meilleurs résultats en les comparant avec les forêts rotationnelles avec ACP ; ce qui valide notre contribution de départ par le fait que ACI donne plus d'avantage que ACP dans plusieurs aspects.

Classifieurs	Arbres	TC%	TE%	SE%	SP%
Forêt aléatoire	5	84.61±15.38	15.38±15.38	87.38 ±7.88	73.79±95.76
	15	86.15±5.91	13.84±5.91	86.24±2.14	86.33±102.09
	20	93.23±8.99	6.76±8.99	96.67±1.10	84.82±96.38
	25	90.76±1.18	9.23±1.18	93.22±2.47	86.49±30.27
	30	86.76±1.89	13.23±1.89	86.60±0.53	88.77±79.76
	35	86.15±1.18	13.84±1.18	86.23±0.71	85.91±21.35
	40	85.53±6.62	14.46±6.62	88.37±1.65	75.32±73.97
	45	86.15±1.18	13.84±1.18	85.97±0.53	87.33±17.03
	50	87.07±1.89	12.92±1.89	86.12±0.60	93.33±37
	60	86.76±1.89	13.23±1.89	86.59±0.86	87.77±19.13
	80	86.76±0.71	13.23±0.71	86.33±0.72	89.33±0.37
	100	87.38±1.65	12.61±1.65	86.43±0.77	93.55±34.81
Moyenne		87.27±4.03	12.73±4.03	88.01±1.66	85.80±50.65
Rotate-ACP	5	85.84±5.20	14.15±5.20	87.86±3.65	77.53±31.47
	15	87.07±0.71	12.92±0.71	86.91±0.46	88.14±12.72
	20	93.53±9.94	6.46±9.94	97.04±1.41	84.44±83.33
	25	91.69±3.07	8.30±3.07	93.95 ±1.92	84.41±12.29
	30	86.45±0.47	13.53±0.47	85.76±0.01	91.11±24.69
	35	86.15±1.18	13.84 ±1.18	85.97±0.53	87.33±17.03
	40	88±6.39	12±6.39	89.92±2.60	80.65±39
	45	85.23±0.71	14.76±0.71	85.55±0.02	83.55±23.70
	50	85.53±0.71	14.46±0.71	85.61±0.02	85.33±23.70
	60	86.15±1.18	13.84±1.18	85.97±0.53	87.33±17.23
	80	85.84±0.47	14.15±0.47	85.92±0.39	85.69±19.51
	100	86.46±1.65	13.53±1.65	86.28±0.82	87.55±18.14
Moyenne		87.32±2.63	12.68±2.63	88.05± 1.02	85.26±21.89
Rotate-ACI	5	88.30±16.09	11.69±16.09	91.94±9.13	77.08±67.14
	15	88.30±24.37	11.69±24.37	89.74±17.27	82.26±94.05
	20	**95.38**±1.18	4.61±1.18	97.91±0.0009	88.35±13.59
	25	92.61±4.02	7.38±4.02	94.42±6.22	86.76±1.22
	30	88±7.57	12±7.57	88.23±8.83	87.30±16.84
	35	86.46±2.48	13.53 ±2.48	86.54±0.99	86±30
	40	89.23±1.18	10.76±1.18	91.02±0.93	82.70±10.18
	45	86.76±6.62	13.23±6.62	87.15±4.36	84.69±32.27
	50	86.76±0.71	13.23±0.71	86.86±0.42	86.50±18.15
	60	86.15±1.18	31.84±1.18	86.76±0.54	83.09±15.53
	80	86.46±0.47	13.53±0.47	86.81±0.39	84.86±17.60
	100	86.46±4.02	13.53±4.02	86.85±4.46	84.70±22.45
Moyenne		**88.40**±5.82	11.6±5.82	89.51±4.45	84.51±28.24

TABLE 4.3 – Performances des différents classifieurs.

FIGURE 4.2 – Taux de classification par rapport au nombre d'arbres.

Comme toutes techniques de classification, les forêts rotationnelles ont des avantages ainsi que des inconvénients. Nous avons constaté que la facilité d'interprétation des arbres est perdue avec les méthodes d'ensemble. En effet, une forêt est l'agrégation de toute une collection d'arbres, donc nous perdons l'aspect structuré du prédicteur obtenu. Pour pallier à ce manque, un autre indice d'importance des variables, spécifique aux forêts est introduit par Breiman. Aussi en terme du temps les méthodes d'ensemble demandent un temps d'exécution énorme mais cela est seulement dans la phase d'apprentissage, en phase de test on remarque qu'il est moins lent.

De nombreuses études recourant la banque de données du Max Little ont été réalisées pour la classification de Parkinson. L'auteur même de la banque [18] a rapporté un taux de classification de 91.4% en appliquant le classifieur individuel SVM, le point fort de cette étude est l'introduction d'une nouvelle méthode de mesure des dysphonies PPE dans sa banque de données. En outre Boublenza et al. [32] ont pu atteindre une performance de 96.88% à l'aide du même classifieur SVM, une méthode de sélection de variables par Relief-F pour réduire le nombre de caractéristiques était mise en œuvre, cela a donné plus d'avantage à l'étude.

Une des principales critiques que l'on puisse faire sur SVM est le manque d'intelligibilité des résultats. En effet, il s'agit d'une technique "boite noire" qui ne fournit pas d'explications ni d'indices quant aux raisons d'une classification. Les résultats doivent être pris tels quels en faisant confiance au système qui les a produits. Pourtant les experts du domaine préfèrent largement une méthode d'apprentissage avec explications et recommandation d'actions plutôt qu'une boite noire, aussi performante et prédictive soit-elle.

Dans un autre volet, en comparant nos résultats avec ceux obtenus dans l'étude d'Akin Ozcift et al. [30] qui ont utilisé la même banque de données du Max Littel avec les forêts rotationnelles, ils ont obtenu une valeur moyenne du taux de classification de 84.4% sans sélection et de 87.1% avec sélection, c'est une augmentation moyenne de 2.7% en précision globale. En revanche, dans la présente étude nous avons pu arriver à un taux moyen de 88.40% l'argument que nous jugeons appréciable dans les méthodes d'ensemble est l'ajout du comportement global au comportement individuelle qui offre plus de robustesse à l'approche.

De même par rapport aux taux de sensibilité et spécificité nous avons pu avoir de bonnes valeurs ce qui veut dire une bonne reconnaissance des faux positifs et des faux négatifs.

7 Conclusion

Les forêts rotationnelles sont l'une des méthodes d'ensemble qui a la spécificité de transformer les données avec différents axes tout en préservant l'information complète. Cette méthode effectue généralement des résultats meilleurs que d'autres classifieurs d'ensemble.

Cette étude présente un modèle de classification par les forêts rotationnelles avec FastICA rapportant une amélioration de performances avec un taux de 88.40% par rapport aux forêts rotationnelles classiques et forêts aléatoires.

La clé de réussite de l'approche abordée est la construction de la matrice de rotation afin d'encourager la précision et la diversité dans l'ensemble.

Conclusion générale

La maladie de Parkinson est une affection chronique due à la disparition progressive de certains neurones dans le cerveau. Cela provoque une baisse de la production du dopamine, une substance qui transmet l'information entre neurones, dans une région du cerveau essentielle au contrôle des mouvements.

Les méthodes d'ensemble constituent une famille ou ensemble d'algorithmes qui génèrent une collection de classifieurs, par la suite les combiner en agrégant leurs prédictions. L'efficacité de la combinaison des classifieurs repose principalement sur leur capacité à tirer les complémentarités des classifieurs individuels dans le but d'améliorer autant que possible les performances en généralisation de l'ensemble. Une explication de ce lien entre complémentarité et performance est donnée par la notion de diversité.

Ce travail aborde le problème de la reconnaissance précoce de la maladie de Parkinson en appliquant la méthode d'ensemble "forêt rotationnelle". Elle est reconnue dans la littérature, comme l'une des techniques de génération d'ensemble d'arbre indépendamment la plus performante. Où la projection des données dans différents nouveaux espaces porte une amélioration simultanée sur l'exactitude individuelle et la diversité au sein de l'ensemble, la diversité est obtenue par l'application d'extraction de caractéristiques pour la rotation des axes et la précision est favorisée en gardant tous les composantes.

Notre objectif a été d'améliorer les performances de cette dernière en utilisant une méthode de transformation linéaire FastICA .

Les forêts rotationnelles ont fait ressortir leur efficacité par rapport aux d'autres méthodes d'ensemble dans la classification supervisées ce qui permet d'apporter une valeur ajoutée dans l'implémentation des systèmes d'aide au diagnostic médical.

Nous avons bien conscience que notre travail ne constitue qu'un début des réflexions qui devront être poursuivies en suivant les nombreuses pistes qui nous sont apparues . À court terme pour ce travail, il nous semble que :

- L'évaluation des performances de l'approche proposée en adoptant d'autres algorithmes instables comme un réseau neuronal, réseau bayésien, étant une piste interessante.
- L'application d'un algorithme de sélection pour trouver les caractéristiques les plus pertinentes peut munir une classification plus performante.
- L'utilisation d'un vote pondéré à la place du vote majoritaire dans la phase d'agrégation des classifieurs peut fournir une amélioration en termes de performances.
- L'introduction d'un critère du choix de la fonction de non linéarité pour influencer sur la vitesse de convergence de Fastica. et ainsi gagner en temps de calculs.

Annexe : Implémentation d'un système d'aide au diagnostic pour la détection de Parkinson

Une interface graphique fournit un contrôle d'une application interactivement avec l'utilisateur. Matlab permet d'écrire assez simplement cette dernière pour faire un produit interactive utilisable par des utilisateurs qui ne sont pas forcément formés à Matlab.

Matlab possède l'outil GUIDE, qui permet de créer facilement des interfaces graphiques. Il suffit pour cela de taper dans la ligne de commande de MATLAB guide et on a alors accès à ses différentes fonctions. On peut créer une interface graphique à partir de modèles déjà existants, d'interfaces sauvegardées ou la réaliser depuis le début, ce que nous avons fait.

Dans un premier temps nous avons réalisé une interface principale de l'application ; la figure suivante 4.3 représente une capture d'écran de cette interface .

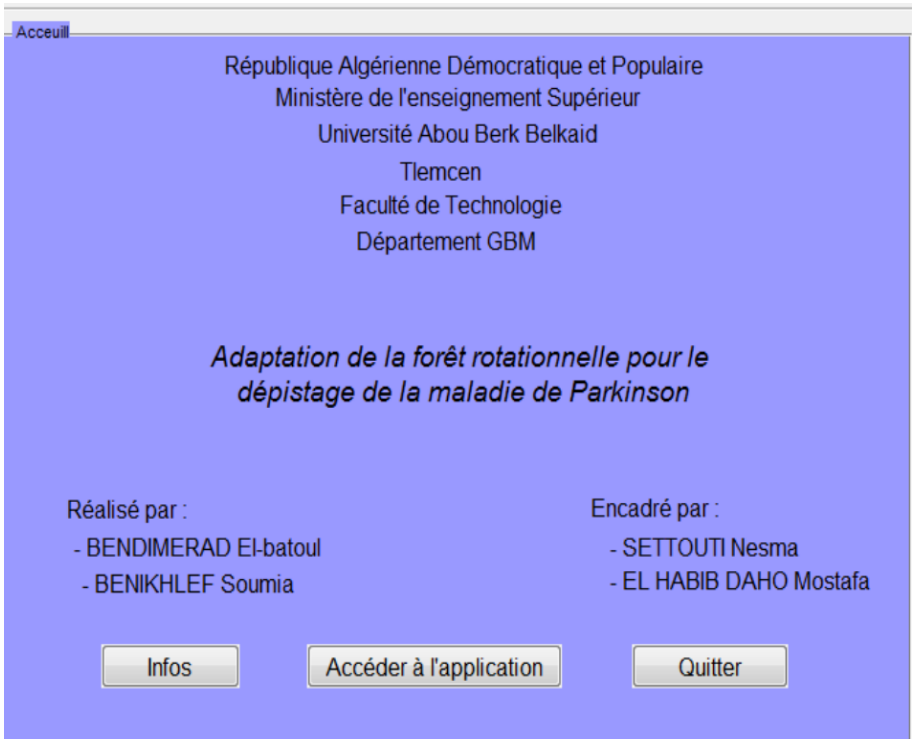

FIGURE 4.3 – Interface principale de l'application.

Cette dernière contient trois boutons :

– Bouton quitter pour sortir complétement.
– Bouton infos contient un aperçu général de cette dernière 4.4.

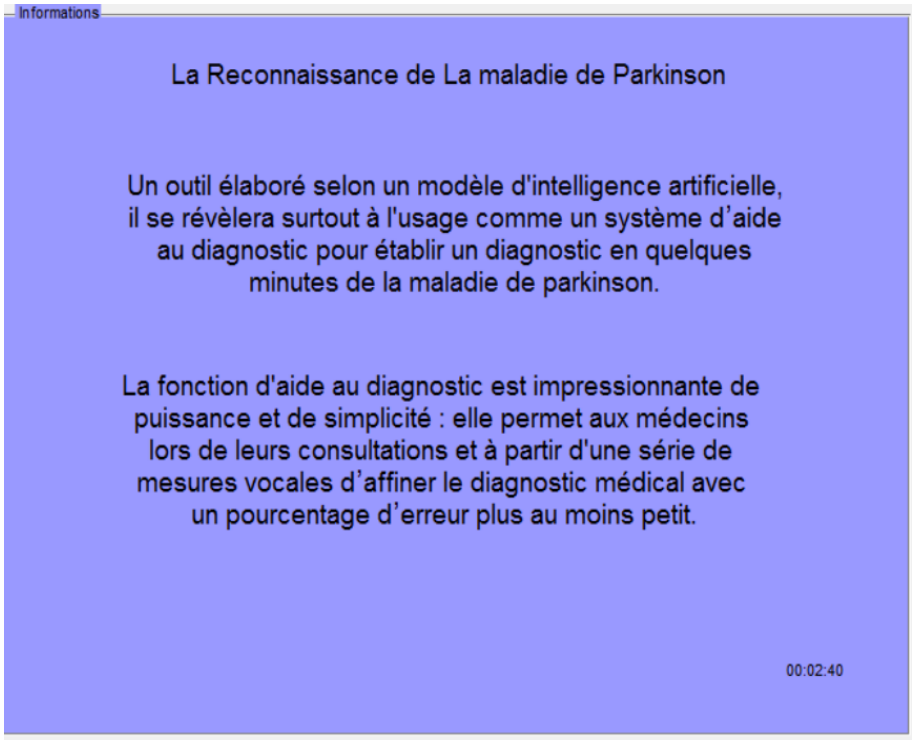

FIGURE 4.4 – Interface informations.

– Et un bouton accéder pour avoir accès à l'application.

Une fois appuyer sur accéder une autre interface de mot de passe fait face a l'utilisateur pour que ce dernier s'authentifier en entrant un mot de passe valide ce que montrera la figure suivante 4.5

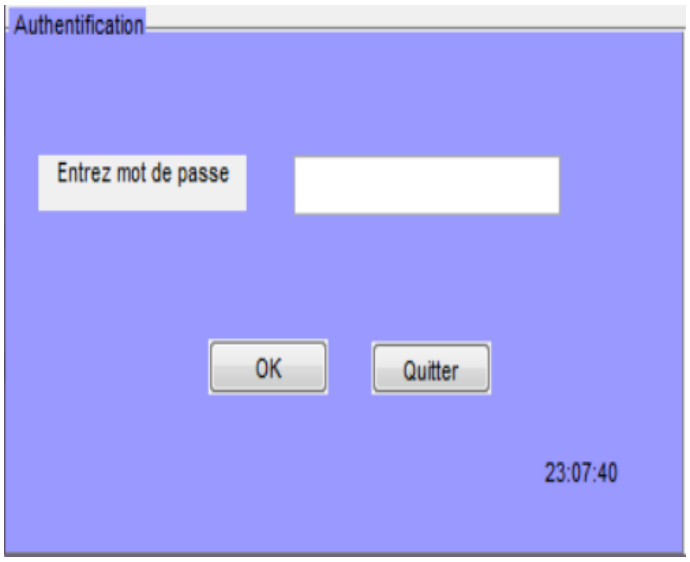

FIGURE 4.5 – Interface de mot de passe.

Lorsque l'utilisateur fait entrer le bon mot de passe, il remarque bien une fenêtre que tous les paramètres de l'application utilisés y figure. 4.6

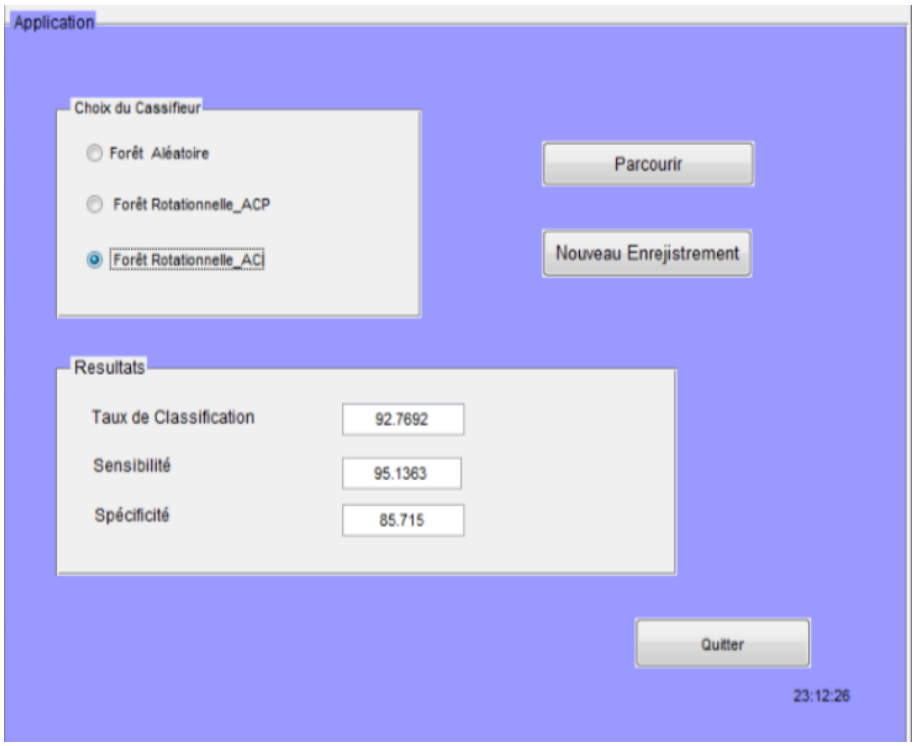

FIGURE 4.6 – Interface de test.

Pour cela il a les possibilités suivantes :
- Un bouton parcourir pour charger la banque de données utilisée dans cette étude.
- Choix du classifieur pour lancer le test où une fois le choix est effectué les différents résultats de la classification s'affichent automatiquement sur leurs champs correspondants.
- Un bouton nouveau enregistrement pour que l'utilisateur effectue une nouvelle instance en tapant les valeurs des différentes variables de la banque 4.7.

FIGURE 4.7 – Interface pour nouveau enregistrement.

Et en appuyant sur le bouton diagnostiquer il pourra prédire l'état du nouveau enregistrement (parkinsonien ou non parkinsonien) suivant les trois classifieurs avec même un degré de précision ; ce que illustre la figure 4.8

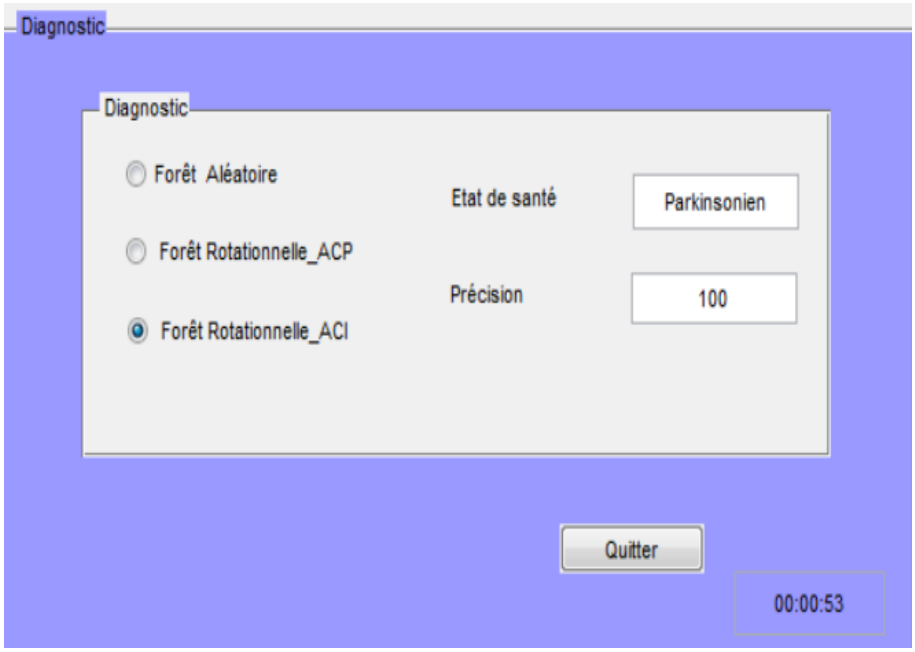

FIGURE 4.8 – Interface pour avoir le diagnostic .

Cette interface graphique permet de fournir un aide au diagnostic pour le dépistage de la maladie de parkinson étant une façon aisée en effectuant quelque clicks pour tirer profit des techniques d'intelligence artificielle plus exactement des méthodes de classification.

Bibliographie

[1] Cécile Chevalier, *Les médicaments dopaminergique de la maldie de Parkinson aux traitements des addictions*, Ph.D. thesis, Université Joseph Fourier, Faculté de pharmacie de Gronoble, May 29,1984.

[2] Philippe Boulu, "Chifres clés et prévalence de maladie de parkinson,(http ://www.carenity.com/maladie/parkinson/chiffres-cles-et-prevalence-de-la-maladie-de-parkinson)," Access 20/02/2015.

[3] Valérie FRAIX, *Motricité,cognition,émotion dans la maladie de parkinson : role des oscillations du noyau subthalamique*, Ph.D. thesis, Université Joseph Fourier-Gernoble I Pole chimi,sciences de la vie et de la santé,Biongénirie, November 28, 2008.

[4] E. R. Dorsey, R. Constantinescu, and J. P. Thompson, "Projected number of people with parkinson disease in the most populous nations, 2005 through 2030," *official journal of the American Academy of Neurology*, vol. 68, pp. 5, January 30, 2007.

[5] L.Defebvre and M.Vérin, "La maladie de parkinson, monographie de neurologie," *Issy-Les-Moulineaux : Elsevier Masson*, pp. 47–64, 2011.

[6] Frey Clémentine and Senepin Chloé, *Création et évaluation d'un logiciel d'entrainement pour les patients parkinsoniens atteints de dysarthrie*, Ph.D. thesis, Université Claude Bernard Lyon1 Institut des sciences et Techniques de réadaptation, June 28,2012.

[7] Antoine Flavien Eger, Christophe Gaudet Blavignac, and Arthur Hammer, *La Maladie de Parkinson*, Ph.D. thesis, Université de Genève, June 26, 2009.

[8] F Viallet and B Teston, "La dysarthrie dans la maladie de parkinson," *Les Dysarthries, P. Auzou (Ed.)*, vol. 13, pp. 169–174, 2007.

[9] S. Pinto, A. Ghio, B. Teston, and F. Viallet, "La dysarthrie au cours de la maladie de parkinson. histoirenaturelle de ses composantes : dysphonie, dysprosodie et dysarthrie," *Elsevier Masson SAS*, vol. 66, pp. 800 – 810, August 26 ,2010.

[10] Max Little, Thanasis Tsanas, and Ladan Baghai-Ravary, "Parkinson's voice initiative,(http ://www.parkinsonsvoice.org/science.php)," (25.07.2012 - 16 h 11),Access 15/03/2013.

[11] Ludivine Olives, "Guérir la maladie de parkinson avec une pilule et un téléphone,(http ://www.slate.fr/lien/59835/parkinson-alzheimer-telephone-pilule)," (26.07.2012 à 10 h 48),Access 15/03/2013.

[12] A. Tsanas, M.A. Little, Patrick E, McSharry, and Lorraine O. Ramig, "Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests," *IEEE Transactions on Biomedical Engineering*, vol. 57(4), pp. 884–893, 2009.

[13] J. Peter Snyder, Michael. Cannizzaro, and Brian.Harel, "Variability in fundamental frequency during speech in prodromal and incipient parkinson's disease : A longitudinal case study," *Brain and Cognition*, vol. 56, pp. 24–29, 2004.

[14] Max Little, Patrick McSharry, Irene Moroza, and Stephen Roberts, "Nonlinear, biophysically-informed speech pathology detection," *ICASSP*, p. 4, 2006.

[15] Max A. Little, *Biomechanically Informed Nonlinear Speech Signal Processing*, Ph.D. thesis, Université d'Oxford, 2006.

[16] Max A Little, Patrick E McSharry, Stephen J Roberts, Declan AE Costello, and Irene M Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomed Engligne*, vol. 6, pp. 30, June 26, 2007.

[17] S.Henry Cheang and D.Marc Pell., "An acoustic investigation of parkinsonian speech in linguistic and emotional contexts," *Journal of Neurolinguistics*, vol. 20, pp. 221–241, 2007.

[18] M.A. Little, Patrick E. Mc Sharry, Eric J. Hunter, Jennifer Spielman, and Lorraine O. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56(4), pp. 1015–1022., November16, 2008.

[19] Jan Rusz, *Acoustic analysis of voice and speech disorders in Parkinson's disease*, Ph.D. thesis, Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Circuit Theory, March 2012.

[20] Patricia Gillivan-Murphy, *Voice tremor in Parkinson's disease (PD) Identification, characterisation and relationship with speech, voice,and disease variables*, Ph.D. thesis, The Institute of Health & Society For the degree of Doctor of Philosophy, January 2013.

[21] Taha Khan, "Running-speech mfcc are better markers of parkinsonian speech deficits than vowel phonation and diadochokinetic," *Mälardalen University, School of Innovation, Design and Engineering,Dalarna University*, p. 16, March 2014.

[22] Athanasios.Tsanas, Max A. Little, Cynthia .Fox, and Lorraine O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in parkinson's disease," *IEEE Engineering in Medicine and Biology Society,*, vol. 22, pp. 10, January 2014.

[23] Athanasios.Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig, "Parkinsons disease symptom severity metric achieve clinically useful quantification of average nonlinear speech analysis algorithms mapped to a standard," *J. R. Soc. Interface published online*, p. 15, November 17, 2010.

[24] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig, "Using the cellular mobile telephone network to remotely monitor parkinson's disease symptom severity," *IEEE Transactions on Biomedical Engineering*, p. 9, 2012.

[25] Max A. Little, Paul. Wicks, Timothy E. Vaughan, and Alex .Sandy Pentland, "Quantifying short term dynamics of parkinson's disease using self-reported symptom data from an internet social network," *Journal of Medical Internet Research*, vol. 15(1), pp. 14, January 24, 2013.

[26] Ricardo Graça, "Parkdetect early diagnosing parkinson's disease", medical measurements and applications (memea)," *IEEE International Symposium*, p. 66, July 19,2013.

[27] Jan ruusz, Roman Cmejla, hnan Ruzickova, and Even Ruzicka, "Objectification of dysarthria in parkinson's disease using bayes theorem," *Recent Researches in Communications, Automation, Signal Processing, Nanotechnology, Astronomy and Nuclear Physics. ISBN : 978-960-474-276-9*, p. 5, 2011.

[28] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's

disease," *2011 Acoustical Society of America.*, vol. 129(1), pp. 350–367, October 7, 2010.

[29] A.Wald, "Sequential tests of statistical hypotheses," *Annals of Mathematical Statistics*, vol. 16, pp. 117–186, 1945.

[30] Akin Ozcift and Arif Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," *computer methods and programs in biomedicine*, vol. 104, pp. 443–451, 2011.

[31] Hananel Hazan, Dan Hilu, Larry Manevitz andLorraine O. Ramig, and Shimon Sapir, "Early diagnosis of parkinson's disease via machine learning on speech data," *IEEE 27-th Convention of Electrical and Electronics Engineers in Israel*, p. 4, November 2012.

[32] A.Boublenza, A.Benosman S.Bouchikhi, and M.A.Chikh, "Parkinson's disease detection with svm classifier and relief-f features selection algorithm," Tech. Rep., Tlemcen University, Algeria, 2012.

[33] S.Arora, MA. Little, V.Venkataraman, S. Donohue, KM. Biglan, and ER. Dorsey, "High-accuracy discrimination of parkinson's disease participants from healthy controls using smartphones," *IEEE Xplore Digital Library*, p. 1, 2013.

[34] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[35] Siddharth Arora, Vinayak Venkataraman, Sean Donohue, Kevin M. Biglan, Earl R. Dorsey, and Max A. Little, "High accuracy discrimination of parkinson's desease participants from healthy controls using smartphone's," *IEEE Xplore Digital Library*, p. 4, 2014.

[36] Max Little, Ph.D.Thanasis Tsanas, and Ladan Baghai-Ravaryand Ph.D M.Sc., "Parkinson's voice initiativ," *http ://www.parkinsonsvoice.org/science.php*, p. 1, August 09,2014.

[37] Soumia Benikhlef, El Batoul Bendimerad, and Settouti Nesma, "Application de l'analyse en composantes indépendantes pour la détection de la maladie de parkinson," Tech. Rep., Université Abou Bekr Belkaid – Tlemcen,Laboratoire Génie Biomédical, 2013.

[38] D.Gopika and B.Azhagusundari, "A novel approach on ensemble classifiers with fast rotation forest algorithm," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, pp. 8, August 2014.

[39] Juan J. Rodrıguez and Ludmila I. Kuncheva, "Rotation forest :a new classifier ensemble method," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, pp. 10, October 2006.

[40] Leo Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[41] Schapire RE Freund Y, "A decision-theoretic generalization of on-line learning and an application to boosting," *J Comput System Sci*, vol. 55(1), pp. 119–139, 1997.

[42] Ludmila I. Kuncheva and Juan J. Rodriguez, "An experimental study on rotation forest ensembles," *Springer-Verlag Berlin Heidelberg*, vol. 4472, pp. 10, 2007.

[43] Kun-Hong Liu and De-Shuang Huang, "Cancer classification using rotation forest," *Computers in Biology and Medicine*, vol. 38, pp. 601 – 610, February 15, 2008.

[44] R. E.Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, pp. 197–227, 1990.

[45] Chun-Xia Zhang and Jiang-She Zhang, "Rotboost : A technique for combining rotation forest and adaboost," *Pattern Recognition Letters*, vol. 29, pp. 1524–1536,, 2008.

[46] Chun-Xia Zhang Æ Jiang-She Zhang, "A variant of rotation forest for constructing ensemble classifiers," *Springer-Verlag London*, p. 19, September 10, 2009.

[47] Sotiris Kotsiantis, "Combining bagging, boosting, rotation forest and random subspace methods," *Springer Science+Business Media B.V*, vol. 35, pp. 223–240, 2010.

[48] Chun-Xia Zhang, Jiang-She Zhang, and Guan-Wei Wang, "An empirical study of using rotation forest to improve regressors," *Applied Mathematics and Computation*, vol. 195, pp. 618–629, 2008.

[49] Aapo Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10(3), pp. 626–634, 1999.

[50] B.Efron and R. J.Tibshirani, *An Introduction to the Bootstrap.*, New Yor, 1993.

[51] Yali Amit and Donald Geman, "Shape quantization and recognition with randomized trees," *Neural Computation*, vol. 9, pp. 1545–1588, 1997.

[52] Juan J. Rodrıguez and Ludmila I. Kuncheva, "Rotation forest :a new classifier ensemble method," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, pp. 1619–1630, October 2006.

[53] Chouaib Hassan, *Sélection de caractéristiques :méthodes et applications*, Ph.D. thesis, Université Paris Descartes, July 8, 2011.

[54] Marc-Olivier Billette, *Analyse en composantes indépendantes avec une matrice de mélange éparse*, Ph.D. thesis, Université de Montréal,Département de mathématiques et de statistique, june 2013.

[55] Littel.M, "Uci machine learning repository parkinsons data set," 2008.

[56] Maude Grueber, "Effets de la stimulation sousthalamique bilatérale sur la voix et la parole de patients parkinsoniens," M.S. thesis, Université de Lausanne, Faculté de biologie et médecine, January 4,2012.

Résumé

Touchant près d'un million de personnes chaque année dans le monde, la maladie de Parkinson a atteint le second rang des maladies dégénératives. Ainsi, le champ de recherche s'est développé énormément pour conduire à un diagnostic médical précis. L'application des techniques d'apprentissage artificielle peut être une piste qui apparaît de plus en plus d'être très prometteuse. Ceci nous a amené à élaborer notre projet qui consiste à traiter la possibilité de l'implémentation d'un système d'aide aux médecins pour la reconnaissance de la maladie dans ces stades précoces.
Dans le cadre de notre projet de fin d'étude nous nous intéresserons à l'amélioration des performances de la classification par les forêts rotationnelles. Cette méthode combine la robustesse des arbres de décision, la puissance de l'extraction des caractéristiques tout en augmentant la précision et la diversité des arbres dans la forêt par l'utilisation de tous les paramètres. Les résultats expérimentaux appliqués sur la banque de données médicales réelles (Parkinson) montrent une efficacité dans la tâche de classification avec exactitude significative vérifiée statistiquement.

Mots clés : La maladie de Parkinson ; Forêt rotationnelle ; Méthode d'ensemble ; Forêt aléatoire ; Analyse en composantes principales ; Analyse en composantes indépendantes

Abstract

Affecting nearly a million people each year in the world, Parkinson's disease has reached the second rank of degenerative diseases. Thus, the research field developed tremendously to lead to a precise medical diagnosis. The application of artificial learning techniques may be a track that appears increasingly to be very promising.This led us to develop our project, which comprises treating the possibility of the implementation of a system of assistance to doctors for the recognition of the disease in the early stages.

As part of our project of graduation, we will focus on improving the performances of classification by the rotational forests. This method combines the robustness of decision trees, the power of the feature extraction while increasing accuracy and tree diversity in the forest by using all parameters. The experimental results applied to the actual medical database (Parkinson) show effectiveness with significant accuracy in the classification verified statistically.

Keywords : Parkinson's disease ; Rotational forest ; Overall method ; Random forest ; Main component analysis ; Independent Component Analysis.