Exploration et classification de données en neurologie

Lucas Randazzo et Naina Razakarison
encadrés par Nicolas Vayatis, Emile Contal, Themistoklis Stefanakis

CMLA, École Normale Supérieure de Cachan

À l'heure actuelle, les neurologues peuvent effectuer un diagnostic efficace à partir d'un test simple : le test de Romberg. Il consiste pour le patient à se tenir debout pendant 5 secondes, une fois les yeux ouverts, une autre les yeux fermés, afin de quantifier son équilibre. Notre projet consiste à élaborer une aide au diagnostic, notamment avec l'acquisition de la trajectoire du centre de pression du patient au cours du test. Cela permet d'afficher le résultat du test pour aider le médecin à visualiser la trajectoire, ainsi que d'enrichir une base de donnée anonyme qui associe une maladie à chaque acquisition. On peut alors effectuer une analyse statistique des données sur un grand ensemble d'acquisitions, dans le but d'être capable de classifier les sujets malades des sujets sains avec des algorithmes d'apprentissage. Cependant, les trajectoires ne peuvent être utilisées en l'état, car les algorithmes d'apprentissage nécessitent une représentation vectorielle des données. Il faut au préalable trouver des descripteurs qui en résument efficacement leur information. Des descripteurs efficaces permettraient une meilleure classification et donnent des information supplémentaires au médecin. Idéalement, on devrait être en mesure de quantifier l'état de la maladie chez un patient est de suivre son évolution au cours du temps, ce qui peut s'avérer utile pour évaluer l'efficacité d'un traitement. Même si le nombre de patients est encore assez faible, la base de données est constamment mise à jour avec de nouvelles acquisitions. Certains descripteurs présentent de bonnes performances, qui semblent déjà dépasser l'état de l'art.

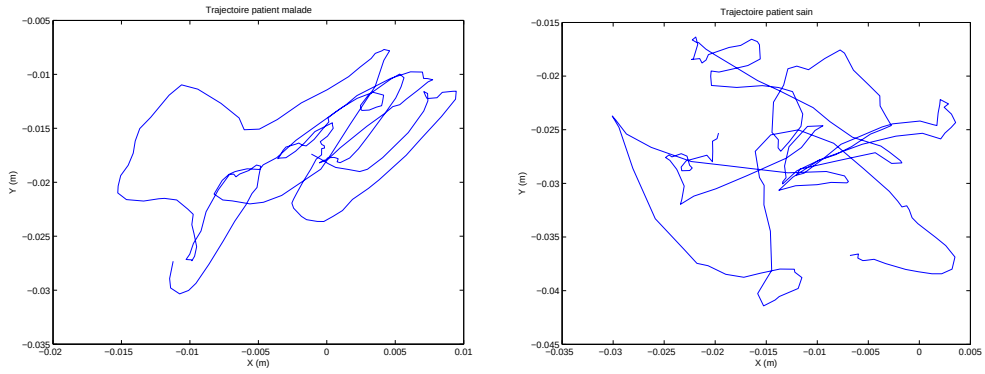
1 Présentation du projet

Le projet consiste à concevoir un logiciel d'aide au diagnostic pour quantifier des maladies neurologiques. A l'aide d'un système d'acquisition facile à trouver dans le commerce, une *wii balance board*, et une interface graphique sur tablette tactile, les neurologues devraient avoir accès une analyse statistique du patient pour le placer parmi un ensemble de données annotées. Les données des acquisitions sont alors envoyées vers un serveur central, situé actuellement au CMLA. Dans un premier temps, seul le Docteur Damien Ricard participait au projet et nous fournissait des informations. L'objectif est qu'un maximum de neurologues participent à enrichir la base de données, afin d'obtenir un nombre conséquent d'acquisitions.

Les données obtenues auprès des neurologues concernent à la fois des patients sains et des patients atteints de maladies neurologiques. Nous avons actuellement quatre maladies différentes, chacune pouvant potentiellement influencer le comportement du patient lors du test de Romberg. Celles-ci sont :

- *Syndrome cérébelleux* : le patient présente des symptômes divers entrainés par une atteinte du cervelet. Il présente généralement une ataxie cérébelleuse.
- *Parkinson*
- *Syndrome vestibulaire* : problème de l'oreille interne
- *Trouble proprioceptif*, qui peut se caractériser par la présence d'une ataxie proprioceptive.

Les trajectoires de centre de pression sont acquises avec une *wii balance board*. La fréquence d'acquisition est de 20Hz. Les données se présentent sous forme de deux trajectoires, une pour chaque acquisition (yeux ouverts, yeux fermés). Les résultats présentés plus loin ont été effectués sur un ensemble de 143 patients, incluant 60 patients sains et 38 présentant un trouble proprioceptif. Les autres catégories de patients ne présentent pas encore assez de cas pour permettre une analyse statistique, avec par exemple seulement 4 cas de patients ayant un syndrome cérébelleux. Sur le long terme, ce nombre devrait augmenter significativement, mais pour l'instant, nous ne nous intéressons qu'à savoir classifier un sujet sain d'un sujet atteint de trouble proprioceptif.



(a) Malade (b) Sain

FIGURE 1: Exemples de trajectoires du centre de pression, yeux fermés

2 Exploration des données

2.1 Descripteurs

Notre travail consiste à étudier la trajectoire du centre de pression des patients afin de déterminer les différences entre un sujet sain et un sujet malade. Cependant à l'œil nu, on ne voit pas de différences notables entre les différentes catégories de personnes. Nous avons donc recherché des descripteurs afin d'avoir un résumé le plus exhaustif possible des informations de chacune des trajectoires. Pour cela, différentes approches, intuitives ou empruntées à la littérature biologique, ont été utilisées.

Dans un premier temps, nous ramenons la trajectoire en deux dimensions à une courbe unidimensionnelle qui la résume. Nous nous intéressons donc aux valeurs suivantes :

- Les deux coordonnées cartésiennes
- Les coordonnées polaires avec comme origine, le centre d'équilibre de la trajectoire. Il est calculé comme étant le barycentre des points de la trajectoire, avec une pondération liée à la vitesse en ce point.
- La courbure
- Le produit vectoriel entre vitesse et accélération
- L'accélération
- Le produit scalaire entre vitesse et accélération qui correspond à la puissance massique instantanée. Intuitivement, cette valeur représente l'effort du patient à atteindre une position d'équilibre et à y rester.
- La corrélation croisée entre la puissance et le rayon, qui permet d'observer la relation entre la quantité de réaction et l'éloignement au centre d'équilibre.

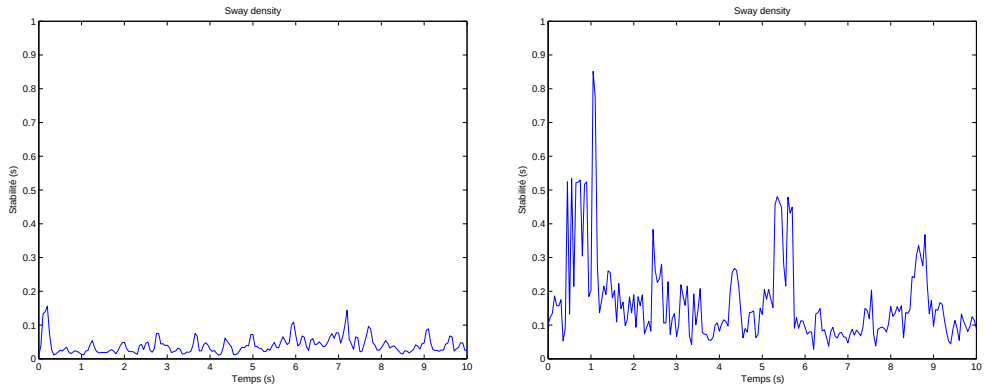
Étant donné que la trajectoire étudiée est celle du centre de pression, et non celle du centre de gravité, le produit scalaire vitesse-accélération ne correspond pas exactement à la puissance délivrée par le sujet. Néanmoins, il est communément admis que la différence entre les deux trajectoires est assez faible pour pouvoir les assimiler [Winter, 1995, Baratto et al., 2002].

2.1.1 Sway Density

Il s'agit de constructions graphiques simples qui tentent d'identifier les actions d'anticipation et de réaction d'une personne. Le balancement lorsque la personne est debout et sans perturbation extérieure, est similaire à une suite de petites chutes ralentie par les anticipations et les réactions. Le principe de ce descripteur est de trouver les zones stables. Le but est de trouver les zones où il y a un amas de points consécutifs et d'interpréter ces zones comme étant des intervalles de forte stabilité. C'est à dire du nombre de point consécutif qui sont, pour un rayon fixé, dans le même cercle. Ainsi, lorsqu'on trace la courbe représentant le *Sway Density*, on remarque une alternance de pics et de vallées. Les pics représentent les instants de stabilités, alors que les vallées représentent les moments d'instabilités, les moments où le corps cherche à rapidement se mettre dans un état stable ([Baratto et al., 2002]).

Les données qui peuvent être intéressantes via cette transformation sont principalement :

- L'amplitude des pics, qui permet d'estimer le degré de stabilité du patient.
- L'intervalle de temps entre deux pics consécutifs, ce qui représentent le taux de production de commande.
- La distance entre deux pics consécutifs, ce qui correspond à l'amplitude des commandes.



(a) Malade (b) Sain

FIGURE 2: Comparaison de courbes de Sway Density entre un sujet malade et un sujet sain

2.1.2 Outils statistiques

Nous calculons ensuite sur ces fonctions extraites différents paramètres statistiques les caractérisant, qui vont nous servir de descripteurs ([Hastie et al., 2009]). Nous empruntons naturellement ces extracteurs au domaine statistique, nous choisissons alors :

- Le minimum
- Le maximum
- La moyenne
- Le kurtosis qui est une quantification de l'aplatissement de ces valeurs.
- Les moments à différents ordres
- La variance
- Le coefficient d'asymétrie.
- L'écart type

En plus de cela, nous étudions la médiane, et les quantiles 0.1 et 0.9 qui peuvent s'interpréter comme des minima et maxima robustes. Ils permettent de limiter des erreurs de capture qui peuvent faire apparaître des valeurs aberrantes, et traduisent un comportement plus général de la trajectoire.

2.1.3 Approche biologique

Il y a quelques descripteurs biologiques classiques tel que la longueur de la trajectoire et l'aire balayée par celle-ci, centrée en le centre d'équilibre ([Hufschmidt et al., 1980]). La direction privilégiée et son importance permettent également de décrire la trajectoire.

2.1.4 Transformée de Fourier

Nous avons également travaillé sur la transformée de Fourier sur chacune des coordonnées, afin de transformer les données discrètes du domaine temporel dans le domaine fréquentiel, puis nous avons étudié les fréquences en dessous desquelles nous avions un certain pourcentage de la fonction. Cependant, cette transformée est pour le moment inutilisable car la fréquence d'acquisition est de 20Hz ce qui est trop faible. Il sera éventuellement possible à l'avenir de doubler cette fréquence et ainsi exploiter cette transformée.

2.1.5 Approximation par ondelettes

L'un des derniers type de descripteurs utilisés est l'erreur entre la courbe et son approximation par transformation par ondelettes, avec différents degrés d'approximation et différentes mesures d'erreur.

Finalement, une vingtaine d'approches différentes a été utilisée. Chacune d'entre elles a été appliquée avec les diverses données et dans les contextes des yeux ouverts ou des yeux fermés puis le ratio entre ces deux contextes. Ce qui donne, au final, 684 descripteurs différents. Parmi ces descripteurs, nous en avons qui viennent du domaine mathématique et statistique, biologique, ou encore de la physique. Cependant, l'efficacité de ces descripteurs pour classifier les sains des malades est variable. Même si certains ont l'air plus intuitif que d'autres, nous avons il ne s'agit pas forcément des plus efficaces.

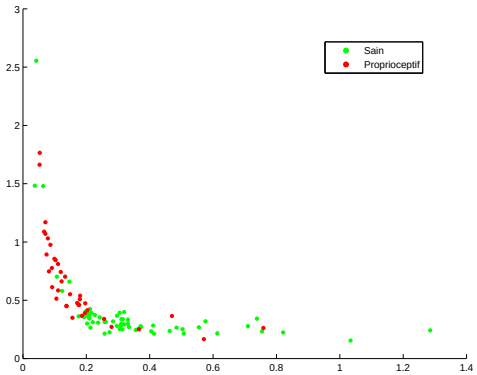


FIGURE 3: Répartition des individus selon les deux meilleurs descripteurs

2.2 Sélection des descripteurs

Le nombre de descripteurs au final est assez conséquent, et ils n'ont pas tous la même efficacité pour la classification. Le fait de sélectionner les descripteurs a plusieurs avantages. Tout d'abord, cela permet d'éliminer les données inutiles ou peu efficaces, afin d'améliorer la classification. En effet, avec un grand nombre de descripteurs, les algorithmes d'apprentissage ont plus de chance de sur-apprendre. De plus, la sélection d'un petit groupe de descripteurs efficaces permet aux neurologues qui auront accès à ces données d'avoir une meilleure compréhension des phénomènes qui régissent le mouvement du patient.

La sélection de ces données a été effectuée avec la méthode des *Random Forests* [Breiman, 2001]. Le principe est de construire un ensemble d'arbres de décision aléatoires sur notre ensemble de données. Chaque arbre de la forêt est entrainé avec un petit sous-ensemble aléatoire de descripteurs et environ deux tiers des acquisitions, choisis aussi aléatoirement. L'évaluation d'un descripteur se fait en calculant la différence d'erreur commise par les arbres sur le tiers des données restantes, et celle commise après avoir permuté aléatoirement les valeurs du descripteur sur les données. Si l'erreur commise après la permutation est plus importante, alors cela signifie que le descripteur était pertinent, et qu'on a perdu de l'information.

Au final, nous obtenons la liste des descripteurs triés selon un ordre d'importance. Une première remarque est que les descripteurs les plus efficaces portent sur l'acquisition yeux fermés. Ce résultat semble normal, étant donné que les troubles de la proprioception se manifestent par une désensibilisation du sujet à la position de ses propres membres. Ce déficit

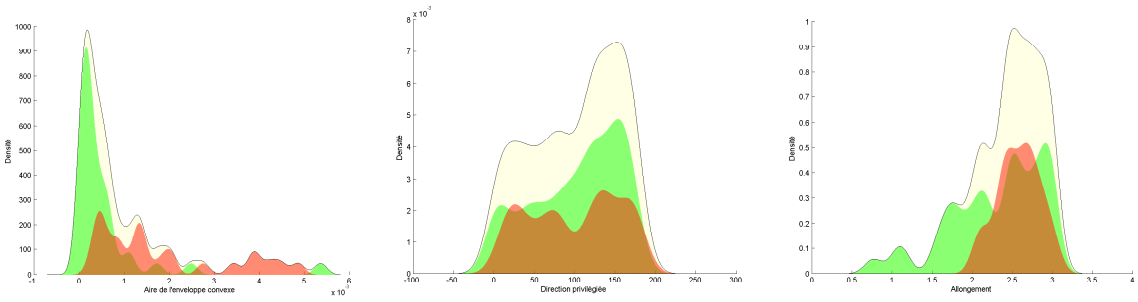
peut être contrebalancé par la vue, mais une fois les yeux fermés, le sujet a des difficultés à se situer spatialement. Nous trouvons alors plus aisément des descripteurs discriminants les sains et les malades lors de l'acquisition yeux fermés que yeux ouverts.

Les meilleurs descripteurs sont généralement des valeurs statistiques extraites des courbes de l'accélération, de la vitesse, et de Sway Density. Ces courbes traduisent le mouvement chaotique du sujet atteint, qui va avoir du mal à contrôler son mouvement pour rester stable.

2.3 Mise en évidence des meilleurs descripteurs

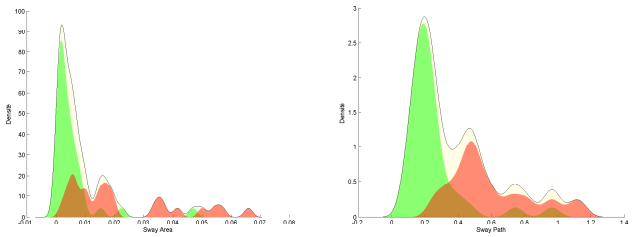
L'étape de sélection des descripteurs permet de mettre en avant des paramètres non classiques, pas forcément intuitifs, et souvent plus efficaces que ceux présentés dans la littérature.

Nous supposons que la répartition des données pour un descripteur correspond à une somme de gaussiennes. Nous effectuons alors une évaluation de densité avec nos données pour observer l'allure de cette répartition. Les courbes vertes correspondent aux patients sains, alors que les rouges correspondent aux patients malades.

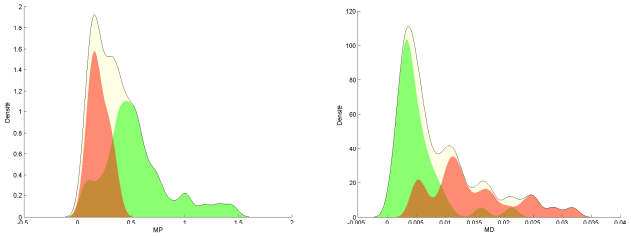


(a) Surface (b) Direction privilégiée (c) Allongement

FIGURE 4: Répartition des densités de trois descripteurs intuitifs

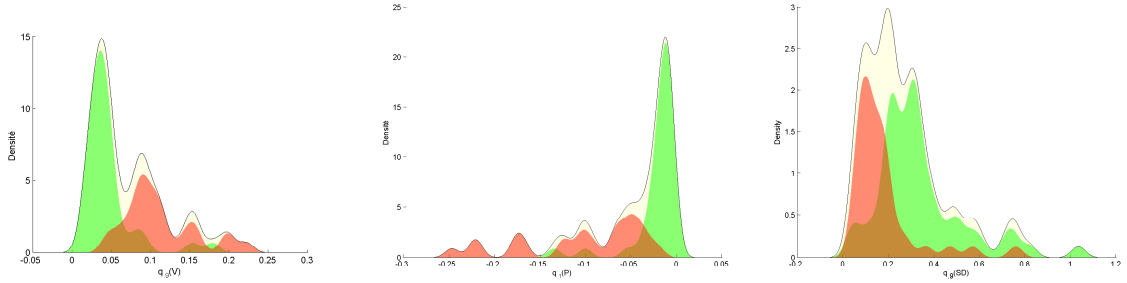


(a) SwayArea (b) SwayPath



(c) MP (d) MD

FIGURE 5: Répartition des densités des quatre descripteurs bibliographiques



(a) 9e décile de la vitesse (b) 1er décile de la puissance (c) 9e Sway Density

FIGURE 6: Répartition des densités de trois descripteurs originaux

Nous observons pour certains descripteurs intuitifs que les zones de fortes densité chez les sains et les malades coïncident assez souvent, ils n'apportent donc aucune information.

Certains descripteurs bibliographiques présentent sur leur courbe de densité des pics de densité qui se chevauchent moins, et traduisent donc mieux la différence sain/malade.

Cependant, avec nos descripteurs, nous obtenons des résultats bien plus intéressants. Les pics de densité étant clairement séparés, cela montre que ces descripteurs traduiraient plus efficacement les différences que nous souhaitions mettre en avant.

3 Classification

3.1 Apprentissage statistique

L'apprentissage statistique est un ensemble de méthodes statistiques pour analyser la relation d'une variable par rapport à une ou plusieurs autres. Le principe est de déterminer la relation entre $X \in \mathcal{X} \subseteq \mathbb{R}^d$ un vecteur de l'espace des descripteurs $\mathcal{X}$, correspondant à une acquisition, et $Y \in \mathcal{Y}$, la maladie associée. On cherche donc une fonction $f : \mathbb{R}^d \to \mathbb{R}$ telle que $Y = f(X)$, alors qu'on observe partiellement f à partir d'un échantillon $\{(x_i, y_i), i = 1..n\}$. Dans notre cas, l'échantillon est de taille 143.

d, le nombre de descripteurs, se trouve être très grand après avoir fait une exploration des données, notamment parce que l'on peut observer chacune des deux trajectoires de l'acquisition, yeux ouverts et fermés, indépendamment ou bien en croisant les résultats. On obtient alors 684 descripteurs différents. On peut alors procéder à une sélection des données avant d'effectuer un apprentissage pour éventuellement améliorer l'efficacité des algorithmes. En effet, le nombre de données actuellement est petit devant le nombre de descripteurs possibles. Dans l'espace des descripteurs, les données sont alors toutes très espacées, par rapport à la distance euclidienne, les unes des autres, ce qui rend certaines méthodes moins efficaces.

3.2 Méthodes et résultats

Plusieurs méthodes sont utilisées pour la classification. Pour chacune, nous effectuons une validation croisée, avec 75% de l'échantillon servant à l'apprentissage, et 25% au test. A titre de comparaison, le classifieur qui prédit que tout individu est sain aurait un taux d'erreur de 40% sur notre échantillon.

Support Vector Machine (SVM) Le principe général des SVM est d'effectuer une transformation non linéaire de l'espace des descripteurs en un espace plus grand dans lequel il y a possiblement une séparatrice linéaire des données.

Cette méthode semble efficace dans notre cas, avec un taux d'erreur compris entre 17% et 22%.

Arbres de décision. L'algorithme CART produit des arbres de décision, utilisés pour répartir notre population de patients en groupes homogènes, grâce à un ensemble réduit de descripteurs. Chaque nœud de ces arbres correspond à un critère binaire : un seuil sur un descripteur, qui crée deux sous-groupes. Cet algorithme est récursif, ce qui signifie que lorsqu'il choisit un critère sur un nœud, il effectue les même opérations sur les sous-groupes obtenus et ne revient jamais en arrière.

Pour sélectionner un critère à une étape donnée, il explore toutes les possibilités et choisit celle qui minimise un certain critère (ici, le coefficient de Gini). De ce fait, le nombre de descripteurs influe peu sur le taux d'erreur commis lors de la classification : entre 18% et 22%. La grande quantité de descripteurs, dont la plupart sont peu efficaces, pourrait cependant entrainer un sur-apprentissage.

Un autre avantage des arbres de décision est la lisibilité de ceux-ci, ce qui est un avantage pour présenter simplement des résultats importants (Fig.7).

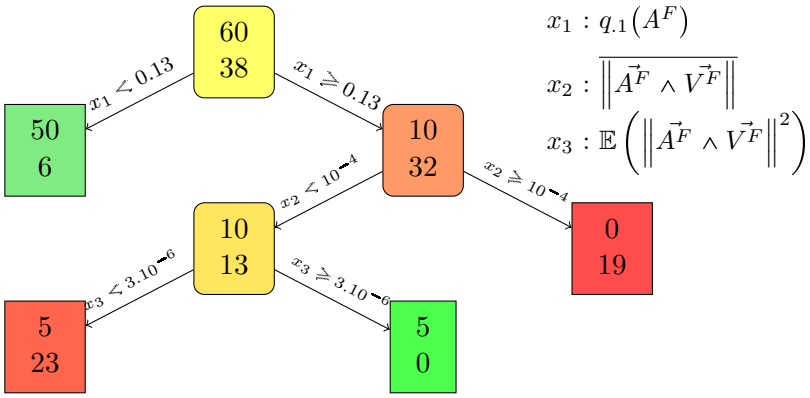


FIGURE 7: Arbre de décision CART avec répartition sains/proprioceptifs dans chaque partition

K plus proches voisins (KNN). Le principe de cet algorithme est le suivant : pour classifier une nouvelle donnée, on considère l'ensemble de ses K plus proches voisins dans l'espace des descripteurs. La classe majoritaire dans cet ensemble correspond à la prédiction associée.

Pour un nombre limité de descripteurs, environ une dizaine, cette méthode fonctionne relativement bien, avec en moyenne 17% d'erreur en classification. Cependant, à mesure que le nombre de descripteurs devient grand, le taux d'erreur augmente jusqu'à atteindre rapidement les 40% dans le meilleur des cas. Cela s'explique par le nombre limité de données, assez faible devant le nombre maximum de descripteurs. Les dizaines de données réparties dans cet espace de grande dimension sont toutes éloignées les une des autres par rapport à la norme euclidienne. De ce fait, il est difficile de classifier une nouvelle donnée par rapport à ses voisins.

Régression Logistique Cette méthode évalue la probabilité d'une donnée à appartenir à une classe. Elle peut s'avérer utile si l'on souhaite quantifier l'état de maladie d'un patient. Cependant, les résultats en classification sont moins bons qu'avec les autres méthodes : environ 22% d'erreur avec moins de 10 descripteurs, et ce taux augmente considérablement avec la dimension.

4 Conclusion

Nos objectifs dans ce stage étaient de pouvoir prédire l'état sain ou malade d'une personne à l'aide d'apprentissage. Nous avons fait une quantification du test de Romberg, dont l'acquisition est améliorée grâce aux nouvelles technologies, en particulier la *Wii Balance Board*. Ce test a l'avantage d'être facile à mettre en place et d'être relativement peu coûteux. En effet, il permet de connaître l'état d'un patient sans qu'elle ait trop d'effort à faire. Certains malades ont du mal à marcher, ainsi le fait que les patients n'aient qu'à rester debout rend le test intéressant et potentiellement utilisable pour quantifier l'état de la maladie. Dans un état avancé de la maladie, les patients ont généralement beaucoup de mal à marcher. Cependant, à un stade précoce, il est plus difficile de la déceler et de distinguer un sujet sain d'un sujet malade. Il serait donc intéressant pour les neurologues d'être en mesure de quantifier la maladie.

L'utilisation du système *Wii Balance Board* est intéressante car elle montre que les nouvelles technologies permettent des avancées dans des domaines a priori très éloignés. Ce qui a été créé en premier lieu pour un jeu, a vu son utilisation détournée pour être utilisée en médecine.

L'un des intérêts majeurs de ce projet était que nous n'étions pas limités aux mathématiques ou à l'informatique. Lors de ce stage, nous avons eu des interactions avec beaucoup de disciplines différentes. En effet, puisque le sujet de notre projet était la réaction du corps dans un état debout, sans perturbation, les yeux ouverts ou fermés, il fallait s'intéresser à la bibliographie correspondante dans le domaine biologique. Ainsi lors de la recherche de nos descripteurs, qui constitue la majorité du projet, nous avons utilisé des articles de statistiques, de biologie et de physique. Cet ensemble hétéroclite de descripteur nous a permis d'avoir le « résumé » le plus exhaustif des informations de la trajectoire.

Même si actuellement, il y a une base de donnée de patients et de sains relativement restreinte, nous avons été introduits à l'apprentissage supervisé. En ne prenant que la maladie ayant un nombre de patient le plus élevé, les proprioceptifs, et les sains, nous avons pu faire une première approche limitant le bruit et le surapprentissage tout en nous permettant d'avoir une base solide pour l'avenir. Travailler uniquement sur deux classes en mettant de coté les trois autres maladies nous a permis d'obtenir une première version effective du projet, ce qui est un premier pas vers son aboutissement.

Dans un avenir proche, le projet va s'étendre, et de nouveaux médecins vont participer ce qui augmentera considérablement la base de donnée qui sera disponible pour les tests. Ainsi, cela nous permettra d'utiliser notre travail sur plusieurs maladies. Il sera aussi possible, et c'est ce qui est le plus intéressant, de voir l'évolution d'un patient au fil d'un traitement. Le projet pourra être utilisé pour quantifier l'état de maladie des patients. Cela permettra de savoir si un traitement est efficace ou non.

Pour terminer, la participation à ce projet a pu nous permettre d'un point de vue personnel de voir comment on travaillait dans un laboratoire de mathématiques et de nous donner un premier aperçu du travail de chercheur. C'est pourquoi nous remercions Nicolas Vayatis, Emile Contal et Themistoklis Stefanakis d'avoir bien voulu nous encadrer lors de ces quelques mois

de stages et d'avoir été disponibles pour répondre à nos questions.

Références

L. Baratto, P.G. Morasso, C. Re, and G. Spada. *A New Look at Posturographic Analysis in the Clinical Context : Sway-Density Versus Other Parameterization Techniques.* Motor Control, 2002.

L. Breiman. *Random Forests.* Machine Learning, 2001.

T. Hastie, R. Tibshirani, and J. Friedman. *Elements Statistical Learning.* Springer, 2009.

A. Hufschmidt, J. Dichgans, K.-H. Mauritz, and M. Hufschmidt. Some methods and parameters of body sway quantification and their neurological applications. *Archiv fur Psychiatrie und Nervenkrankheiten*, 1980.

D A Winter. Human balance and posture control during standing and walking. *Gait and Posture*, 1995.