Étude de cas 2: Analyse de l'activité cérébrale via l'électroencéphalographie

LIEDRI Ibtissam
KOTTI Hend
HASSINI Houda

19 février 2020

1 Introduction

La maladie d'Alzheimer est une pathologie très difficile à diagnostiquer et à distinguer des autres maladies comme la démence et les troubles cognitifs subjectifs. Il est nécessaire d'utiliser des bio-marqueurs du liquide céphalorachidien ou de tomographie informatisée à émission de photons uniques afin de détecter cette maladie de façon efficace.

2 Objectif

L'objectif de cette étude est l'analyse de l'activité cérébrale via l'électroencéphalographie dans le contexte de pathologie neurodégénérative 'Alzheimer' dans un but d'une classification supervisée des individus pour diagnotiser l'Alzheimer puis dans le but d'une classification non supervisée afin de s'assurer que les classes affectées par des spécialistes sont cohérentes avec les résultats obtenus.

3 Bases de données

Nous disposons de 3 types de profils d'individus :

- 22 personnes SCI: **Subjective Cognitive Impairement**
- 22 personnes MCI: **Mild Cognitive Impairment** ou **Trouble cognitif léger**
- 28 personnes AD :**Attente de Démence**

Pour chaque individus des trois profils nous sommes en possession de matrice de la mesure de désynchronisation.Cette matrice est obtenue en plaçant 30 capteurs sur le cuir chevelu de l'individu selon une disposition spécifique puis nous mesurons les désynchronisations entre ces trente capteurs deux à deux. On obtient une matrice de taille 30x30 symétrique.
Ces matrices ont été réalisées pour différentes fréquences α, β, δ et θ qui correspondent au basses fréquences et dans cette étude, on ne s'intéressera qu'aux basse fréquences comprise entre 0-30 Hz (basses fréquences).

4 Traitements des données

Comme nous l'avons mentionné dans la partie précédente, chaque individu est représenté par une matrice symétrique de taille 30x30, il est donc nécessaire d'effectuer un travail de traitement avant de commencer la classification, l'approche proposée pour notre étude est la suivante:

- Nous commençons par choisir le Zoom : Étude de synchronisations, désynchronisations ou les deux. Notre groupe a choisi d'étudier la désynchronisation entre les électrodes.
- Les matrices des individus doivent être transformées en un vecteur descriptif avec 30 variables (représentants les électrodes). Pour ceci on utilisera plusieurs mesures telles que la moyenne, la variance DCF et le nombre de capteurs désynchronisés et on procédera pour cela en deux approches:
  1. **Approche 01:** On étudiera la désynchronisation en utilisant séparément les trois mesures: moyenne, variance DCF et nombre de capteurs désynchronisés.
  2. **Approche 02:** Combinaison des deux descripteurs: DCF et le nombre de capteurs désynchronisés.
- Nous voulons étudier ses individus selon différents seuils de synchronisation (100%, 70%, 50%, 30%, 20% et 10%), on répète donc l'étude pour chaque seuil. A la fin nous obtenons 4(bande de fréquences)x6(seuils)x30(électrodes) descripteurs pour chaque individu, c'est à dire 720 variables. Compte tenu du nombre de variables nous sommes face une base de données de type **"Large data"**, très peu d'observations et beaucoup de variables. Nous risquons un sur-apprentissage par la suite.
- Avant de réaliser une classification, il est nécessaire de remédier au problème de **"Large data"** rencontré précédemment. On effectue une sélection de variables comme suivant pour éviter le sur apprentissage :
  - Choix de la variable la plus importante pour chaque bande de fréquence et pour chaque seuil
  - Construction d'un jeu de données des 4x6 variables les plus pertinentes.
  - Re-sélection des variables en choisissant les 5 plus importantes variables.

5 Sélection de variables

Comme nous l'avons décrit dans la section précédente, la sélection de variable dans cette étude se fait en deux temps à cause de nombre important des variables que nous possédons. Pour réaliser cette sélection, nous utilisons l'algorithme des arbres aléatoires.
Les arbres aléatoires sont une catégorie d'arbres utilisée dans l'exploration de données et en informatique décisionnelle. Ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions en vue de la prédiction d'un résultat ou d'une classe. Chaque individu, qui doit être attribué à une classe, est décrit par un ensemble de variables qui sont testées dans les nœuds de l'arbre. Les tests s'effectuent dans les nœuds internes et les décisions sont prise dans les nœuds feuille. L'arbre est construit par partition récursive de chaque nœud en fonction de la valeur de l'attribut testé à chaque itération (top-down induction). Le critère optimisé est la homogénéité des descendants par

rapport à la variable cible. La variable qui est testée dans un nœud sera celle qui maximise cette homogénéité. Le processus s'arrête quand les éléments d'un nœud ont la même valeur pour la variable cible (homogénéité). Cette intégration de l'homogénéité dans le choix des variables retenus est le critère qui nous a poussé à choisir cette méthode. Étant donnée que nous ne possédons pas un grand nombre d'individus, nous choisissons de travailler avec une faible profondeur d'arbre et de préconiser un apprentissage avec la méthode bootstrap pour augmenter le nombre d'individus et réduire la variance de la méthode.
Voici un exemple de cette méthode de sélection de variables:

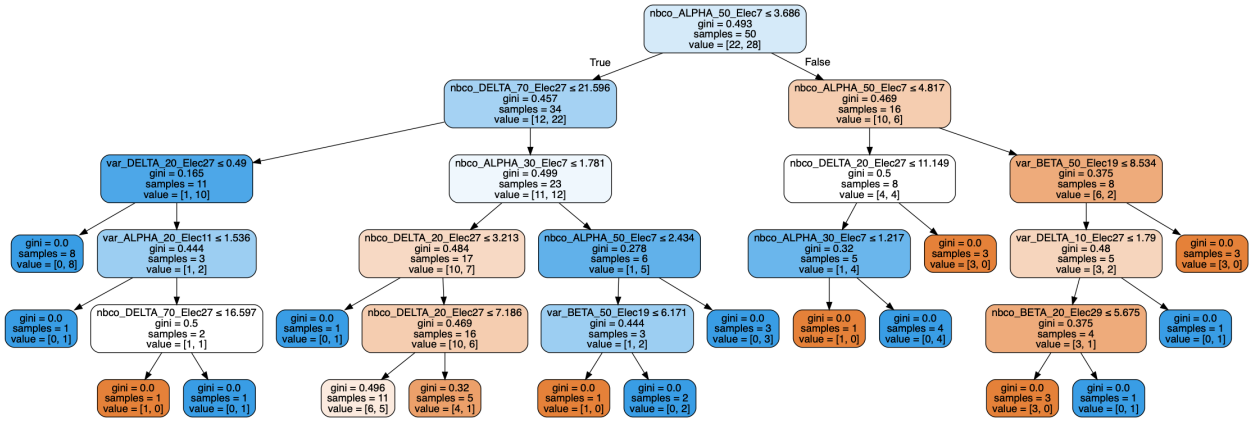


Figure 1: Extrait de l'arbre de décision pour la sélection sur le couple AD et SCI

Cette figure montre comment les variables permettent de créer l'arbre et comment celles-ci sont choisies par la suite dans l'ordre de pertinence .
En effet on remarque bien que la racine de l'arbre contient toutes les observations **"samples=50" : [22,28]**, ses deux nœuds fils sont associés aux deux sous-parties **"samples=34" et "samples=16"** obtenues par la première découpe du partitionnement, et ainsi de suite. Par exemple dans la première découpe on a toutes les observations avec une valeur de la variable **Nb-cnx-delta-70-elec-27** plus petite que $d = 21.596$ vont dans le nœud fils de gauche, et toutes celles avec une valeur de la variable **Nb-cnx-alpha-50-elec-7** plus petite que $d = 4.817$ vont dans le nœud fils de droite. La méthode sélectionne alors la meilleure découpe, c'est-à-dire le couple (variable , d) qui minimise une certaine fonction de coût. Dans notre cas il s'agit de chercher à diminuer la fonction de pureté de Gini:
$\sum_{i=k}^{K} \hat{p}_N^k \times (1 - \hat{p}_N^k)$ **où** $\hat{p}_N^k$ **est la proportion des éléments de classe k dans le nœud N**. Et donc à augmenter l'homogénéité des nœuds obtenus, un nœud étant parfaitement homogène s'il ne contient que des observations de la même classe.
Ainsi l'algorithme sélectionne les variables qui classifie le mieux les individus du couple AD et SCI dans cet exemple.

En plus des choix fait précédemment, nous avons opté pour garder les 5 variables les plus pertinentes lors de la seconde étape de la sélection. Tout d'abord nous avons utilisé une variables sonde comme indicateur du nombre de variable à sélectionner, cette approche avait tendance à nous donner un nombre de variable assez conséquents (12 variables) ce qui n'éliminait pas le risques de sur-apprentissage. Quand la première approche qui s'est avéré peu concluante nous avons opté pour un testé du nombre optimale des variables à sélectionner, nous avons testé des valeurs entre 2 et 12 et pour des valeurs de 5 à 7 variables nous obtenons de performances significativement bonnes avec des scores plutôt élevé mais

similaires. Nous privilégions donc un choix de 5 variables dans la suite de notre étude.

6 Classification supervisée

6.1 Première approche:

Dans un second temps, nous avons procédé à une classification supervisée visant à trouver les frontières entre les classes toujours dans un but de comprendre ce qui caractérise le plus un types de patient. Nous utilisons comme variables d'entrée celles sélectionnées dans l'étape précédente. Cette classification se fait par paires de deux classes d'individus, nous allons procédé à une classification qui sépare les **AD** des **MCI**, puis **MCI** des **SCI** et finalement les **AD** des **SCI**. Compte tenu du peu d'observations dont on dispose, on a utilisé les deux algorithmes **SVM** et **KNN** du fait de leur robustesse sur les données de petite taille.
Il faudra s'assurer qu'on a une bonne performance totale du classifieur avec un équilibre entre la valeur de la spécificité et la sensibilité. La sensibilité d'un algorithme est sa capacité à donner un résultat positif lorsqu'une hypothèse est vérifiée. Elle s'oppose à la spécificité, qui mesure la capacité d'un algorithme à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée.
On s'est basé sur ces deux métriques, ainsi que sur le score, qui est le taux des individus bien classés parmi le nombre total d'individus, pour comparer les modèles entre eux et en choisir le plus performant.
On peut résumer la spécificité et la sensibilité dans les formules suivantes:
Avec :

- Spécificité $\frac{VP}{VP + FN}$
- Sensibilité $\frac{VN}{VN + FP}$

Figure 2: Sensibilité et spécificité

- **VP** représente le nombre d'individus vérifiant l'hypothèse, avec un test positif
- **FP** représente le nombre d'individus ne vérifiant pas une hypothèse, avec un test positif
- **FN** représente le nombre d'individus vérifiant l'hypothèse, avec un test négatif
- **VN** représente le nombre d'individus ne vérifiant pas l'hypothèse, avec un test négatif

Cette classification nous a donné les résultats ci-dessous:

	Moyenne						**Variance CFD**						**Nombre de connexions**					
	knn(metric='minkowski',n = 5)			svm(kernel='poly',deg= 2)			knn(metric='minkowski',n = 5)			svm(kernel='poly',deg= 2)			knn(metric='minkowski',n = 5)			svm(kernel='poly',deg= 2)		
	acc	spe	sen	acc	spe	sen	acc	spe	sen	acc	spe	sen	acc	spe	sen	acc	spe	sen
ad/mci	0.6	0.5	0.7	0.5	0.5	0.7	0.7	0.6	0.7	**0.7**	**0.7**	**0.8**	0.6	0.6	0.8	**0.7**	**0.7**	**0.7**
sci/mci	0.8	0.8	0.8	0.8	0.6	0.7	0.8	0.8	0.8	**0.8**	**0.8**	**0.8**	0.8	0.8	0.8	**0.8**	**0.8**	**0.8**
sci/ad	0.5	0.5	0.7	0.7	0.5	0.6	0.6	0.6	0.6	**0.6**	**0.6**	**0.6**	0.7	0.7	0.7	**0.6**	**0.6**	**0.6**

Figure 3: Résultats de la classification supervisée (1ère approche)

On remarque en observant les résultats du tableau ci-dessus qu'une classification sur la base du descripteur **moyenne** donne de faibles performances, contrairement à celle sur la base de la variance DCF ou du nombre de connexions, qui ont des performances presque similaires. Le descripteur **moyenne** ne permet donc pas d'avoir assez d'information afin de caractériser les 3 type de patients, elle n'est pas adapté à notre problématique.
Le SVM donne de meilleures performances par rapport au KNN, cependant a plus de mal à séparer les **SCI** des **AD**, mais sépare bien les **SCI** des **MCI**.
Pour comparer les performances on s'est basé sur la sensibilité et la spécificité comme indiqué précédemment. En effet plus le test est sensible, moins il comportera de faux négatifs, et mieux il permettra, s'il est négatif, d'exclure la maladie. Ainsi plus le test est spécifique, moins il occasionnera de faux positifs, et mieux il permettra, s'il est positif, de confirmer la maladie.
On cherche alors les tests qui ont à la fois une sensibilité et une spécificité très élevées.

6.2 Deuxième approche :

Suite à ces résultats, nous avons envisagé de combiner deux descripteurs qui sont **la variance DCF** et **le nombre de connexions entre les électrodes**, la combinaison de ces deux descripteurs n'est pas forcément celle qui nous donnera les meilleures performances, nous l'avons cependant privilégié pour son évidence mais il faudra notamment étudier les autres possibilités que nous n'avons pas pu exploiter faute de temps.
Pour cette deuxième approche qui consiste à se baser sur la combinaison des deux descripteurs variance DCF et le nombre de connexions, on a envisagé une sélection de variables en trois étapes comme suit:

- Tri des variables par ordre d'importance en utilisant les arbres de décision séparément pour le nombre de connexion et la variance CDF.
- Choix de la variable la plus importante pendant l'étape 1. A l'issu de cette étape, on a 4x6 variables sélectionnées pour chacun des descripteur.

- Ré-application d'un arbre de décision et choix de 5 variables les plus importantes parmi les 24 pour chaque descripteur.
- Concaténation des variables sélectionnées pour chaque descripteur.
- Re-sélection finale de 5 variables par arbre de décision en utilisant l'ensemble des variables issu de la concaténation.

Les variables sélectionnées sont les suivantes:

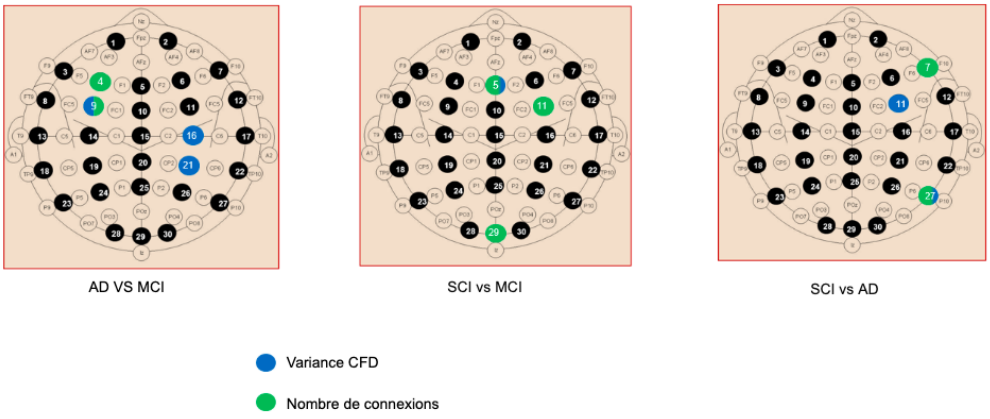


Figure 4: Variables sélectionnées(2ème approche)

On remarque que les zones des variables sélectionnées diffèrent d'une paires d'individus à une autre. Pour le couple des individus AD/MCI, on constate que les variables sélectionnées sont celles situées dans la partie frontale et la partie centrale. Le couple SCI/MCI se caractérise lui par des électrodes sélectionnées dans la partie occipitale alors que le couple SCI/AD par des électrodes dans la partie temporelle. En effet, les fréquences alpha sont généralement distribuées dans la zone occipitale pour les sujets sains; chez les patients atteints de AD, ces fréquences sont plus localisées dans des zones antérieures à mesure que la maladie progresse.
Les performances des classificateurs SVM et KNN ont été estimées par **la validation croisée** par paires de sujets, qui est connue pour fournir une estimation non biaisée du score.
Les résultats de cette deuxième approche sont résumés dans le tableau ci-dessus:

	SVM (Linear)			**KNN (4 voisins)**		
	acc	spe	sen	acc	spe	sen
ad/mci	0.80	0.75	0.85	**0.80**	**0.75**	**0.85**
sci/mci	0.86	0.85	0.95	**0.89**	**0.85**	**0.95**
sci/ad	0.85	0.80	0.85	**0.80**	**0.80**	**0.84**

Figure 5: Résultats de la classification supervisée (2ème approche)

On remarque que les résultats de la classification supervisée pour cette deuxième approche sont similaires pour les deux algorithmes. Les performances sont cependant meilleures que celles obtenues lors de la première approche.
Au vu de la quasi-similarité des performances des deux algorithmes, on privilégiera le **KNN** pour sa simplicité comparée à celle du **SVM**.

Les résultats de la classification ont montré qu'une classification correcte avec un taux de 80% est atteinte lors de la discrimination des patients AD des sujets SCI, avec une spécificité (proportion de sujets AD bien classés) de 80% et une sensibilité (proportion de patients SCI bien classés) de 84%. Ce résultat démontre une fiabilité significative quoique pas à 100% fiable, des fonctionnalités utilisées et la méthode proposée pour détecter la maladie d'Alzheimer. Le résultat montre également une très bonne détection des sujets MCI avec une spécificité aux alentours supérieure à 85% malgré le fait que les sujets MCI ne sont pas des sujets totalement sains puisqu'ils souffrent de problèmes de mémoire.

Pour le cas des SCI/MCI, on atteint un score de 86% avec une spécificité(proportion de sujets SCI bien classés) de 85% et une sensibilité(proportion de patients MCI bien classés) de 95%. Cette séparation entre les **MCI** et les **SCI** est meilleure dans cette deuxième approche comparée à la première approche qui se basait sur l'étude de chaque descripteur séparément.

7 Sous-profils cognitifs

L'étude réalisée jusqu'à présent a permis de voir qu'il est possible de séparer les trois types de patients de façon assez correcte, certaines performances ont atteint 80% de scores surtout pour les types de patients les plus éloignés comme **AD** et **SCI**. Cependant il en ressort que la classification n'est pas fiable à 100% et une difficulté plus significative à séparer les malades de types **AD** des **MCI**, ces deux types de patients sont très similaires et les experts ont aussi une certaine difficulté à les séparer. Ces raisons ont motivé la réalisation d'une classification non supervisée qui servira dans un premier temps pour analyser où et pourquoi les erreurs de la classification supervisée occurrent, cette étude permettra aussi de mettre en lumière la labellisation réalisée par les experts et analyser sa pertinence et sa fiabilité car nous pouvons imaginer que la complexité de la maladie a causé quelques labellisations erronées.

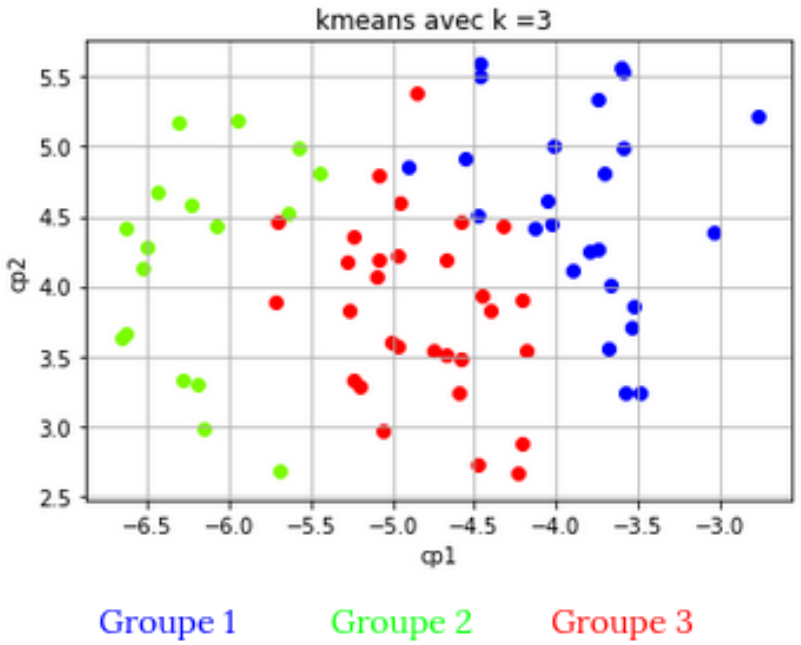
7.1 Cadre général de la réalisation

Afin de réaliser l'étude précédemment décrite, nous avons fait le choix de travailler sur l'ensemble des 15 variables sélectionnées lors de la classification. Nous allons travailler sur une base de données constituée de 72 individus (28 **AD**, 22 **MCI** et 22**SCI**) décrits par l'ensemble des 15 variables.

7.2 Classification par algorithme des K-moyennes

Dans un premier temps, nous avons choisi de réaliser une classification par l'algorithme des K-moyennes, ce choix est motivé principalement par la simplicité de l'interprétation de ses résultats.
Cependant, l'algorithme suppose une hypothèse forte qui consiste en un choix préalable du nombre de classes. Nous allons procéder à plusieurs tests avec des valeurs différentes pour

le nombre de groupes (K=3,4,....)
À l'aide de l'algorithme **t-SNE**, qui est une méthode non-linéaire permettant de représenter un ensemble de points d'un espace à grande dimension dans un espace à deux dimensions, nous allons représenter les résultats des K-moyennes dans un plan. Une étude des relations linéaires entre les variables sélectionnées à montrer que celle-ci sont trop faible ce qui justifie le choix d'un algorithme de réduction de dimensions non linéaires. Nous allons d'abord analyser le résultat pour K=3, ce nombre étant celui des types proposés par les experts. La figure suivante représente les groupes obtenus:



Comme la figure le montre les trois classes sont fortement déséquilibrées, cependant en analysant de près les répartitions on remarque que les classes suggèrent plus ou moins des dominances d'une classes dans chaque groupe. Les résultats sont synthétisés dans les tableaux suivants:

Groupe 1	Groupe 2	Groupe 3
24	33	15

Figure 7: Tableau détaillant le nombre d'individus dans chaque groupe

Groupe 1	Groupe 2	Groupe 3
6 AD, 3 MCI et 15 SCI	16 AD, 10 MCI et 7 SCI	6 AD et 9 MCI

Figure 8: Tableau détaillant la nature de chaque individus dans chaque cluster

On remarque que malgré ces déséquilibres les résultats montrent une dominance des **SCI** dans le groupe 1, de **AD** dans le groupe 2 et une dominance de **MCI** dans le groupe 3.

On peut aussi voir que dans le groupe à dominance **MCI** nous avons une forte présence des **AD** aussi ce qui démontre entre autre la difficulté particulière de séparer les deux types de patients.

Les groupes sont tous dominés par un type spécial de patients, ils peuvent donc être caractéristique de chacun de ces types. La figure suivante résume les résultats les plus marquants et le plus caractéristique des types de patients d'après la classification non supervisée:

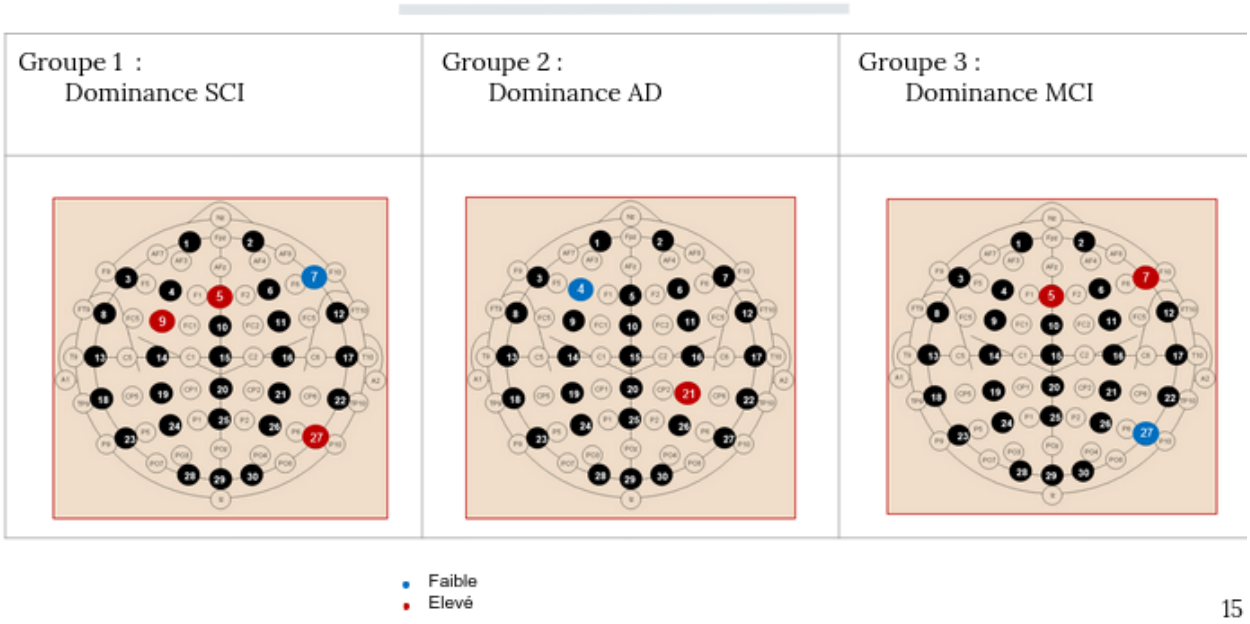
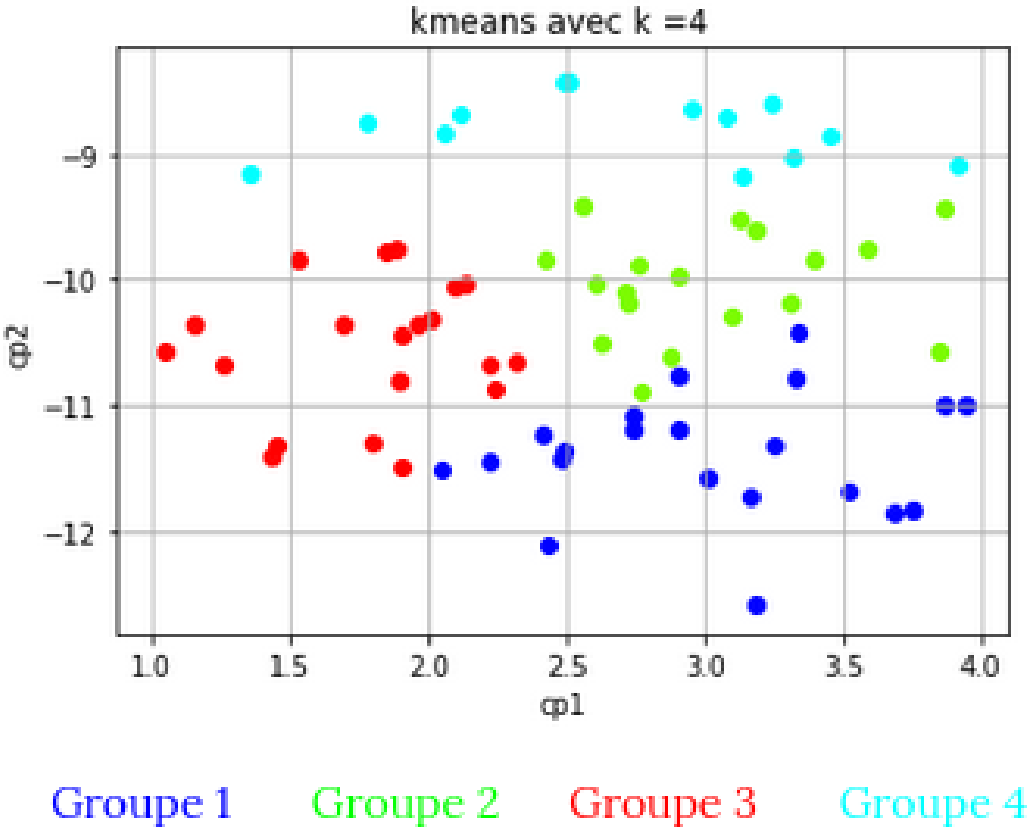


Figure 9: Caractéristiques des 3 groupes

On peut voir sur cette figure que les patients de type **SCI** sont caractérisés par de fortes valeurs du nombres de connexion et de variance CDF au niveau des électrodes frontales (Électrode 5 et 9) mais aussi de fortes valeurs de variance CDF au niveau de l'électrode T6. Tandis que le groupe à dominance **AD** sont des sujets avec des valeurs du nombre de connexion et de variance CDF très faibles aux niveaux de l'électrode frontale 4 et une forte valeur du nombre de connexion et de variance CDF au niveau de l'électrode 21 et le groupe 3 qui est dominé par les patients de type **MCI** sont caractérisés par de fortes valeurs du nombres de connexion et de variance CDF au niveau des électrodes frontales (Électrode 5 et 7) mais de faibles valeurs de variance CDF au niveau de l'électrode T6.

Le déséquilibre observé lors de la classification précédente ainsi que l'ensemble des résultats pourraient suggérer la présence d'une classe intermédiaire, cette hypothèse nous pousse à explorer la classification avec un nombre de classe supérieur à 3, on présentera dans la suite les résultats pour K=4. Les résultats de cette classification sont les suivants:



La répartition des groupes et des individus par groupes sont résumés dans les deux tableaux suivants:

Groupe 1	Groupe 2	Groupe 3	Groupe 4
21	18	20	13

Figure 11: Tableau détaillant le nombre d'individus dans chaque groupe

Groupe 1	Groupe 2	Groupe 3	Groupe 4
6 AD, 2 MCI et 13 SCI	15 AD, 2 MCI et 1 SCI	3 AD et 9 MCI et 8 SCI	4 AD et 9 MCI

Figure 12: Tableau détaillant la nature de chaque individus dans chaque cluster

On remarque ici que la répartition des individus entre les 4 groupes est beaucoup plus équilibrée. La dominance d'un type de patient est plus marqué avec cette classification sur les groupes 1,2 et 4. Or le groupe 3 reste un groupe avec un mélange hétérogène de typologie de patient, il contient des représentants des trois types étudiés. En effectuant 100 fois l'algorithme, on remarque que les individus de ce groupe se retrouvent toujours ensemble au sein d'un même groupe, sur l'image précédente ces individus sont représentés en couleurs bleu cyan.
Dans l'ombre de ses résultats, on s'intéresse de plus près à ces individus et des résultats

qu'ils ont obtenus lors des classifications supervisées réalisées précédemment. On s'aperçoit que ces individus sont des patients sur lesquels nous obtenons une erreur de classification supervisée. Ce résultat nous a poussé à explorer de plus prêt leurs caractéristiques de ces individus sur l'ensemble des 15 variables.
Il en ressort que ces individus se distinguent par une particularité par rapport au nombre de connexion de l'électrode 5 pour la fréquence α et la seuil 10% et la variance de l'électrode 21 pour la fréquence β et le seuil 50%. Le tableau suivant résume les valeurs observées pour l'ensemble des 20 individus du groupe 3 sur les deux électrodes:

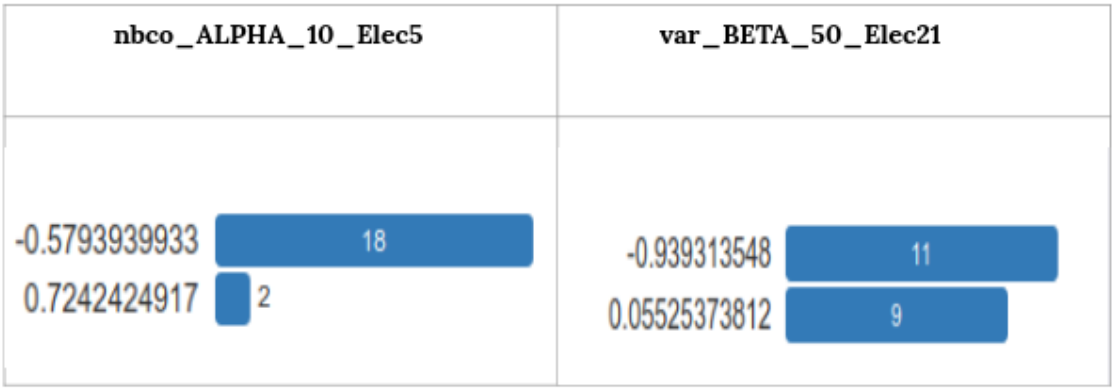


Figure 13: Particularité du groupe 4

On peut voir que 18 sur 20 individus ont la même valeur du nombre de connexion sur l'électrode 5 et les deux restant ont se caractérisent par une valeur commune, et pour l'électrode 21 on voit que 11 individus ont partagé une première valeur et les 9 autres ont partagé une autre. Cette particularité peut justifier les erreurs commises lors de la classification de ses individus, il peut s'agir d'individus dans un état intermédiaire entre deux groupes mais aussi à un fausse labelisation par manque d'information de la part des experts et ça peut être dû aussi à un manque d'information qui induit la machine en erreur. Il est donc nécessaire d'établir une qualité de la labélisation réalisée par les experts.

7.3 Classification par l'algorithme des cartes de Kohonen

La classification avec l'algorithme des K-moyennes a instauré un doute quant à la qualité de labelisation réalisée par les experts. Pour s'en assurer nous avons choisi d'utiliser les cartes topologiques ou auto-organisatrices qui font partie de la famille des modèles dite à «apprentissage non supervisé», c'est-à-dire qui s'appliquent sur des données dont on connaît le domaine sur lequel porte le recueil statistique, mais pour lesquelles les connaissances a priori ne sont pas totalement organisées et ceci dans un but de s'assurer des labelisation faites par les experts mais aussi de savoir s'il existe des groupes avec des états intermédiaires qui pourrait justifier entre autre les erreurs de diagnostic remarqué.
La classification de nos individus est réalisé en passant d'abord par une étape d'apprentissage, nous allons utilisé une structure de carte de grande dimensionnalité afin de réduire la sensibilité de la carte aux points aberrants. Ensuite, nous effectuons une affectation par vote majoritaire et nous réalisons une classification des neurones par une classification hiérarchique ascendante (CAH) et on utilise les résultats obtenus afin de labeliser les individus du jeu de données initial.
Après l'apprentissage d'une carte de taille 10x10 neurones, nous avons procédé à la classification de ces neurones par une CAH, nous avons obtenu le dendrogramme suivant:

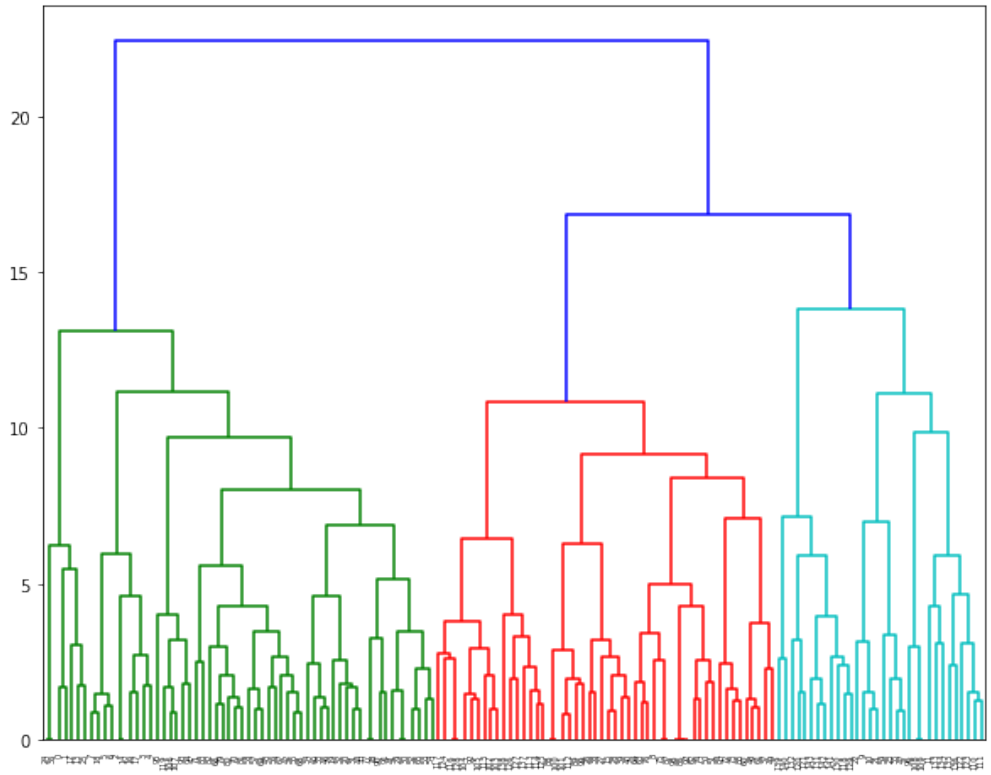


Figure 14: Dendrogramme de la classification des neurones

Ce dendrogramme suggère une classification en 3 classes, nous réalisons une classification en 3 classes des neurones et nous projetons le résultat sur les individus que nous représentons sur un espace à deux dimensions obtenu par l'algorithme t-SNE obtenu précédemment.
Il en résulte la représentation suivante :

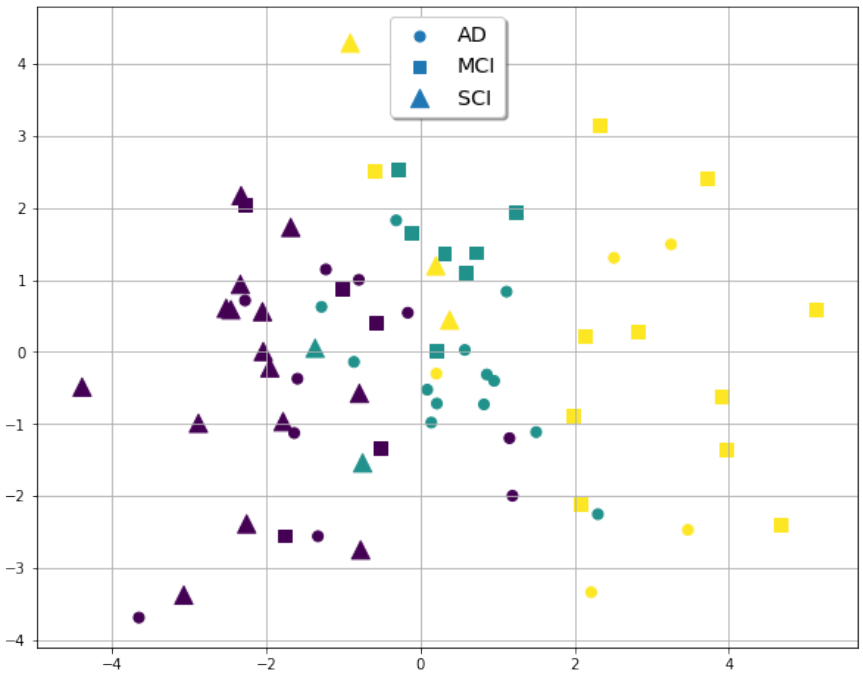
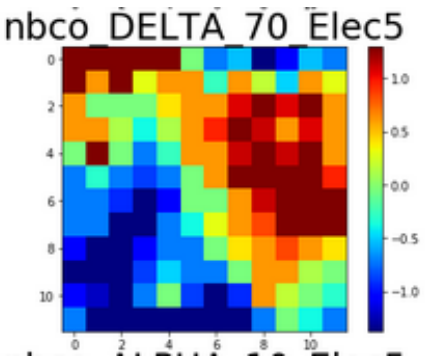
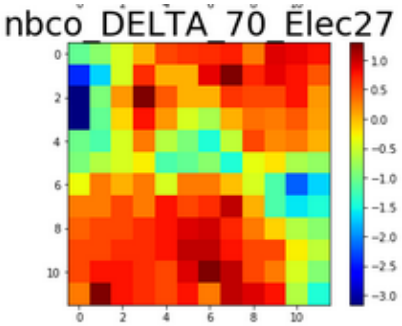


Figure 15: Représentation des cluster obtenus grâce aux cartes de Kohonen

Nous pouvons voir que les groupes montrent une meilleure séparation des individus selon leur labélisation faites par les experts néanmoins on remarque que certains individus restent mal classés. Cette remarque est surtout valable pour la classe de dominance **AD** où nous voyons plusieurs individus **MCI** et **SCI** qui sont mal classés. On peut alors conclure que les données que nous traitons contiennent des erreurs de labelisation ou des individus dans un état transitoire entre deux types. On peut aussi penser que l'information contenue dans notre jeu de données devrait être complété pour mieux caractériser la maladie d'Alzheimer.

La classification réalisée par les cartes de Kohonen a permis aussi de mettre la lumière sur certaines variables en particulier :





Les activations les plus fortes sont remarquées pour les variables concernant l'électrode 5 et l'électrode 27, ce sont les variables qui contribuent le plus à différencier les 3 types de patients. Nous avons vu que tout au long de l'étude l'électrode 27 (T6) permet de séparer les individus **AD** des **SCI**, elle a d'ailleurs été sélectionnée plusieurs fois pour des seuils et des bandes de fréquences différentes. De même, nous avons vu que les électrodes de la zone

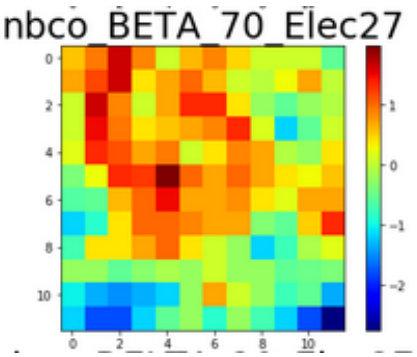


Figure 16: Activation des neurones par variables

frontale permettent de caractérisés les individus **AD** et de les séparer des **MCI**.

8 Conclusion

L'étude portant sur l'analyse de l'activité cérébrale via l'électroencéphalographie a permis de faire ressortir plusieurs spécificité concernant la pathologie de l'Alzheimer. Par exemple nous avons vu que la distinction entre un individu **SCI** et un individu **AD** est faite principalement selon les caractéristiques de l'électrode 27 (T6), la distinction entre les **MCI** et les **AD** est quant à elle faite au niveau des électrodes de la zone frontale.

Cette étude a mis en avant l'importance du choix du descripteur, comme nous l'avons vu la moyenne n'était pas adapté à la problématique elle ne permettait pas d'extraire une information assez explicative pour caractériser les 3 trois types de patients. Les résultats que nous avons obtenus sont faible avec ce descripteur mais le changement et l'orientation vers le nombre de connexion et de variance CDF permet une grande amélioration de ces résultats.

L'étude non supervisée a remis en cause les labelisations faites par les experts et a permis de faire ressortir des individus spéciaux qui était souvent sujet d'une fausse classification lors de l'étude supervisée. On pourrai suggérer que ces individus se trouvent dans un état intermédiaire ce qui justifierait la difficulté de classification mais nous pouvons aussi imaginer que nous ne disposons pas d'information suffisamment significatif pour les caractériser.

Finalement, l'étude que nous avons réalisée peut être largement améliorée et complétée en tenant compte aussi de la synchronisation et non seulement l'a-synchronisation par exemple. Mais ce que nous avons retenu c'est de toujours analyser les résultats obtenus par un algorithme et analyser là où celui-ci se trompe car un algorithme de classification supervisée s'efforce à extraire des caractéristiques qui permettent de justifier la labelisation même si celle-ci est fausse car il n'a aucune connaissance du métier et du cas d'étude traité. Il est donc important d'être particulièrement méticuleux afin d'éviter d'expliquer des résultats par des faits complètement incohérents.