UQAC
Université du Québec
à Chicoutimi

EXPLORATION DE L'APPRENTISSAGE AUTOMATIQUE POUR LA PRÉDICTION DU DEGRÉ D'ATTEINTE DE PERSONNES SOUFFRANT D'ARSACS À PARTIR DE SIGNAUX EMG

PAR IDIR BOUABID

MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI EN VUE DE L'OBTENTION DU GRADE DE MAÎTRE ÈS SCIENCES (M. SC.) EN INFORMATIQUE

QUÉBEC, CANADA

RÉSUMÉ

Ce mémoire explore l'utilisation de l'apprentissage automatique pour prédire le degré d'atteinte des personnes atteintes d'Ataxie Récessive Spastique Autosomique de Charlevoix-Saguenay (ARSACS) à partir de signaux électromyographiques (EMG). L'ARSACS, une maladie neurodégénérative rare, se manifeste par une dénervation sévère des muscles, détectable grâce à l'EMG.

L'objectif principal de cette recherche est d'identifier des marqueurs prédictifs permettant de suivre le degré d'atteinte des patients atteints d'ARSACS et de prévoir la progression de la maladie sur plusieurs années, contribuant ainsi à une meilleure compréhension de cette pathologie. Pour atteindre cet objectif, des approches de *clustering* et de régression ont été appliquées aux données EMG et cliniques, dans le but de tenter de découvrir des groupements inédits de profils de patients présentant des similitudes, et de prédire et estimer la progression de la maladie à des instants futurs.

Les résultats montrent que le *clustering* permet d'identifier des sous-groupes cliniquement significatifs, bien que la variabilité au sein des groupes puisse limiter la discrimination des degrés d'atteinte, reflétant ainsi la complexité de la maladie.

Les modèles de régression utilisés pour prédire la progression de la maladie ont rencontré des difficultés à capturer efficacement les interactions complexes entre les variables cliniques et les données EMG. La variabilité importante des performances des modèles et le risque élevé de surapprentissage mettent en évidence leur manque de robustesse. Ces résultats sont principalement dus à la taille réduite de l'échantillon de données et à la forte hétérogénéité des patients, qui limitent considérablement la généralisation des modèles.

Malgré ces limites, cette recherche représente une avancée importante dans l'application de l'apprentissage automatique à l'ARSACS et ouvre la voie à des études futures en intégrant des approches plus sophistiquées et un volume de données plus important pour mieux comprendre la progression de la maladie.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

ARSACS Ataxie Récessive Spastique Autosomique de Charlevoix-Saguenay
GRIMN Groupe de Recherche Interdisciplinaire sur les Maladies Neuromusculaires
CMNM Clinique des Maladies Neuromusculaires
POP Procédure Opérationnelle Permanente
LIARA Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités
EMG Électromyographie
sEMG Électromyographie de Surface
RMS *Root Mean Square*
PCA Analyse en Composantes Principales
MAE Erreur Absolue Moyenne
R^2 Coefficient de Détermination
SVM Machine à Vecteurs de Support
k-NN *k-Nearest Neighbors*
CNN Réseau de Neurones Convolutifs
t-SNE *t-distributed Stochastic Neighbor Embedding*
LEMOCOT Test de Coordination Motrice des Membres Inférieurs
UPDRS *Unified Parkinson's Disease Rating Scale*
CA Ataxie Cérébelleuse
DBS Stimulation Cérébrale Profonde

REMERCIEMENTS

Tout au long de ce projet, j'ai eu la chance de pouvoir compter sur l'aide, la bienveillance et le soutien de personnes dont l'influence a été cruciale dans ce parcours de recherche. Je tiens à leur exprimer ici toute ma reconnaissance.

Tout d'abord, je souhaite remercier chaleureusement mon directeur, Sébastien Gaboury, et mon codirecteur, Julien Maître, pour leur encadrement, leurs conseils avisés et leur accompagnement constant tout au long de cette recherche. Leur expertise et leur patience ont été une source inestimable de motivation et de progrès.

Je tiens également à exprimer ma reconnaissance envers Florentin Thullier, dont l'aide précieuse et les suggestions constructives ont largement contribué à l'avancement de ce travail.

Ma recherche a été menée au sein du Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités, que je remercie pour m'avoir accueilli et pour m'avoir fourni les ressources nécessaires à l'accomplissement de ce projet. Je suis également reconnaissant envers le Groupe de Recherche Interdisciplinaire sur les Maladies Neuromusculaires pour sa collaboration, qui a été essentielle à la réussite de cette étude.

Un remerciement particulier est adressé aux participants atteints de l'Ataxie Récessive Spastique Autosomique de Charlevoix-Saguenay, dont la contribution précieuse a permis d'approfondir notre compréhension de cette maladie. Leur participation a été essentielle à cette recherche, et je suis profondément reconnaissant de leur engagement.

Enfin, je ne saurais terminer ces remerciements sans adresser mes plus sincères remerciements à ma famille et mes amis, dont le soutien moral et les encouragements ont été essentiels. Ma sœur, Thanina, mérite un remerciement spécial pour sa patience et sa bienveillance, qui m'ont été d'un grand réconfort et ont souvent renouvelé ma motivation.

CHAPITRE I

INTRODUCTION

1.1 CONTEXTE

L'Ataxie Récessive Spastique Autosomique de Charlevoix-Saguenay (ARSACS) est une maladie neurodégénérative, découverte pour la première fois dans les régions de Charlevoix et du Saguenay-Lac-Saint-Jean au Québec, Canada, qui est due à l'effet fondateur [2]. Elle est causée par des mutations dans le gène SACS [3], codant pour la protéine sacsin, essentielle au bon fonctionnement des cellules nerveuses.

Les patients atteints d'ARSACS présentent généralement des symptômes dès l'enfance, incluant une difficulté lors de l'initiation de la marche, une ataxie spastique des membres inférieurs et supérieurs, une dysarthrie ainsi qu'une amyotrophie distale [2]. L'ARSACS se manifeste également par des signes de dénervation sévère dans les muscles distaux observés par Électromyographie (EMG) [4], dus à une dégénérescence axonale dans les systèmes nerveux central et périphérique [5]. Un des principaux défis posés par l'ARSACS réside dans la grande variabilité interindividuelle, tant au niveau de la présentation clinique que de la gravité des symptômes et de l'évolution de la maladie [6].

Plusieurs études ont exploré l'ARSACS à travers l'analyse des signaux électromyographiques, notamment celle de Bouchard et al. [4], qui examine l'électromyographie et la conduction nerveuse chez des patients atteints d'ataxies, y compris l'ARSACS. Les résultats de cette étude montre que les patients atteints d'ARSACS présentent des signes plus prononcés de dénervation et une vitesse de conduction motrice plus lente que les autres formes d'ataxie, ce qui souligne les particularités électrophysiologiques de cette maladie [4]. Ces observations soulignent l'importance des données EMG dans la compréhension et l'évaluation de la progres-

sion de l'ARSACS. En effet, les signaux EMG sont particulièrement riches en informations, car ils enregistrent en temps réel l'activité électrique de plusieurs muscles [7], souvent lors de différents exercices physiques. Ces données permettent de capturer les interactions complexes entre le système nerveux et les muscles [7], dans le cas de cette étude, affectés par l'ARSACS. En outre, pour chaque patient, plusieurs tests sont généralement effectués aux mêmes et à différents moments, générant ainsi des ensembles de données volumineux et variés. Cependant, malgré la richesse des informations contenues dans les signaux EMG, leur analyse manuelle reste complexe en raison de la grande variabilité des réponses musculaires et du volume conséquent de données générées. C'est dans ce contexte que l'apprentissage automatique apparaît comme une approche prometteuse.

L'apprentissage automatique est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données et de prendre des décisions sans nécessiter une programmation explicite pour chaque tâche spécifique [8]. L'apprentissage automatique utilise des algorithmes pour traiter et analyser de vastes ensembles de données, permettant ainsi de découvrir des motifs et des tendances complexes. Cette capacité à extraire des informations pertinentes à partir de données hétérogènes et volumineuses en fait un atout puissant pour de nombreuses applications, notamment dans le domaine de la santé.

L'utilisation de l'apprentissage automatique dans le domaine de la santé, en particulier pour les maladies neurodégénératives, pourrait aider à comprendre l'évolution de ces maladies ainsi qu'à améliorer les approches diagnostiques et thérapeutiques [9]. L'application de l'apprentissage automatique aux données cliniques et longitudinales collectées sur les patients peut également identifier des sous-groupes de patients aux caractéristiques similaires, offrant ainsi une meilleure compréhension de l'évolution des maladies [9].

1.2 PROBLÉMATIQUE ET OBJECTIFS

Bien que l'apprentissage automatique ait démontré son potentiel pour améliorer la compréhension de l'évolution des maladies neurodégénératives, son application à l'ARSACS demeure limitée. La variabilité interindividuelle dans la présentation clinique de cette maladie neurodégénérative, ainsi que l'absence d'un grand volume de données sur cette pathologie spécifique, représentent des défis majeurs pour l'élaboration de modèles prédictifs robustes. À notre connaissance, aucune recherche n'a encore exploré l'utilisation de l'apprentissage automatique, et en particulier l'analyse des signaux électromyographiques, pour prédire la progression de l'ARSACS ainsi que le degré d'atteinte des personnes affectées.

Cependant, des études récentes sur d'autres maladies neurodégénératives ont montré des résultats prometteurs grâce à l'analyse des données cliniques et des signaux électromyographiques, suggérant que des approches similaires pourraient être appliquées à l'ARSACS. Cette recherche vise donc à combler cette lacune en explorant l'utilisation des algorithmes d'apprentissage automatique pour prédire le degré d'atteinte des patients souffrant d'ARSACS et la progression de la maladie.

Les objectifs de cette recherche sont d'abord de découvrir des sous-groupes de patients atteints d'ARSACS en utilisant des approches de *clustering*, afin de mieux comprendre la variabilité de la maladie et d'identifier des profils de patients présentant des similitudes. Ensuite, des modèles de régression sont développés pour prédire et estimer la progression de la maladie. Ces modèles incluent des prédictions ponctuelles de scores cliniques ainsi que des modèles dynamiques intégrant des données longitudinales, dans le but d'anticiper l'évolution de la maladie. Enfin, l'objectif est d'identifier des marqueurs prédictifs permettant de suivre le degré d'atteinte des patients et de prévoir l'évolution de la maladie sur plusieurs années,

contribuant ainsi à une meilleure compréhension de l'ARSACS et à un suivi plus précis de sa progression.

1.3 CONTRIBUTION DE LA RECHERCHE

Cette recherche constitue une première dans plusieurs aspects de l'étude de l'ARSACS, notamment par l'application de techniques d'apprentissage automatique telles que le *clustering* et la régression à l'analyse des données cliniques et des signaux électromyographiques pour cette maladie neurodégénérative. À ce jour, aucune étude n'a exploré l'utilisation des signaux EMG en conjonction avec des modèles d'apprentissage automatique pour prédire l'évolution et le degré d'atteinte de personnes souffrant d'ARSACS.

L'une des contributions majeures de cette recherche est l'utilisation du *clustering* pour tenter de découvrir des groupements inédits de profils de patients atteints d'ARSACS, ce qui pourrait permettre une meilleure compréhension des différents sous-groupes de la maladie. En parallèle, des modèles de régression sont utilisés pour tenter de prédire et d'estimer la progression de l'ARSACS chez les patients, à la fois par des prédictions ponctuelles de scores cliniques et par des modèles dynamiques intégrant des données longitudinales pour anticiper l'évolution de la maladie à des instants futurs. Ces approches visent à identifier des marqueurs prédictifs du degré d'atteinte des personnes souffrant d'ARSACS, ainsi que d'essayer de suivre l'évolution de la maladie sur plusieurs années.

1.4 MÉTHODOLOGIE DE LA RECHERCHE

La méthodologie adoptée dans cette recherche se décompose en plusieurs phases, allant de la récolte de données à l'analyse par des modèles d'apprentissage automatique, en passant par leur traitement et leur préparation pour les expérimentations.

1.4.1 RÉCOLTE DES DONNÉES

Les données utilisées dans cette recherche ont été recueillies via des capteurs d'électromyographie placés sur des participants atteints d'ARSACS et des individus sains. Ces capteurs ont enregistré l'activité musculaire des membres inférieurs lors de la réalisation de plusieurs exercices, tels que l'extension et la flexion du genou. La collecte des données a été réalisée avec l'assistance de cliniciens dans un cadre contrôlé, en suivant les recommandations du SENIAM [10] pour le placement des électrodes, assurant ainsi la qualité et la fiabilité des enregistrements.

1.4.2 TRAITEMENT DES DONNÉES

Les données EMG recueillies ont été traitées à l'aide d'une application développée au sein du LIARA, principalement pour analyser les coefficients de co-contraction, qui mesurent l'activation simultanée de muscles agonistes et antagonistes. Le traitement a commencé par la conversion des fichiers de données dans un format exploitable, suivie d'un suréchantillonnage des signaux EMG pour assurer leur alignement. Ensuite, les données ont été filtrées pour éliminer les bruits indésirables, puis normalisées et préparées pour l'analyse. Ces étapes permettent de garantir une qualité optimale des données pour les phases prédictives et exploratoires.

1.4.3 PRÉPARATION DES DONNÉES

Dans cette étude, les participants ont réalisé plusieurs répétitions des exercices d'extension et de flexion du genou. Pour chaque participant, les trois meilleures itérations ont été sélectionnées à chaque temps de mesure distinct, constituant ainsi trois instances par temps, bien que certains participants n'aient pas été présents à tous les moments de mesure. Les données recueillies incluent des informations électromyographiques et cliniques. Les données

cliniques comprennent des informations sociodémographiques (comme l'âge et le niveau de mobilité des participants) ainsi que des mesures d'évaluation standardisées, telles que les scores d'équilibre, de coordination motrice et de gravité de la maladie.

La préparation des ensembles de données a consisté à fusionner les données EMG et cliniques pour créer des ensembles contenant des informations pertinentes pour chaque expérimentation. Chaque instance inclut les signaux EMG de plusieurs muscles ainsi que les données cliniques associées. Une fois fusionnées, les données ont été ajustées pour chaque expérimentation en sélectionnant les variables les plus pertinentes. Des techniques d'ingénierie des caractéristiques ont été utilisées pour ajouter des variables supplémentaires, telles que des mesures anticipées, afin d'explorer les relations temporelles. Enfin, des étapes de gestion des données manquantes et de normalisation ont été appliquées, garantissant des ensembles de données optimisés et cohérents pour les analyses ultérieures.

1.5 ORGANISATION DU DOCUMENT

Ce mémoire est structuré en six chapitres qui suivent une progression logique pour atteindre les objectifs de ce projet de recherche. Le Chapitre 1, qui constitue l'introduction, présente le contexte général, la problématique abordée et les objectifs spécifiques, ainsi que la contribution et la méthodologie de la recherche. L'organisation du reste du mémoire est détaillée dans cette section.

Le Chapitre 2, consacré à l'état de l'art, s'ouvre par une section dédiée à l'EMG, qui explore les différences entre l'EMG à aiguille et l'Électromyographie de Surface (sEMG), les caractéristiques des signaux sEMG, ainsi que l'importance de l'sEMG dans l'étude de l'ARSACS. Cette section se conclut par une sous-section sur l'apport de l'apprentissage automatique impliquant des données sEMG. Par la suite, le chapitre passe en revue les travaux

de recherche existants relatifs à l'utilisation des techniques d'apprentissage automatique pour les maladies neurodégénératives, permettant de tirer des leçons des techniques et résultats observés dans d'autres contextes similaires, qui peuvent être adaptés à notre étude, notamment celle de l'ARSACS.

Dans le Chapitre 3, intitulé traitement des données EMG, l'application MyoRatio développée par le LIARA est d'abord présentée dans la première section. Cette dernière commence par une description détaillée du processus de collecte des données, en abordant la méthodologie et le système d'acquisition utilisés. Ensuite, les types de données enregistrées au cours des différentes phases du projet de recherche sont exposés. Le processus de traitement de ces données, essentiel pour réaliser des analyses précises, est ensuite expliqué en détail. Enfin, les différents types de fichiers générés par l'application sont présentés, notamment les métadonnées, les fichiers utilisateur et les rapports d'analyse. La deuxième section de ce chapitre présente une nouvelle fonctionnalité ajoutée à l'application MyoRatio, permettant de générer un nouveau type de fichier à partir des fichiers d'analyse, appelé fichier de synthèse. Pour finir, la troisième et dernière section de ce chapitre porte sur la préparation des données en vue de la construction des ensembles de données à partir des fichiers d'analyse. Ces ensembles de données sont nécessaires pour expérimenter différentes approches d'apprentissage automatique, qui seront présentées dans les Chapitres 4 et 5.

Les Chapitres 4 et 5 présentent les différentes expérimentations menées ainsi que les résultats obtenus, respectivement pour le *clustering* et la régression. Les différentes étapes de chaque expérimentation sont détaillées et justifiées. De même, les résultats obtenus pour chaque expérimentation sont présentés et interprétés.

Enfin, une conclusion incluant une discussion générale est présentée dans le dernier chapitre. Les résultats obtenus y sont interprétés, justifiés et discutés en profondeur, mentionnant leurs implications, les limitations rencontrées ainsi que les pistes pour les recherches futures.

CHAPITRE II

ÉTAT DE L'ART

Ce chapitre est divisé en deux sections, la première section explore l'Électromyographie (EMG), en distinguant l'EMG à aiguille de l'Électromyographie de Surface (sEMG), ainsi que les caractéristiques des signaux sEMG et leur importance pour l'étude de l'ARSACS. Une sous-section aborde ensuite l'apport de l'apprentissage automatique dans l'analyse des données sEMG. La deuxième section de ce chapitre passe en revue les travaux de recherche existants relatifs à l'utilisation des techniques d'apprentissage automatique pour les maladies neurodégénératives.

2.1 ELECTROMYOGRAPHIE

L'électromyographie est une technique utilisée pour analyser et enregistrer l'activité électrique générée par les muscles squelettiques. Cette méthode mesure la réaction des muscles aux signaux nerveux, offrant ainsi des données essentielles sur l'état du système nerveux et musculaire [7]. On distingue principalement deux types d'EMG : l'électromyographie à aiguille, qui est invasive, et l'électromyographie de surface, qui est non invasive.

2.1.1 ELECTROMYOGRAPHIE VS ELECTROMYOGRAPHIE DE SURFACE

L'EMG à aiguille consiste à insérer une fine aiguille directement dans le muscle pour mesurer l'activité électrique au niveau intracellulaire. Ce procédé permet d'obtenir des données très précises et localisées, mais son caractère invasif le rend moins adapté pour des études longitudinales ou impliquant des mesures répétées [11]. L'sEMG, quant à elle, utilise des électrodes placées sur la peau pour enregistrer les signaux musculaires sous-jacents. Les

premières électrodes de surface étaient de simples plaques métalliques recouvertes d'une fine couche de gel électrolytique, fixées à la peau à l'aide de ruban adhésif. Bien que cette conception soit susceptible de générer des artefacts de mouvement dus aux perturbations mécaniques, elle reste efficace si bien appliquée [12]. De nos jours, la conception des électrodes flottantes permet de minimiser ces artefacts en éliminant le contact direct entre la surface métallique et la peau, tout en maintenant une liaison conductrice via un pont électrolytique [13]. Cette approche est non invasive et plus pratique pour l'enregistrement simultané de plusieurs muscles, bien qu'elle soit moins précise que l'EMG à aiguille en termes de localisation [14]. L'sEMG est couramment utilisée pour analyser les mouvements musculaires globaux et est particulièrement adaptée aux études impliquant des cohortes de patients, comme celles sur les maladies neurodégénératives [15].

La Figure 2.1 illustre cette comparaison entre l'sEMG et l'EMG à aiguille. Dans (a), les électrodes de surface sont montrées, tandis que (b) et (c) représentent respectivement des configurations d'électrodes à aiguille concentrées et bipolaires, montrant la précision accrue mais l'invasivité de cette méthode.

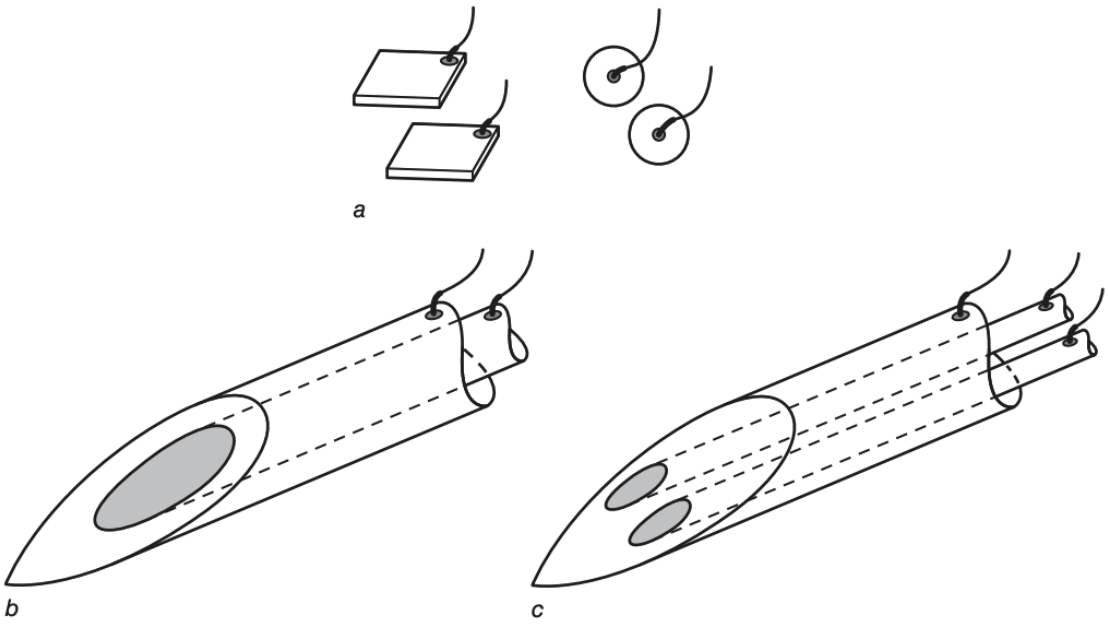

FIGURE 2.1 : Comparaison entre l'électromyographie de surface et l'électromyographie à aiguille : Configurations des électrodes (tirée de [1]).

Il est important de mentionner qu'à partir du Chapitre 3 et dans le reste du mémoire, l'abréviation EMG sera utilisée exclusivement pour désigner l'électromyographie de surface, afin d'alléger et de simplifier le texte.

2.1.2 CARACTÉRISTIQUES DES SIGNAUX EMG DE SURFACE

Les signaux sEMG reflètent l'activité électrique générée par les muscles pendant leur contraction. Ces signaux sont influencés par divers facteurs tels que l'épaisseur des tissus entre les électrodes et les muscles, la position des électrodes, ainsi que l'état de fatigue musculaire [16]. L'amplitude du signal EMG est une caractéristique fondamentale, couramment utilisée pour évaluer l'intensité de l'activation musculaire. Elle est directement liée à la force de contraction musculaire et permet d'observer l'effort fourni par le muscle à un instant donné. Toutefois, cette mesure peut être affectée par des variations liées à la position des électrodes ou

aux changements d'état du muscle, rendant son interprétation parfois plus complexe [16]. Pour obtenir une estimation plus stable et représentative de l'ensemble de l'activation musculaire, le *Root Mean Square* (RMS), également appelé la valeur quadratique moyenne, est souvent utilisé en complément de l'amplitude. Le RMS représente la moyenne quadratique des valeurs du signal sur une période donnée et fournit une estimation fiable de la puissance musculaire globale. Contrairement à l'amplitude instantanée, le RMS prend en compte l'ensemble des variations du signal, offrant ainsi une vue d'ensemble de l'activité musculaire [15]. Une augmentation du RMS reflète une activation musculaire plus importante, tandis qu'une diminution peut indiquer une fatigue progressive. Les caractéristiques fréquentielles, telles que la fréquence médiane et la fréquence moyenne, sont également utilisées pour évaluer la fatigue musculaire, car les fréquences du signal tendent à diminuer lorsque le muscle se fatigue [14]. De plus, l'analyse du spectre de puissance permet de décomposer le signal EMG en différentes composantes fréquentielles, fournissant des informations sur les types de fibres musculaires activées lors des contractions [7]. Enfin, des caractéristiques temporelles telles que le temps de montée, correspondant au temps nécessaire pour qu'un muscle atteigne son niveau maximal d'activation après le début de la contraction, et le temps de relaxation, représentant la durée nécessaire pour qu'un muscle retourne à son état de repos après une contraction, peuvent être utilisées pour évaluer la dynamique de l'activation musculaire, ce qui est particulièrement pertinent dans les études portant sur la capacité motrice [16].

2.1.3 IMPORTANCE DE L'ELECTROMYOGRAPHIE DE SURFACE DANS L'ÉTUDE DE L'ARSACS

L'Ataxie Récessive Spastique Autosomique de Charlevoix-Saguenay est une maladie neurodégénérative qui se caractérise par des signes de dénervation sévère des muscles distaux, observés par EMG [4], conséquence d'une dégénérescence axonale affectant à la fois les

systèmes nerveux central et périphérique [5]. L'EMG permet de mesurer avec précision la dénervation et les anomalies motrices, fournissant ainsi des données essentielles sur l'évolution de la maladie et le degré de dysfonctionnement musculaire. Cependant, il est important de noter que très peu d'études se sont concentrées sur l'utilisation de l'sEMG pour l'analyse des patients atteints d'ARSACS, malgré son potentiel à révéler des informations clés sur la dynamique musculaire et la progression de la maladie.

Une des rares études dans ce domaine a été menée par Richards et al. [17] auprès de 17 patients, dont 8 atteints d'ARSACS et le reste d'ataxie de Friedreich, afin d'évaluer la dynamique musculaire fonctionnelle. Les chercheurs ont mesuré le couple musculaire au niveau du genou lors de mouvements isocinétiques volontaires et enregistré l'activité sEMG de cinq muscles des membres inférieurs. Une réduction de la force dynamique et des anomalies dans les patrons sEMG ont été observées. La coactivation musculaire, mesurée en comparant l'activité sEMG des muscles antagonistes et agonistes lors des contractions isocinétiques, a montré une augmentation significative chez les deux groupes de patients, fournissant des données essentielles pour l'évaluation thérapeutique future. Bien que cette étude apporte des informations précieuses, elle met en évidence la rareté des recherches utilisant l'sEMG pour comprendre les particularités de l'ARSACS, suggérant ainsi un besoin d'explorations plus approfondies dans ce domaine.

2.1.4 APPORT DE L'APPRENTISSAGE AUTOMATIQUE IMPLIQUANT DES DONNÉES D'ÉLECTROMYOGRAPHIE DE SURFACE

Les signaux EMG représentent des données temporelles riches en informations, mais complexes à analyser manuellement en raison de leur variabilité et du volume important de données générées. L'utilisation de l'apprentissage automatique permet d'analyser ces données de manière plus efficace, en détectant des motifs subtils et des corrélations entre les

signaux EMG et l'évolution clinique des patients. Par exemple, dans une étude portant sur la reconnaissance des *patterns* d'activation musculaire des épaules, un Réseau de Neurones Convolutifs (CNN) a été utilisé pour traiter des signaux sEMG provenant de 12 muscles afin de reconnaître différents mouvements du bras supérieur, tels que l'abduction ou les mouvements avant/arrière. Le modèle a atteint une justesse allant jusqu'à 97,57 % pour les mouvements à vitesse normale et 97,07 % pour les mouvements rapides, démontrant l'efficacité des CNN dans le traitement des signaux sEMG pour des applications de rééducation robotisée [18].

Dans une autre étude, des algorithmes d'apprentissage automatique comme les Machine à Vecteurs de Support (SVM) et *k-Nearest Neighbors* (k-NN) ont été employés pour classifier 20 actions physiques quotidiennes à partir des caractéristiques temporelles, fréquentielles et statistiques inter-canaux des signaux sEMG. Cette approche a atteint une justesse de 100 % pour la classification de 10 actions normales avec le SVM, et 98,91 % pour la classification de 10 actions agressives avec k-NN. Une combinaison hybride de ces deux méthodes a permis d'atteindre une justesse globale de 98,97 % pour la classification des 20 actions, montrant ainsi la polyvalence des méthodes d'apprentissage automatique dans l'analyse des signaux sEMG [19].

De manière plus générale, l'utilisation de l'apprentissage automatique dans le domaine de la santé, notamment pour les maladies neurodégénératives, peut améliorer la compréhension de l'évolution des maladies en exploitant des données cliniques et longitudinales [9]. Ces approches permettent non seulement d'automatiser l'analyse de grands volumes de données, mais aussi de révéler des motifs qui ne seraient pas perceptibles par une analyse conventionnelle.

2.2 TRAVAUX CONNEXES

Bien que l'ARSACS soit une maladie neurodégénérative bien caractérisée, il existe une lacune notable dans la littérature scientifique concernant l'utilisation de l'apprentissage automatique pour prédire le degré d'atteinte des personnes atteintes d'ARSACS et la progression de la maladie. En l'absence d'articles spécifiques sur l'ARSACS, les critères de sélection ont été étendus afin d'inclure les recherches portant sur l'utilisation de l'apprentissage automatique pour les maladies neurodégénératives, en donnant la priorité à celles utilisant des données EMG. Cette approche permet de tirer des leçons des techniques et résultats obtenus dans d'autres contextes similaires, qui peuvent être adaptés à l'étude de l'ARSACS.

L'étude de Kleinholdermann et al. [20] a exploré l'utilisation de l'sEMG pour mesurer de manière objective les symptômes moteurs chez les patients atteints de la maladie de *Parkinson*. Dans cette recherche, des données ont été recueillies auprès de 45 patients atteints de *Parkinson* en utilisant des électrodes sEMG fixées à un bracelet, tandis que les patients exécutaient une tâche de tapotement, avec et sans médication dopaminergique. L'objectif était de prédire les scores de l'échelle *Unified Parkinson's Disease Rating Scale* (UPDRS) à partir des caractéristiques sEMG en utilisant divers modèles de régression et techniques d'apprentissage automatique. Parmi les modèles testés, la régression par forêts aléatoires s'est révélée la plus performante, avec une corrélation de 0,739 entre les valeurs UPDRS réelles et prédites. Ces résultats démontrent que les données sEMG peuvent être utilisées pour prédire l'affection motrice chez les patients atteints de *Parkinson*, et suggèrent que de tels enregistrements pourraient aider à ajuster les traitements médicamenteux. Cette étude présente un avantage significatif en démontrant l'efficacité des techniques d'apprentissage automatique appliquées aux données sEMG pour prédire des mesures cliniques importantes [20].

Dans une autre étude, Ferreira et al. [21] ont examiné les données des paramètres spatio-temporels de la marche de 63 personnes atteintes de *Parkinson* et de 63 individus du groupe témoin, en utilisant des algorithmes d'apprentissage automatique pour distinguer les personnes atteintes de *Parkinson* des individus sains et pour discriminer les stades de la maladie de *Parkinson*. Pour le diagnostic de la *Parkinson*, l'algorithme *Naïve Bayes* a atteint une justesse de 84,6%, avec une précision de 0,923 et un rappel de 0,800. En ce qui concerne l'identification des stades de la maladie de *Parkinson*, l'algorithme *Random Forest* a surpassé les autres algorithmes étudiés, atteignant une aire sous la courbe ROC de 0,786. Les résultats obtenus démontrent le potentiel des algorithmes d'apprentissage automatique pour diagnostiquer et identifier les stades de la maladie de *Parkinson* à travers l'analyse des paramètres de marche [21].

La recherche de Oliveira et al. [22] explore l'application du *t-distributed Stochastic Neighbor Embedding* (t-SNE), un algorithme de réduction non linéaire de la dimension et la visualisation des données, pour distinguer les individus neurologiquement sains de ceux atteints de la maladie de *Parkinson* traités par lévodopa et Stimulation Cérébrale Profonde (DBS). Les participants ont exécuté une séquence de quatre tâches motrices afin de recueillir des données inertielle et électromyographique. L'étude compare les performances de classification d'une SVM en discriminant des ensembles de caractéristiques bidimensionnelles estimées à partir du PCA, du mappage de Sammon et du t-SNE. Les meilleurs résultats de classification ont été obtenus avec t-SNE, atteignant 96,9% pour l'ensemble d'entraînement et 76,6% pour l'ensemble de test. Les résultats de dipersion des individus via t-SNE pourrait permettre de mesurer la divergence entre les comportements moteurs normaux et anormaux, permettant ainsi la personnalisation et l'ajustement des traitements [22].

L'Ataxie Cérébelleuse (CA) regroupe des maladies affectant le cervelet, responsable de la coordination des mouvements. Cette condition provoque des mouvements désordonnés et

peut également impacter l'équilibre, la parole et les mouvements oculaires [23]. Abeysekara et al. [24] ont examiné les mesures cinématiques des mouvements des membres supérieurs pendant un test de suivi balistique, se concentrant principalement sur les mouvements de l'articulation de l'épaule. L'objectif de cette étude était de comprendre les défis liés à l'identification et à l'évaluation de la sévérité de la CA à travers ces mouvements. Pour développer des modèles d'apprentissage automatique pour la classification et la régression, les chercheurs ont utilisé des caractéristiques statistiques des signaux cinématiques. Pour la classification, le modèle de *Gradient Boosting Classifier* a atteint une justesse de 74%. En revanche, les modèles de régression ont présenté une spécificité limitée et des performances plutôt médiocres [24].

L'étude de Purnawan et al. [25] explore la classification des maladies neuromusculaires, en particulier la maladie de *Parkinson*, pour laquelle des signaux sEMG ont été recueillis. Dans cette étude, quatre stades typiques de la maladie de *Parkinson* ont été inclus : sain, possible, probable et certain, avec 10 participants pour chaque stade. Les signaux sEMG de chaque participant ont été traités en utilisant 17 caractéristiques temporelles, puis réduits en quatre composantes principales à l'aide de PCA. Les paires de ces nouvelles composantes principales ont été utilisées pour l'entraînement et les tests à l'aide d'une SVM. Les résultats obtenus ont montré que la justesse de la classification a été améliorée grâce à l'utilisation de PCA [25].

Un autre travail notable est celui de Anselmino et al. [26], qui a exploré l'utilisation de dispositifs portables pour améliorer la mobilité des personnes souffrant de troubles tels que les amputations et les maladies neurodégénératives. Cette recherche propose un contrôleur avancé basé sur des SVM utilisant des signaux sEMG des muscles de la cuisse pour détecter l'initiation et la direction des pas. En testant ce système sur dix sujets sains, marchant dans quatre directions, dans trois contextes différents et dans diverses conditions, les chercheurs ont atteint une justesse médiane de 83.34 % pour la détection de l'initiation du pas et de 73.92%

jusqu'à 92.91 % pour la direction du pas. Ces résultats suggèrent que ce système pourrait améliorer la réactivité et la liberté de mouvement des dispositifs portables, tout en étant facile à intégrer dans des dispositifs compacts [26].

Enfin, l'étude de Castelli Gattinara Di Zubiena et al. [27] qui a utilisé la posturographie dynamique avec des capteurs portables pour détecter les anomalies de l'équilibre chez les patients atteints de la maladie de *Parkinson*. Les chercheurs ont appliqué des techniques d'apprentissage automatique afin de distinguer les patients atteints de la maladie de *Parkinson* des sujets sains, ainsi que de distinguer les patients sous traitement dopaminergique et ceux qui ne le sont pas, et ce, en testant 52 classificateurs issus de plusieurs algorithmes (arbres de décision, *K-Nearest Neighbor*, machines à vecteurs de support, réseaux de neurones artificiels) pour analyser les réponses posturales enregistrées. Ils ont utilisé des critères de sélection basés sur la justesse, le rappel et la précision ainsi qu'un critère d'évaluation, et ont identifié le *Fine K-Nearest Neighbor* comme le plus efficace pour distinguer les patients atteints de la maladie de *Parkinson* des sujets sains. Cependant, pour la comparaison entre les patients sous traitement dopaminergique et ceux qui ne le sont pas, aucun des des classificateurs testés n'a satisfait les critères de sélection [27].

CHAPITRE III

TRAITEMENT DES DONNÉES EMG

3.1 INTRODUCTION

Dans ce chapitre, nous présentons tout d'abord l'application **MyoRatio** [1] déjà développée par le LIARA. Ensuite, nous parlons des nouvelles fonctionnalités ajoutées à cette application. Enfin, nous expliquons notre processus de collecte et de préparation d'ensembles de données, qui est une étape cruciale dans notre projet de recherche.

3.2 APPLICATION MYORATIO

Cette section présente l'application **MyoRatio**, développée par le LIARA, en mettant en avant les différentes phases de son développement et de son fonctionnement. Cela inclut notamment la récolte de données, les types de données récoltées, le traitement de ces données, ainsi que les types et structures des fichiers générés à partir de ces données.

3.2.1 RÉCOLTE DES DONNÉES

Cette phase, essentielle pour le développement de l'application, se compose de deux parties. La première partie, intitulée *méthodologie*, décrit les exercices réalisés par les participants pour la collecte de données, tout en mettant en évidence les principes directeurs et les critères qui ont orienté ce processus. La seconde partie, intitulée *système d'acquisition*, se focalise sur les aspects techniques et les outils utilisés pour la collecte et l'enregistrement des données.

---

1. https ://github.com/FlorentinTh/MyoRatio

MÉTHODOLOGIE

La récolte des données a été réalisée en faisant intervenir une soixantaine de participants. Parmi ceux-ci, on peut notamment retrouver différentes catégories :

- Les participants **sains (S)** : ce sont des participants qui ne sont pas atteints d'une maladie neuromusculaire.
- Les participants **marcheurs (M)** : ce sont les participants dont le degré d'atteinte de la maladie est moindre, ils sont capables de marcher, parfois avec une marchette ou une canne.
- Les participants **non marcheurs (N)** : ce sont les participants dont le degré d'atteinte de la maladie est plus important et qui sont, pour la plupart, en fauteuil roulant.

Dans le cas de cette étude, il a été demandé aux participants de réaliser trois exercices lors de séances accompagnées de cliniciens. Cependant, seuls deux de ces trois exercices seront décrits dans la suite de ce mémoire, car le dernier exercice ne fait pas l'objet de cette étude. Les deux exercices pris en compte lors des expérimentations sont les suivants :

1. **L'extension :** lors de cet exercice, les patients ont été installés assis sur un siège avec les deux jambes dans le vide, tel que représenté dans le premier encart de la Figure 3.1. Ensuite, il a été demandé aux participants de tendre la jambe pour que cette dernière arrive dans la position la plus tendue possible. Les participants devaient alors maintenir leur jambe tendue pendant au moins 3 secondes dans la mesure du possible selon leur condition puis revenir en position initiale, tels qu'illustrés par le second et le dernier encart de la Figure 3.1. Le mouvement de transition entre la position initiale et le maintien constitue la phase concentrique du mouvement tandis que le retour à la position initiale depuis le maintien est la phase excentrique du mouvement.

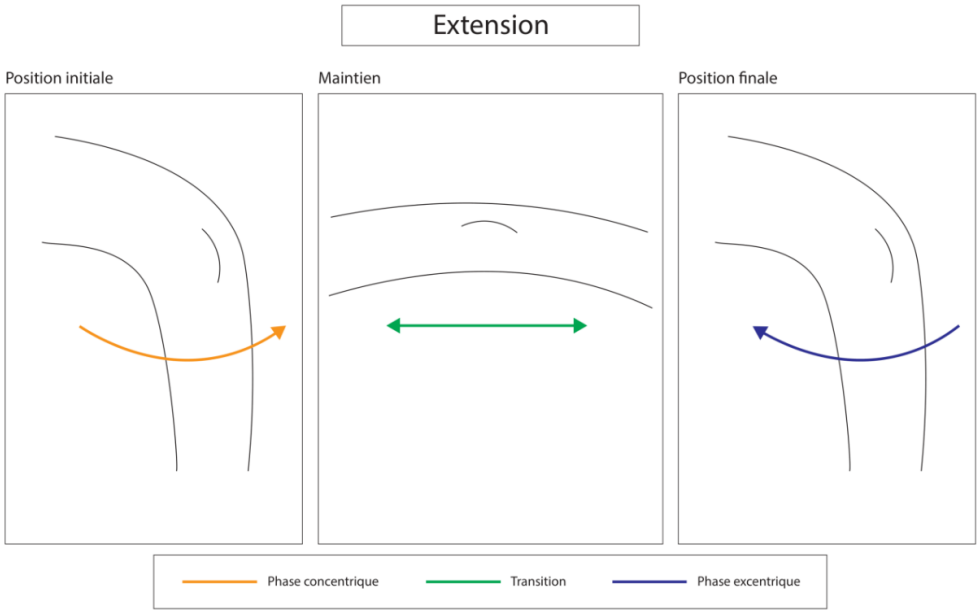

FIGURE 3.1 : Illustration du mouvement d'extension du genou.

2. **La flexion :** lors de cet exercice, les patients ont été installés allongés sur le ventre avec les deux jambes en appui sur un matelas. Ensuite, il a été demandé aux participants de plier la jambe pour que cette dernière arrive dans une position la plus perpendiculaire possible au matelas. Les participants devaient alors maintenir leur jambe pliée pendant au moins 3 secondes dans la mesure du possible selon leur condition puis revenir en position initiale, tels qu'illustrés par le second et le dernier encart de la Figure 3.2. Le mouvement de transition entre la position initiale et le maintien constitue la phase concentrique du mouvement tandis que le retour à la position initiale depuis le maintien est la phase excentrique du mouvement.

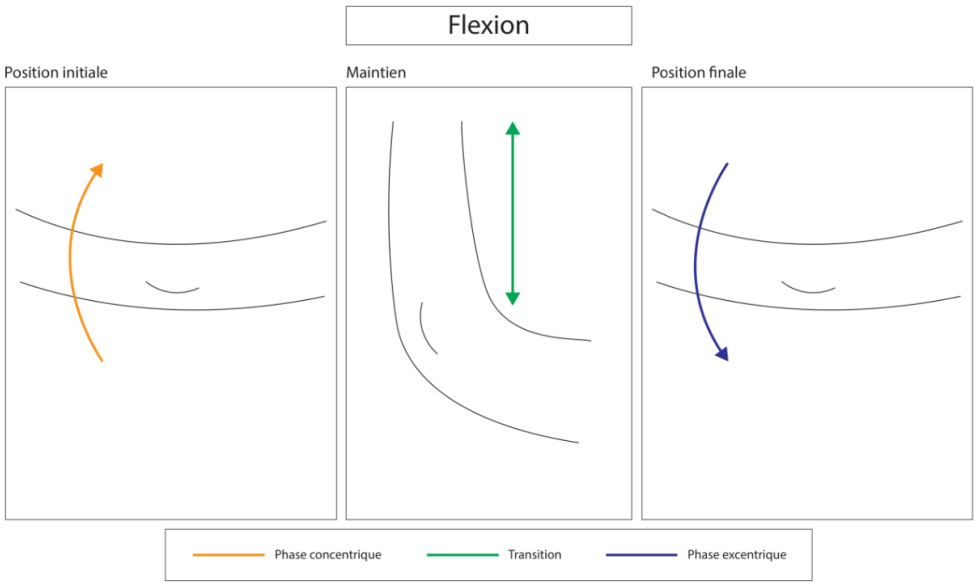

FIGURE 3.2 : Illustration du mouvement de flexion du genou.

SYSTÈME D'ACQUISITION

L'enregistrement des exercices décrits dans la section précédente s'est déroulé, avec tous les participants, par les membres du Groupe de Recherche Interdisciplinaire sur les Maladies Neuromusculaires (GRIMN) à la Clinique des Maladies Neuromusculaires (CMNM) de l'hôpital de Jonquière. La mesure de l'activation des muscles des membres inférieurs a été effectuée par le biais de capteurs d'EMG. Après une préparation de la peau, un ou plusieurs cliniciens du GRIMN avaient la charge de placer deux électrodes bipolaires de surface, en direction des fibres musculaires, et ce conformément aux recommandations du SENIAM [10], sur le rectus femoris et le biceps femoris du membre inférieur droit des participants. Ensuite, ils devaient leur faire réaliser un minimum de trois fois chacun des exercices précédemment énoncés (flexion et extension du genou et assis debout). Bien qu'étant capables de communiquer leurs informations sans fils, les données récoltées par les électrodes

étaient cependant transmises à l'ordinateur d'acquisition par le biais de la station *Trigno* [2], elle-même reliée par un câble à l'ordinateur. Ces données ont alors été compilées et stockées sur cette machine dans un format propriétaire à la société qui fabrique ce matériel spécifiquement conçu pour la recherche médicale.

3.2.2 TYPES DE DONNÉES

Parmi les données enregistrées lors des acquisitions pour toutes les phases du projet de recherche, il est possible de distinguer deux types de capteurs. La majorité des électrodes qui ont été disposées sur les participants sont des capteurs EMG dotés d'une fréquence d'échantillonnage de 1925.93 Hz auxquels est associé, dans le même boitier, un accéléromètre cadencé à 148.15 Hz. Par ailleurs, une électrode supplémentaire, légèrement plus complexe, d'une fréquence d'échantillonnage de de 1111.11 Hz, était également positionnée sur le participant dans le but d'obtenir plus précisément la mesure de l'angle du mouvement selon l'exercice réalisé. Comparativement à l'électrode standard, elle embarque une centrale inertielle complète soit, un gyroscope et un magnétomètre en plus de l'accéléromètre. Le détail des données générées par ces deux types d'électrodes est présenté dans le Tableau 3.1.

TABLEAU 3.1 : Liste des capteurs et leurs fréquences d'échantillonnage respectives.

Capteurs	Fréquences	Données
EMG + accéléromètre	1925.93 Hz 148.15 Hz	Temps, EMG, axes [x, y, z] de l'accéléromètre
EMG + IMU 9 Axes	1111.11 Hz	Temps, EMG, axes [x, y, z] de l'accéléromètre, du gyroscope et du magnétomètre

2. https ://delsys.com/trigno/

3.2.3 TRAITEMENT DES DONNÉES

L'application développée au sein du LIARA a pour but de traiter les données issues de cette expérimentation afin de principalement réaliser une analyse complète des coefficients de co-contraction, c'est-à-dire, l'activation simultanée de deux muscles agoniste/antagoniste (muscles ou groupes de muscles qui s'opposent les uns les autres) lors d'un mouvement.

Pour ce faire, la première manipulation que propose cette application est la conversion des fichiers de données enregistrés par les cliniciens d'un format HPF propriétaire vers un format exploitable facilement pour les autres traitements proposés par l'application (i.e. CSV). Ce choix de format n'a pas été fait par préférence, mais par nécessité, étant donné qu'il s'agit du seul format exploitable par la suite, malgré les problématiques inhérentes qu'il pourrait présenter dans certains cas. Une fois les fichiers de données convertis, un traitement est effectué. D'abord, un sur-échantillonnage (*upsampling*) à 1925.93 Hz est réalisé sur les données de l'électrode complexe (EMG + IMU 9 axes), ce qui permet d'augmenter la fréquence des données dont la fréquence initiale est plus faible afin de les aligner sur la fréquence des données EMG.

Les données EMG, une fois à la même fréquence, sont ensuite concaténées. Les valeurs moyennes sont retirées pour centrer le signal autour de zéro, et un filtre de *Butterworth* [28] est appliqué pour retirer les fréquences sans intérêt, avant l'extraction de l'enveloppe RMS [29] par le biais du produit de convolution appliqué sur une fenêtre rectangulaire de taille 0.2 seconde. Parallèlement, une courbe représentant les angles bruts est générée. Les données sont alors réduites, et une version réduite de la courbe des angles bruts est produite. Un filtre de *Savitzky-Golay* [30], permettant de lisser les données, c'est-à-dire d'augmenter la précision des données sans distordre la tendance du signal, est appliqué sur ces données réduites avant la création de la courbe des angles filtrés et réduits. À partir de cette dernière

et de la courbe réduite des angles bruts, des points sont sélectionnés par l'utilisateur via l'interface de l'application. Une extraction des données est ensuite réalisée dans l'intervalle des points sélectionnés.

À partir de l'enveloppe RMS et des données extraites, les enveloppes EMG sont normalisées dans un intervalle de 1000 points. Une enveloppe moyenne de ces enveloppes normalisées est ensuite créée. La valeur des aires sous la courbe est calculée pour chaque essai et pour chaque enveloppe moyenne en calculant la somme des valeurs des 1000 points. Par la suite, les ratios de co-contraction pour chaque essai et pour chaque angle moyen sont calculés. En parallèle, les angles sont également normalisés sur un intervalle de 1000 points et une moyenne de ces angles normalisés est calculée.

Le processus de traitement des données est décrit par le schéma fonctionnel simplifié donné en Figure 3.3.

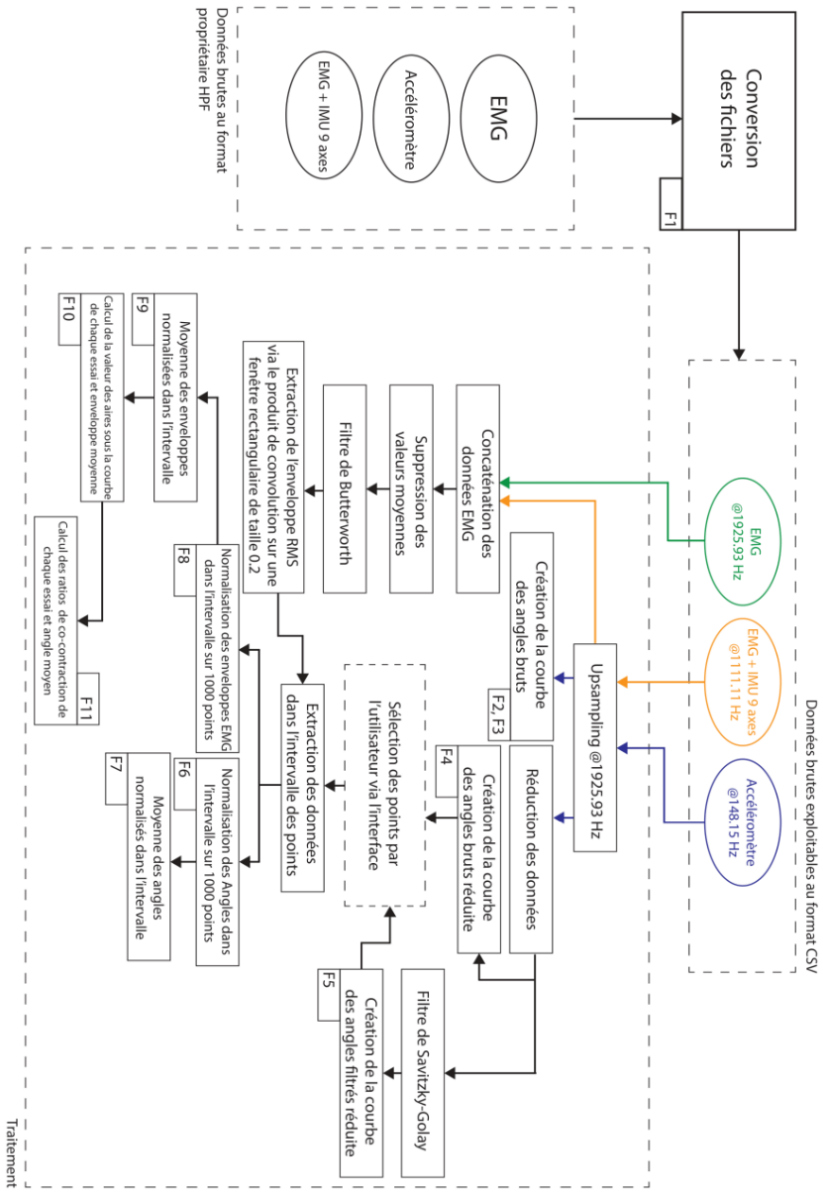

FIGURE 3.3 : Schéma fonctionnel simplifié du processus de traitement des données réalisé par l'application développée par le LIARA.

3.2.4 FICHIERS GÉNÉRÉS

Tout au long de son cycle de vie, l'application développée s'appuie sur l'écriture de fichiers de données intermédiaires qui constitue la base de connaissance de l'application et

qui est indexée selon les identifiants des participants. Les fichiers générés dans cette base de connaissance peuvent être de plusieurs types. Premièrement, il y a les fichiers de métadonnées qui sont utiles au programme et requis dans les traitements effectués par celui-ci, mais qui sont masqués aux utilisateurs de l'application. Ensuite, il y a les fichiers utilisateurs que ce dernier a alors la possibilité de manipuler s'il le souhaite (édition, suppression, *etc.*) et enfin, il y a les rapports d'analyse qui sont également exploitables par les utilisateurs et qui constituent des fichiers Excel qui récapitulent les opérations faites par l'application. Les différents fichiers sont indiqués sur le schéma fonctionnel par un encart qui comporte la lettre « F » suivie d'un chiffre ou d'un nombre.

MÉTADONNÉES

La liste des fichiers générés par l'application et qui correspondent aux métadonnées de cette dernière est donnée par le Tableau 3.2. Ces fichiers sont tous enfant du dossier :

```
Root_Folder/Analysis/.metadata/[analysis]/[participant]
```

où `[analysis]` correspond à l'exercice réalisé par le participant et `[participant]` est l'identifiant du participant en question.

TABLEAU 3.2 : Liste des fichiers de métadonnées générés par l'application.

ID	Fichier	Description
F7	`angles.csv`	Contient la moyenne des angles normalisés sur 1000 points dans l'intervalle sélectionné par l'utilisateur.
F10	`areas_[stage].json`	Contient l'ensemble des valeurs correspondantes aux aires sous la courbe des enveloppes pour chaque essai ainsi que pour l'enveloppe moyenne selon la phase (*i.e.* concentrique ou excentrique)
F8	`envelope_[stage]_[file].csv`	Contient les enveloppes normalisées sur 1000 points dans l'intervalle des angles sélectionnés par l'utilisateur selon la phase et l'essai.
F9	`envelopes_[stage].csv`	Contient l'enveloppe moyenne normalisée sur 1000 points dans l'intervalle des angles sélectionnés par l'utilisateur selon la phase.
F5	`filtered_angle_[file].json`	Contient les données des angles une fois réduits et filtrés selon l'essai.
F2	`full_angle_[file].csv`	Contient les données des angles bruts selon l'essai.
F6	`normalized_angles_[stage]_-[file].csv`	Contient les données des angles normalisés sur 1000 points dans l'intervalle des angles sélectionnés par l'utilisateur selon la phase et l'essai.
F11	`ratios_[stage].csv`	Contient l'ensemble des valeurs correspondantes aux ratios de co-contraction pour chaque essai ainsi que pour l'enveloppe moyenne selon la phase.
F4	`small_angle_[file].json`	Contient les données des angles bruts une fois réduits selon l'essai.

FICHIERS UTILISATEUR

La liste des fichiers générés par l'application et qui correspondent aux fichiers utilisateurs est donnée par le Tableau 3.3. Ces fichiers sont tous enfant du dossier :

```
Root_Folder/Analysis/[analysis]/[participant]
```

où `[analysis]` correspond à l'exercice réalisé par le participant et `[participant]` est l'identifiant du participant en question.

TABLEAU 3.3 : Liste des fichiers utilisateurs générés par l'application.

ID	Fichier	Description
F1	`[file].csv`	Contient l'ensemble des données brutes selon l'essai obtenu après la conversion depuis le format HPF propriétaire.
F3	`plot_angle_[file].svg`	Graphique de la courbe des angles selon l'essai généré à partir du fichier de méta-données `full_angle_[file].csv`.

RAPPORTS D'ANALYSE

La liste des fichiers générés par l'application et qui correspondent aux rapports d'analyse est donnée par le Tableau 3.4. Ces fichiers sont tous enfant du dossier :

```
Root_Folder/Results/[analysis]/[stage]
```

où `[analysis]` correspond à l'exercice réalisé par le participant et `[stage]` est la phase (*i.e.* concentrique ou excentrique) du mouvement réalisé pendant l'exercice.

TABLEAU 3.4 : Liste des fichiers, issus de l'analyse, produits par l'application.

Fichier	Description
`report_[analysis]_[stage]_-[participant].xlsx`	Rapport Excel produit à partir des différents traitements des données décrits auparavant selon l'exercice, la phase ainsi que le participant. Ce dernier contient trois onglets : — Le premier contient les données de la moyenne des enveloppes ainsi que des angles, lorsque normalisés sur 1000 points dans l'intervalle sélectionné par l'utilisateur. — Le second contient différentes analyses statistiques ainsi que les aires sous la courbe et les matrices des ratios de co-contraction pour chacun des essais. — Le dernier contient les différents graphiques de l'activité musculaire pour les muscles agonistes et antagonistes ainsi que la valeur de l'angle en fonction du temps.

3.3 NOUVELLE FONCTIONNALITÉ AJOUTÉE

Après la prise en main et la maîtrise de toutes les fonctionnalités de l'application, un nouveau type de fichiers à générer a été ajouté. Il s'agit des fichiers de synthèse qui sont générés à partir des rapports d'analyse et sont exploitables par les utilisateurs, ils constituent des fichiers Excel qui résument les résultats d'analyse. Le Tableau 3.5 donne la liste des fichiers générés qui correspondent aux fichiers de synthèse. Ces fichiers sont enfant du dossier :

Root_Folder/Results

TABLEAU 3.5 : Liste des fichiers, synthétisant l'analyse, produits par l'application.

Fichier	Description
`summary_[analysis]_[stage].xlsx`	Résumé Excel produit à partir des rapports d'analyse des participants selon l'exercice et la phase. Ce dernier se compose de trois tableaux distincts, chacun représentant une catégorie de participants (*sains*, *marcheurs* et *non-marcheurs*). Chaque tableau liste les participants de cette catégorie et présente trois colonnes de données : le *ratio antagoniste/agoniste* moyen, la *durée* moyenne et l'*amplitude* moyenne.

3.4 PRÉPARATION DES DONNÉES

Dans le but de pouvoir expérimenter différentes approches d'apprentissage automatiques, qui seront présentées utlérieurement dans les Chapitres 4 et 5, une préparation de données est requise afin d'obtenir les ensembles de données nécessaires. Le processus de préparation de données est adapté aux spécificités de chaque expérimentation.

3.4.1 COLLECTE DES DONNÉES

Tel que précisé précédemment, chaque participant a réalisé chaque exercice (extension et flexion du genou) au moins trois fois. Parmi ces répétitions, les trois itérations validées cliniquement ont été retenues. Ainsi, chaque itération retenue pour chaque participant représente une instance de l'ensemble de données à un instant T. Ce protocole d'acquisition a été répété à

trois temps différents, désignés par T3, T4 et T5, chacun espacé de deux ans. Par conséquent, chaque participant peut avoir trois itérations pour chacun de ces trois temps. Il est important de noter que certains participants n'ont pas été présents à tous les temps de mesure.

Lors de chaque exercice effectué, les données EMG des muscles sont enregistrées. Chaque enregistrement est constitué de 1000 points de données, représentant les variations d'activité musculaire mesurées à des intervalles réguliers pendant l'exercice. Dans l'ensemble de données, l'EMG de chaque muscle est représenté par l'aire sous la courbe (le graphe), soit la somme des 1000 points constituant l'EMG. Les ensembles de données comportent plusieurs colonnes, incluant des données EMG de chaque muscle. Ces données sont disponibles pour l'analyse des mouvements d'extension et de flexion en phase concentrique. En plus des données des aires d'activités musculaires, les ensembles de données comprennent des données cliniques recueillies par les cliniciens. Les données des aires d'activités musculaires pris en compte dans nos ensembles de données concernent les muscles suivants :

— **Muscle biceps fémoral** (*biceps femoris*) : Le biceps fémoral est un des trois muscles situés dans le compartiment postérieur de la cuisse. Ce muscle permet la fléxion ainsi que la rotation latérale de la jambe au niveau du genou. il permet également d'étendre et tourner latéralement la cuisse au niveau de la hanche [31].

— **Muscle droit fémoral** (*rectus femoris*) : Le muscle droit fémoral est situé dans le compartiment antérieur de la cuisse. Ce muscle est un des chefs du quadriceps fémoral et c'est le seul de ce groupe qui croise la hanche et le genou, contrairement aux autres qui croisent seulement le genou. Ce muscle contribue à la flexion de la cuisse au niveau de la hanche et, avec l'aide des autres muscles formant le quadriceps, permet l'extension de la jambe au niveau du genou [31].

— **Muscle droit de l'abdomen** (*rectus abdominis*) : Le muscle droit de l'abdomen est situé dans la paroi musculaire antérolatérale de l'abdomen. Ce muscle joue un rôle crucial dans la flexion de la colonne vertébrale [31], contribuant ainsi de manière significative à la stabilité du corps.

— **Muscle tenseur du fascia lata** (*tensor fascia lata*) : Le tenseur du fascia lata est un des muscles du groupe superficiel de la région glutéale. Ce muscle permet la stabilisation du genou lors de son extension [31].

— **Muscle grand glutéal** (*gluteus maximus*) : Le grand glutéal est également un des muscles du groupe superficiel de la région glutéale, et est le plus gros de cette région. Ce muscle permet l'extension du fémur lors de la fléxion de la hanche ainsi que la stabilisation de la hanche et du genou. Il permet également l'abdéction et la rotation latérale de la cuisse [31].

— **Muscle long adducteur** (*adductor longus*) : Le long adducteur est un des six muscles du compartiment médial de la cuisse. Ce muscle contribue autant qu'adducteur et rotateur médial de la cuisse au niveau de la hanche [31].

— **Muscle tibial antérieur** (*tibialis anterior*) : Le tibial antérieur est un des quatre muscles du compartiment antérieur de la jambe. Ce muscle assure l'inversion du pied au niveau de la cheville et permet le support dynamique de l'arche médiale du pied pendant la marche [31].

Les données cliniques sont recueillies *via* l'application d'un questionnaire sociodémographique aux participants, permettant de receuillir des informations, notamment l'age, le sexe et le niveau de mobilité intérieure du participant qui peut être classé en l'une des deux catégories suivantes : marche sans aide ou marche avec l'aide d'une canne ou d'un déambulateur à deux ou quatre roues. Les données cliniques sont également recueillies par des mesures

d'évaluations, passées aux participants par un seul et unique physiothérapeute qualifié, dans un ordre standardisé et en suivant une Procédure Opérationnelle Permanente (POP) [32].

Parmi les mesures d'évaluations passées aux participants et considérées dans les ensembles de données :

— **Échelle d'évaluation de l'équilibre de Berg** : Ce test permet de déterminer l'équilibre statique et dynamique des participants. Il se compose de quatorze épreuves, chacune pouvant recevoir une note allant de 0 (pour le pire cas) à 4 (pour le meilleur cas) dépendamment de l'exécution de l'épreuve de façon indépendante ainsi que de la durée et de la distance effectuée. Par conséquent, le score global peut varier entre 0 et 56 points, plus le score est élevé plus l'équilibre est meilleur [33].

— **Test de dix mètres de marche** : Ce test consiste à calculer la vitesse de marche du participant sur dix mètres [34]. Les participants ont effectué le test avec une vitesse confortable et avec leur vitesse maximale.

— **Test de coordination motrice des membres inférieurs** : Ce test permet de mesurer la coordination des membres inférieurs en bougeant le plus vite possible le membre inférieur d'une cible A à une cible B, distantes de trente cm, et vice versa, pendant vingt secondes. Le score représente le nombre de fois où les cibles sont touchées [35]. Les participants ont effectué le test sur chaque membre inférieur, à savoir le droit et le gauche.

— **Indice de gravité de l'ARSACS** : Ce test permet de déterminer le degré de sévérité de la maladie chez les participants en utilisant l'indice de gravité de la maladie pour l'ataxie récessive spastique autosomique de Charlevoix-Saguenay [36]. Il se compose de huit épreuves, le score global peut varier entre 0 et 38 points, moins le score est élevé moins la maladie est sévère.

Afin de disposer d'un ensemble de données complet, comprenant toutes les informations nécessaires, un script *Python* a été développé pour permettre la fusion des données cliniques et des données des aires d'activités musculaires recueillies auprès des participants, et ce, selon chaque itération et à chaque instant. De cette façon, chaque instance dans l'ensemble de données représente une des trois itérations d'un participant à un temps précis et est constitué des données des aires d'activités musculaires de chaque muscle lors de l'extension en phase concentrique, des données des aires d'activités musculaires de chaque muscle lors de la flexion en phase concentrique ainsi que les données cliniques.

Une fois l'ensemble de données constitué, il est spécifiquement ajusté pour l'expérimentation en cours en sélectionnant et en conservant les variables les plus pertinentes et significatives pour résoudre le problème en question.

Pour aborder certains problèmes, de nouvelles caractéristiques sont créées. Cette étape constitue l'ingénierie des caractéristiques (*feature engineering*), processus qui inclut l'approche de création de caractéristiques anticipées (*lead features*). Dans ce cas, cette approche consiste à ajouter des caractéristiques représentant des mesures d'évaluation pour un instant futur. Par exemple, pour une instance représentant un patient à un instant T, on ajoute une caractéristique qui représente une certaine mesure d'évaluation (selon le problème à résoudre) pour ce même patient à l'instant $T+1$. Pour les instances représentant les patients au dernier instant observé, cette caractéristique prendra une valeur vide, laquelle sera ensuite traitée lors de l'étape de nettoyage des données. Cette approche est pertinente, étant donné qu'elle est utilisée pour des analyses exploratoires dans le but de comprendre les relations temporelles et les tendances au sein des données.

3.4.2 NETTOYAGE DES DONNÉES

Étant donné le protocole de collecte des données et le matériel utilisé, il arrive parfois que des données soient manquantes. Cette étape consiste principalement à gérer les valeurs manquantes afin de garantir l'intégrité et la qualité de l'analyse. Tenant compte de la nature des données et de la faible proportion de valeurs manquantes par rapport aux données complètes et leur caractère aléatoire, les instances comportant des données manquantes ont été supprimées. Certaines incohérences et erreurs sont corrigées lors de la création des ensembles de données à l'aide du script *Python*.

3.4.3 TRANSFORMATION DES DONNÉES

Cette étape implique principalement la normalisation et la mise à l'échelle des données pour les placer dans une plage commune, ce qui est crucial pour permettre une comparaison précise et équitable des caractéristiques. Cette étape est essentielle car des échelles différentes peuvent entraîner des interprétations erronées lors des analyses. Comme mentionné précédemment, les différentes étapes de préparation des données sont adaptées en fonction des besoins spécifiques de chaque expérimentation. Ainsi, la normalisation et la mise à l'échelle sont effectuées selon les exigences particulières de l'expérience en cours.

3.4.4 RÉDUCTION DE LA DIMENSIONNALITÉ

En raison des 30 variables présentes dans nos ensembles de données, il est pertinent de procéder à une réduction de la dimensionnalité afin de conserver l'essentiel des informations tout en simplifiant l'analyse. La technique utilisée pour effectuer la réduction de la dimensionnalité est l'PCA [37], qui consiste à transformer les variables de l'ensemble de données en composantes principales, des combinaisons linéaires des variables originales. Le nombre

de composantes principales est déterminé en fonction de la dimensionnalité souhaitée. Cette étape est essentielle pour certaines expérimentations, car elle permet de simplifier les modèles sans compromettre leurs performances.

3.5 CONCLUSION

Ce chapitre a détaillé le processus rigoureux de collecte, de préparation et de traitement des données EMG, qui constitue une étape cruciale dans notre projet de recherche. À travers l'application MyoRatio, nous avons non seulement exploité les données issues des exercices réalisés par les participants, mais aussi optimisé ces données pour l'analyse des coefficients de co-contraction, facilitant ainsi la conversion, le traitement et la normalisation des données EMG. Ces étapes permettent de garantir la qualité et la pertinence des ensembles de données utilisés dans les expérimentations d'apprentissage automatique présentées dans les Chapitres 4 et 5 suivants. En intégrant des données cliniques et EMG, nous avons ainsi construit des ensembles de données cohérents et complets, prêts pour l'exploration des approches de *clustering* et de régression qui visent à identifier des marqueurs prédictifs et à comprendre la progression de l'ARSACS.

CHAPITRE IV
EXPÉRIMENTATIONS ET RÉSULTATS DE CLUSTERING

4.1 INTRODUCTION

Dans ce chapitre, nous exposons et analysons les différentes expérimentations de *clustering* menées et les résultats obtenus. Les approches de *clustering* ont été utilisées dans le but de tenter d'identifier des groupements inédits de profils de participants sans nécessiter d'étiquettes prédéfinies. Autrement dit, elles visent à essayer de découvrir des sous-groupes de participants partageant des similitudes, pas nécessairement évidentes au premier abord, soit en termes de ressemblance de leurs attributs, soit en termes de progression de la maladie.

4.2 CLUSTERING

Dans cette première approche, l'ensemble de données utilisé concerne les données au temps T3 et contient les caractéristiques suivantes : la durée, l'amplitude, les données des aires d'activités musculaires du muscle biceps fémoral et du muscle droit fémoral, et ce pour la phase concentrique de l'extension et de la flexion. L'ensemble de données inclut également deux mesures d'évaluation, à savoir le score de Berg ainsi que l'amplitude active en phase d'extension.

L'ensemble de données contient initialement la caractéristique représentant l'état du patient, une variable catégorielle de trois catégories : *healthy* pour un participant sain, *walker* pour un participant atteint d'ARSACS marcheur et *non-walker* pour un participant atteint d'ARSACS non-marcheur. À partir de cette variable, les instances où l'état du participant est identifié comme *walker* sont ajustées pour diviser cette catégorie en trois sous-catégories. Cette division est fondée sur le score de Berg associé à chaque instance : la sous-catégorie *walker-1*

est attribuée aux instances pour lesquelles le score de Berg est inférieur à 25, *walker-2* pour celles dont le score se situe entre 25 et 45 inclusivement, et *walker-3* pour celles avec un score supérieur à 45.

Dans cette approche, les variables considérées sont les données des aires d'activités musculaires du muscle biceps fémoral et du muscle droit fémoral lors de la phase concentrique de l'extension et de la flexion. Comme décrit dans le chapitre précédent, l'ensemble de données contient initialement trois instances par participant, soit une instance pour chaque itération, pour un total de trois itérations par participant. Avec un total de 60 participants dans l'étude au temps T3, cela équivaut à 180 instances dans l'ensemble de données avant l'étape de nettoyage de données. 10% de ces instances, soit 18 instances, présentaient au moins une valeur manquante. En raison de la nature sensible de l'étude et de l'importance cruciale de l'intégrité des données dans ce contexte, ces instances comportant des valeurs manquantes ont été supprimées. L'imputation a été évitée pour ne pas introduire de potentielles inexactitudes, ce qui est particulièrement préoccupant vu les exigences strictes de précision et de fiabilité des données. Au final, après suppression des instances avec des valeurs manquantes, l'ensemble de données contient 162 instances complètes.

Après l'étape de nettoyage des données, une normalisation de celles-ci a été réalisée en raison de la sensibilité des algorithmes de *clustering* à l'échelle des caractéristiques. En effet, le *clustering* repose sur des mesures de distance pour regrouper les points de données en clusters. Par conséquent, la normalisation permet à chaque caractéristique de contribuer de manière équivalente au calcul de la distance. L'approche de normalisation utilisée est le *Min-Max Scaling*, qui transforme les caractéristiques de telle sorte qu'elles soient sur une échelle de 0 à 1.

Afin d'identifier les relations linéaires entre les variables, la matrice de corrélation deux-à-deux est calculée et présentée à la Figure 4.1. Les coefficients de corrélation se situent dans un intervalle allant de -1 à 1. Un coefficient de corrélation proche de 1 indique que deux variables ont une forte relation linéaire positive tandis qu'un coefficient de corrélation proche de -1 indique que deux variables ont une forte relation linéaire négative. Un coefficient de corrélation proche de 0 indique une absence de relation linéaire. Les variables fortement corrélées sont éliminées pour réduire la dimensionnalité de l'ensemble de données et éviter le problème de multicolinéarité. De plus, certains algorithmes de *clustering*, notamment le *K-means*, l'algorithme utilisé dans cette expérience, supposent que les variables sont indépendantes.

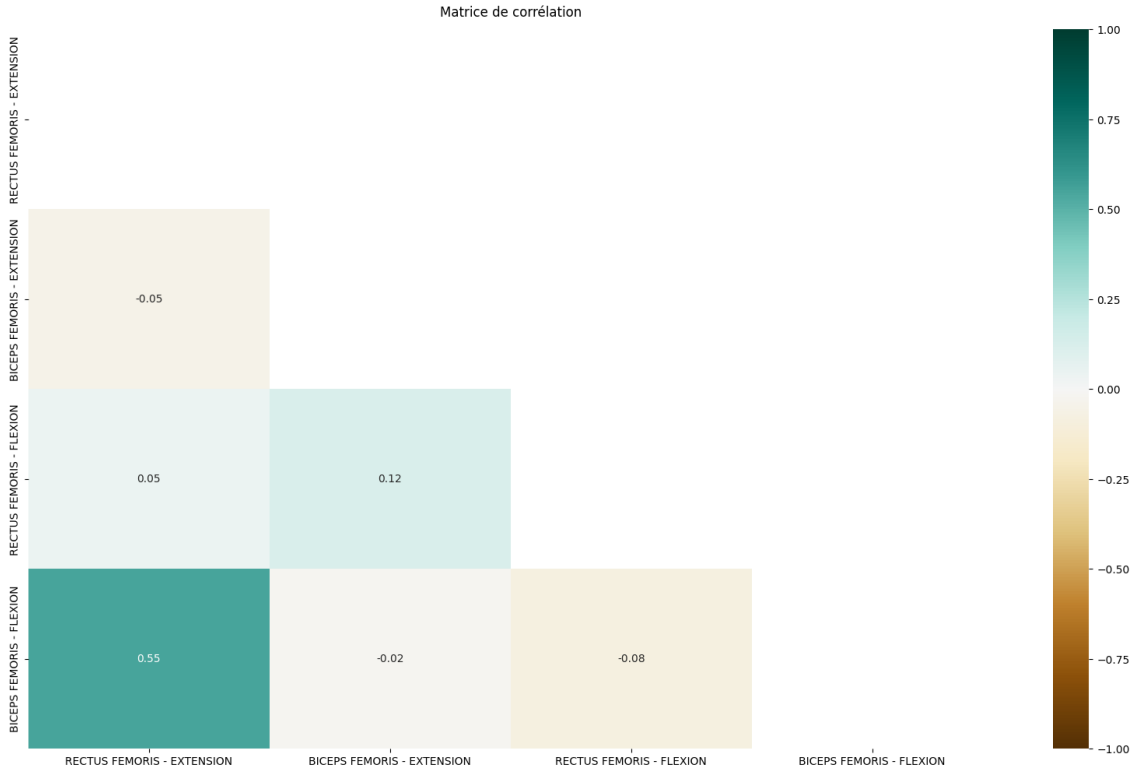

FIGURE 4.1 : Matrice de corrélation.

Afin de déterminer le nombre optimal de clusters (groupes), la méthode du coude a été utilisée. Cette méthode vise à trouver le point où l'ajout de clusters supplémentaires n'améliore pas significativement la variance expliquée par les données. La courbe est tracée pour identifier ce point, également appelé le coude. Ce point correspond au nombre de clusters qui offre le meilleur équilibre entre complexité du modèle et qualité de la classification des données.

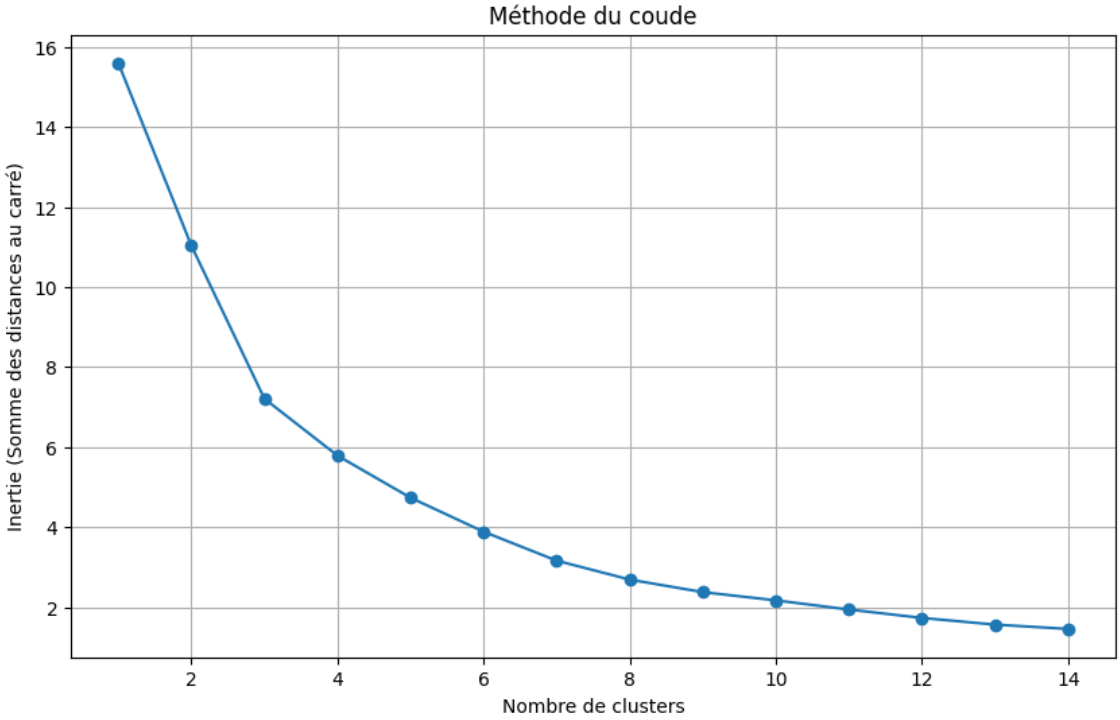

FIGURE 4.2 : Représentation graphique de la courbe du coude.

La Figure 4.2 présente un graphique représentant la courbe du coude. Selon celle-ci, le nombre optimal de clusters est de trois. L'algorithme *K-means* a donc été utilisé afin de segmenter les données en 3 groupes.

Une fois les clusters obtenus, la technique de PCA est utilisée pour réduire la dimensionnalité aux deux premières composantes principales pour une visualisation en 2D, et ensuite

aux trois premières pour une visualisation en 3D. La PCA a été choisie pour sa capacité à simplifier les données tout en conservant les aspects les plus significatifs, facilitant ainsi la compréhension des tendances générales [37].

Une autre technique a été utilisée pour la visualisation des clusters obtenus, l'algorithme t-SNE (*t-distributed stochastic neighbor embedding*) permettant la réduction de dimensionnalité tout en préservant la structure locale des données. Cette technique garde les instances similaires proches les unes des autres et les instances dissimilaires éloignées dans l'espace de dimension réduite, offrant une visualisation détaillée des clusters [37].

Ces deux méthodes sont donc complémentaires, la PCA fournit une vue d'ensemble simplifiée, tandis que le t-SNE permet d'examiner plus en détail les relations locales entre les points de données.

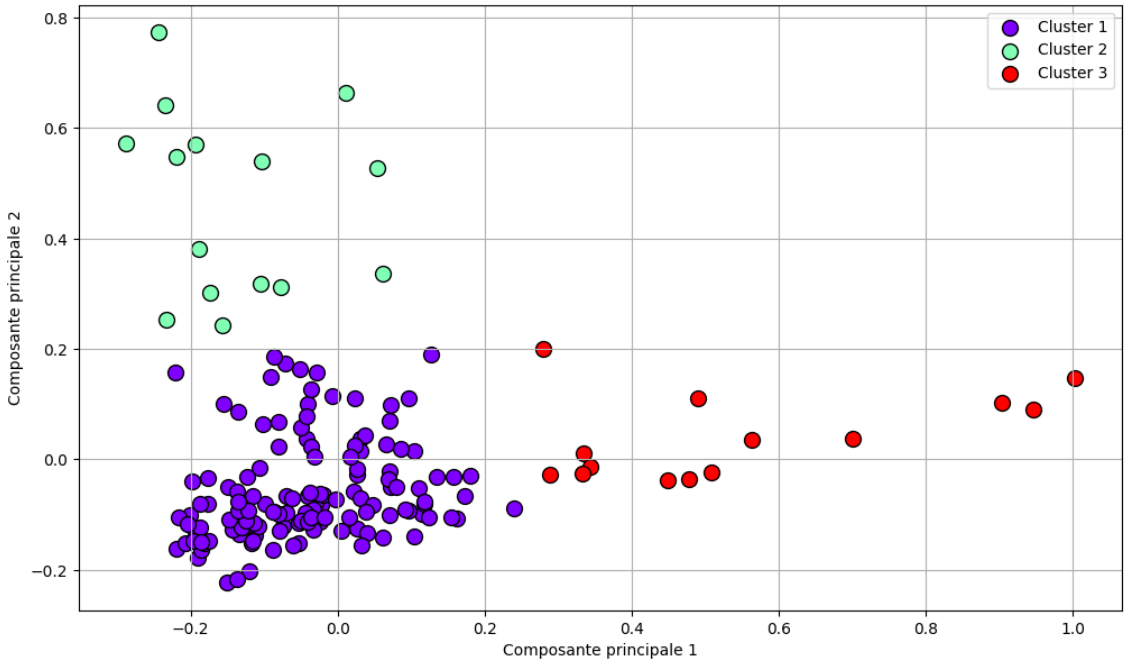

FIGURE 4.3 : Représentation bidimensionnelle des clusters obtenus en utilisant PCA.

La Figure 4.3 illustre la répartition des instances en trois clusters distincts, résultant de l'application de l'algorithme *K-means*. Les données sont réduites en deux dimensions via PCA, et sont représentées par deux axes correspondant aux deux premières composantes principales qui capturent la plus grande variance des données. Chaque point représente une instance de l'ensemble de données et sa couleur indique l'appartenance à un des trois clusters : violet pour le Cluster 1, vert pour le Cluster 2 et rouge pour le Cluster 3. Cette visualisation met en évidence la séparation entre les différents groupes de patients basée sur les caractéristiques EMG analysées. On observe une distinction claire entre les clusters, particulièrement entre le Cluster 1 et les deux autres. Le Cluster 1 semble densément peuplé et centré autour des valeurs proches de zéro sur les deux composantes principales. Le Cluster 2 est dispersé sur une plage plus large sur l'axe de la deuxième composante principale, suggérant une plus grande variabilité au sein de ce groupe. Enfin, le Cluster 3 est positionné principalement sur des valeurs plus élevées de la première composante principale, suggérant une différence distincte dans les caractéristiques de profil par rapport aux deux autres clusters.

Le regroupement dense observé au sein du Cluster 1 pourrait indiquer une uniformité dans les caractéristiques EMG des participants. En revanche, la dispersion observée au sein du Cluster 2 pourrait refléter une hétérogénéité plus marquée dans les caractéristiques des patients de ce groupe. Quant au positionnement distinct du Cluster 3, il suggère que ce groupe pourrait représenter une catégorie de patients présentant des caractéristiques EMG nettement différentes, suggérant probablement qu'il s'agit de participants sains.

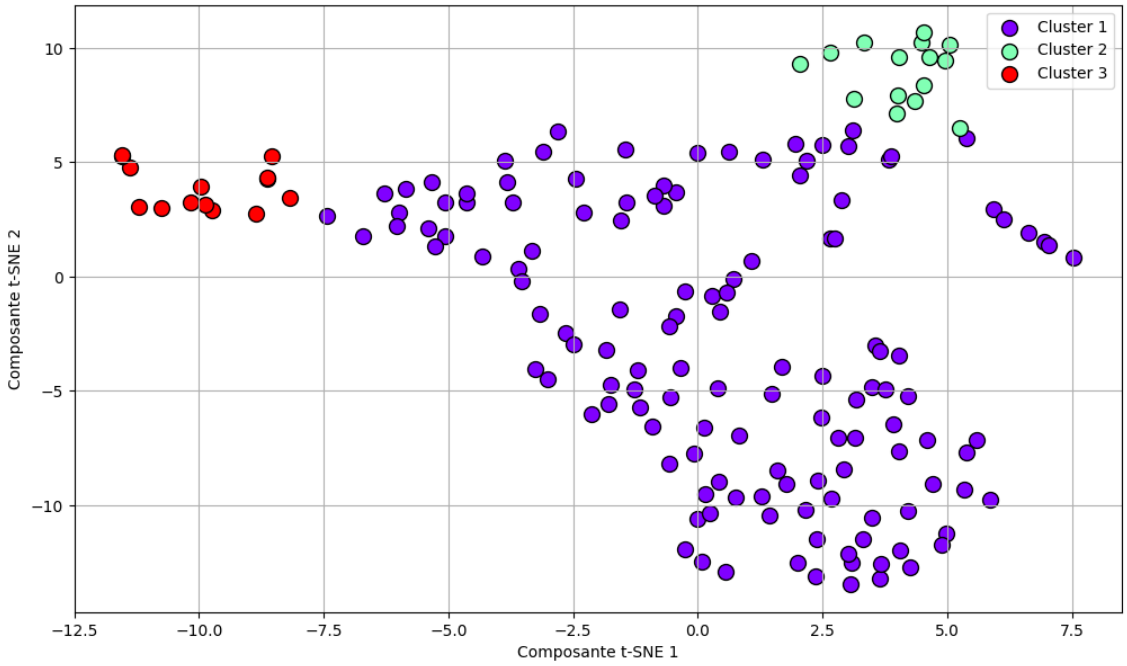

FIGURE 4.4 : Représentation bidimensionnelle des clusters obtenus en appliquant l'algorithme t-SNE.

La Figure 4.4 illustre la répartition des données en clusters suite à l'application de l'algorithme t-SNE. Les points sont colorés en fonction des trois clusters identifiés, avec les Clusters 1, 2, et 3 représentés en violet, vert et rouge respectivement. Dans cette représentation bidimensionnelle, on observe que le Cluster 1 s'étend principalement autour de la région centrale, avec une dispersion modérée dans l'espace t-SNE. Le Cluster 2 est clairement séparé sur l'axe t-SNE 2, démontrant un regroupement distinct en hauteur. Le Cluster 3 est également bien distinct, situé à l'extrémité négative de l'axe t-SNE 1, ce qui suggère un ensemble de caractéristiques fortement différent de celui des deux autres clusters.

Bien que les représentations en deux dimensions soient suffisamment explicites et que les clusters identifiés apparaissent cohérents, l'ajout d'une troisième dimension pourrait potentiellement confirmer la stabilité des clusters déjà identifiés. Elle permettrait également d'examiner l'impact d'une troisième composante principale sur la séparation des clusters.

Cependant, dans cette approche, les visualisations en trois dimensions des clusters ne sont pas présentées, car elles n'altèrent pas l'interprétation des données ni la compréhension des clusters. Par conséquent, elles ne fournissent pas d'informations supplémentaires pertinentes pour notre analyse.

Afin de mieux comprendre les clusters obtenus et comment ils peuvent refléter des niveaux de sévérité différents dans l'ARSACS, il est essentiel d'examiner la répartition des différentes catégories d'état des patients (healthy, non-walker, walker-1, walker-2, walker-3) à travers les clusters identifiés, comme l'illustre la Figure 4.5.

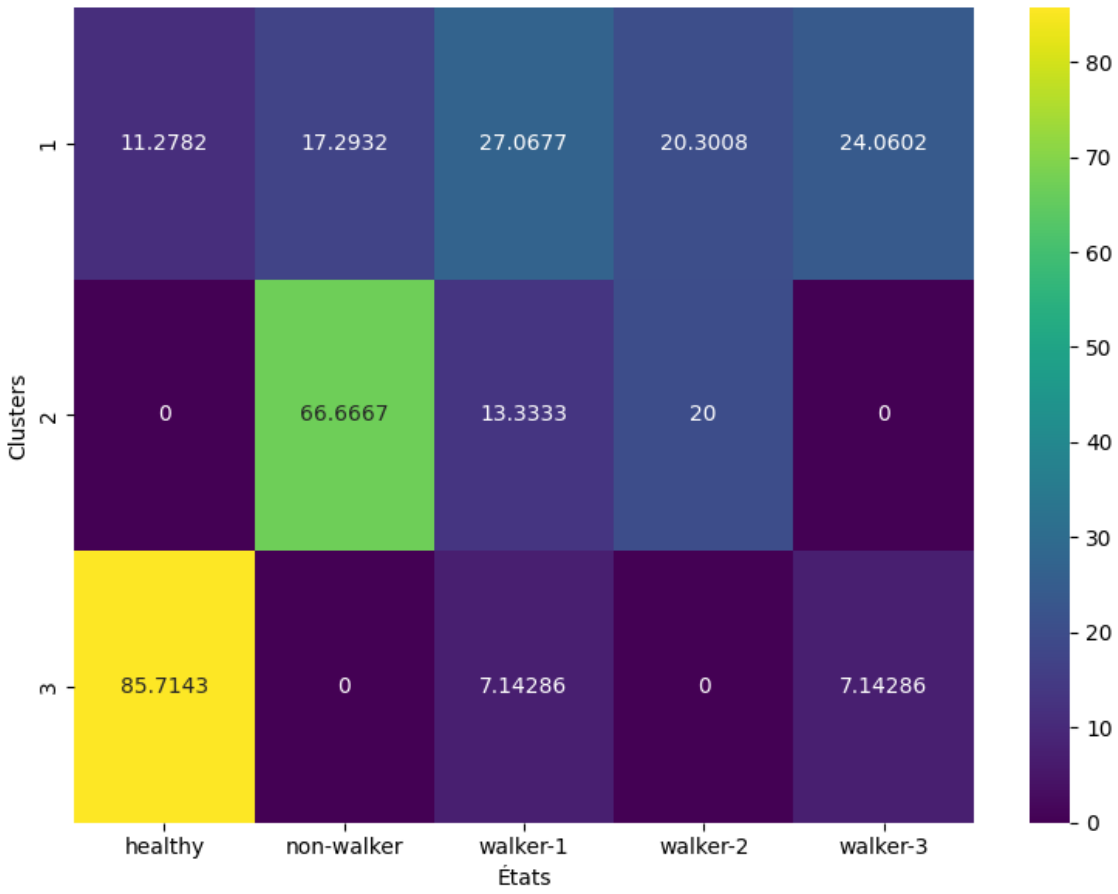

FIGURE 4.5 : Répartition des différentes catégories d'états des participants à travers les clusters identifiés.

Le Cluster 1 contient un mélange de tous les états, ce qui pourrait indiquer qu'il regroupe des participants avec une variabilité importante dans les caractéristiques. Le Cluster 2 inclut majoritairement des participants non marcheurs, ainsi que des participants avec un score de Berg bas ou moyen, ce qui peut suggérer que ce cluster regroupe les participants dont la mobilité est plus sévèrement affectée. Quant au Cluster 3, il regroupe essentiellement des participants sains, indiquant ainsi que leurs caractéristiques se distinguent nettement de celles des participants atteints d'ARSACS. Compte tenu de cette analyse, le *clustering* peut aider à identifier des marqueurs EMG spécifiques qui distinguent les catégories d'état des patients et par conséquant leur degré d'atteinte.

Pour une analyse plus approfondie de la répartition des instances dans chaque cluster, une visualisation 3D, résultant d'une analyse en composantes principales, est réalisée afin de démontrer la distribution des différentes catégories d'états des participants au sein de chaque cluster. Pour ce faire, les clusters sont segmentés en sous-groupes basés sur les catégories d'états présents dans le cluster. Chacun de ces sous-groupes correspond aux participants partageant un même état au sein d'un cluster spécifique. Chaque point représente une instance de l'ensemble de données et la couleur de chaque point correspond à une combinaison d'états de santé et de cluster.

L'analyse des points distribués dans l'espace tridimensionnel des caractéristiques EMG montre une superposition notable entre les clusters, révélant la variabilité des caractéristiques EMG chez les individus atteints de l'ARSACS ainsi que chez ceux en bonne santé. Cette observation souligne que les délimitations entre les états de santé ne sont pas nettement définies dans l'espace des caractéristiques EMG, reflétant une complexité et une variabilité inhérente à la maladie de l'ARSACS ainsi qu'aux caractéristiques EMG elles-mêmes.

Ce mélange des états de santé au sein des clusters pourrait refléter une similarité des profils EMG parmi des patients de différents états de santé, suggérant ainsi que les marqueurs EMG ne sont pas suffisamment discriminants. Cela peut être dû à la nature de la maladie d'ARSACS, caractérisée par une variabilité intrinsèque où les différences entre les états de santé ne sont pas toujours claires ni linéairement séparables. En effet, cela souligne l'ampleur de la complexité dépassant la simple catégorisation des états de santé sur la base des caractéristiques EMG, reflétant ainsi la difficulté de distinguer l'état de santé et le degré d'atteinte d'une personne souffrant d'ARSACS.

4.3 CLUSTERING AVEC TROIS MUSCLES

Dans cette approche, l'ensemble de données utilisé est le même que celui utilisé dans l'approche précédente. Il concerne les données au temps T3 et contient les mêmes caractéristiques. De la même manière que dans l'approche précédente, la catégorie *walker* de l'ensemble de données initial, décrivant l'état de santé des participants atteints d'ARSACS marcheurs, est segmentée en trois sous-catégories : *walker-1*, *walker-2* et *walker-3* basées sur le score de Berg.

Quant aux variables considérées dans cette nouvelle approche, elle se distingue de la précédente par une sélection plus restreinte des caractéristiques étudiées. Alors que l'approche antérieure englobait les données des aires d'activités musculaires du muscle biceps fémoral et du muscle droit fémoral, tant pendant la phase concentrique de l'extension que de la flexion, l'approche actuelle se base sur le choix aléatoire de seulement trois des quatre caractéristiques précédemment considérées. Le choix a été effectué de manière aléatoire, étant donné l'incertitude quant à la détermination de la combinaison optimale des caractéristiques nécessaires pour réaliser une sélection efficace. De plus, cette limitation à trois caractéristiques est adoptée afin de faciliter les visualisations tridimensionnelles directes, sans recourir aux techniques de réduction de dimensionalité telles que PCA et t-SNE utilisées précédemment. Cette approche est justifiée par les limites inhérentes à ces techniques de réduction, notamment leur potentiel à obscurcir les relations intrinsèques entre les variables en compressant l'information dans un espace de dimensions réduites, ce qui peut parfois mener à une interprétation erronée des dynamiques complexes entre les caractéristiques étudiées. Cela pourrait également offrir une meilleure intuition sur la séparation et la distribution des clusters dans un espace tridimensionnel.

Les étapes de prétraitement de données appliquées dans cette approche sont identiques à celles appliquées dans l'approche précédente. Pour commencer, un nettoyage de données consistant à supprimer les instances comportant des valeurs manquantes a été effectué afin d'avoir des données cohérentes. Après ce processus, un total de 162 instances a été retenu pour l'analyse.

Après que les données aient été nettoyées, elles sont normalisées afin que chaque caractéristique contribue de manière équivalente. La technique de normalisation utilisée est le *Min-Max Scaling*, qui transforme les caractéristiques de telle sorte qu'elles se trouvent sur une échelle identique, soit entre 0 et 1 dans le cas de cette expérimentation. L'absence de valeurs aberrantes dans notre ensemble de données réduit le principal inconvénient de cette technique de normalisation, qui est sa sensibilité aux valeurs extrêmes, rendant ainsi cette méthode un choix judicieux pour cette expérimentation.

Cette étape est cruciale étant donné que les modèles basés sur les distances, tel que l'algorithme *K-means* utilisé dans cette expérimentation, supposent que toutes les caractéristiques sont normalisées et permettent donc un fonctionnement optimal lorsque cette condition est remplie.

Ensuite, la matrice de corrélation est calculée pour mesurer le degré de relation entre les caractéristiques deux à deux. Les valeurs obtenues, proches de 0, indiquent une faible corrélation entre les variables. La Figure 4.6 illustre cette matrice de corrélation, où l'on peut observer les coefficients de corrélation entre les différentes variables.

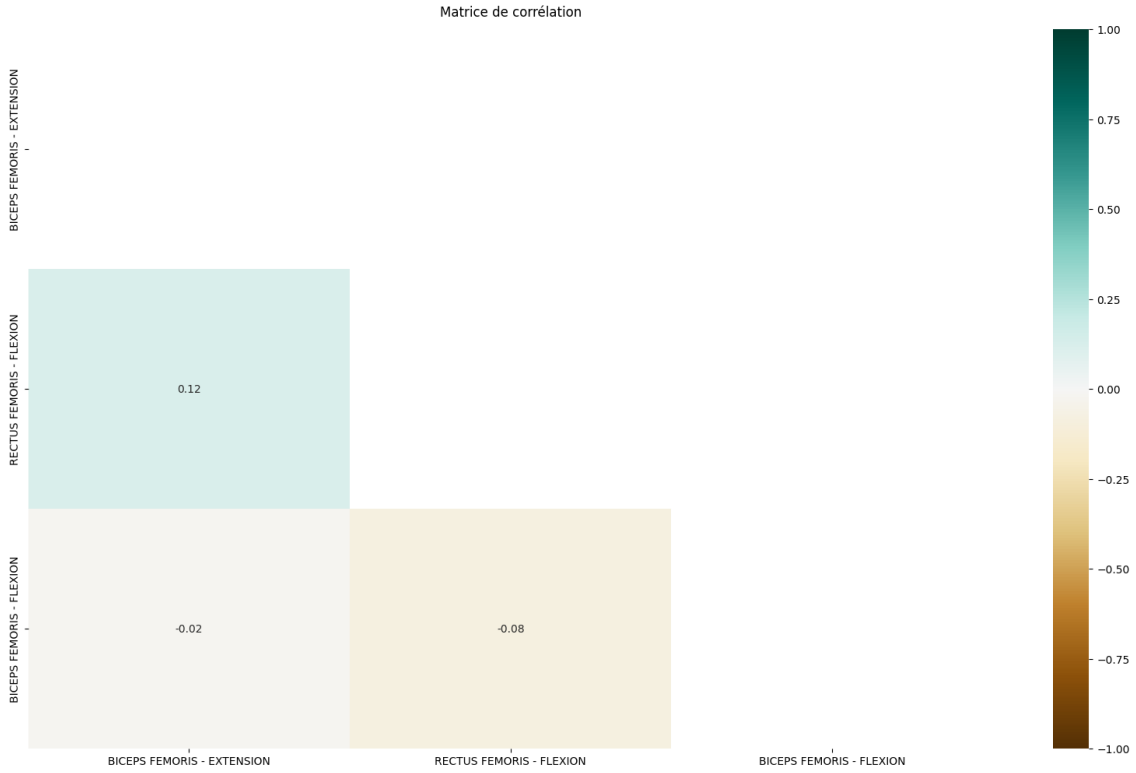

FIGURE 4.6 : Matrice de corrélation.

Une fois l'étape de préparation de données complétée, la méthode du coude est utilisée afin de déterminer le nombre optimal de clusters dans cette expérimentation. Selon la courbe du coude, le nombre optimal de clusters est de trois.

Pour le *clustering* des données, l'algorithme *K-means* a été choisi pour sa simplicité et son efficacité. De plus, *K-means* permet de fournir des résultats clairs et interprétables, malgré la complexité des données. Ceci est important dans le cadre de notre analyse afin d'explorer en profondeur les patterns et les clusters au sein des données. Une fois les clusters obtenus, une visualisation tridimensionnelle a été réalisée, prenant chaque caractéristique comme un des trois axes de la visualisation.

La Figure 4.7 illustre la répartition des instances, résultant de l'application de l'algorithme *K-means* avec K égal à trois. L'axe X représente les données des aires d'activités musculaires du muscle biceps fémoral en phase d'extension, l'axe Y représente les données des aires d'activités musculaires du muscle droit fémoral en phase de flexion et l'axe Z représente les données des aires d'activités musculaires du muscle biceps fémoral en phase de flexion.

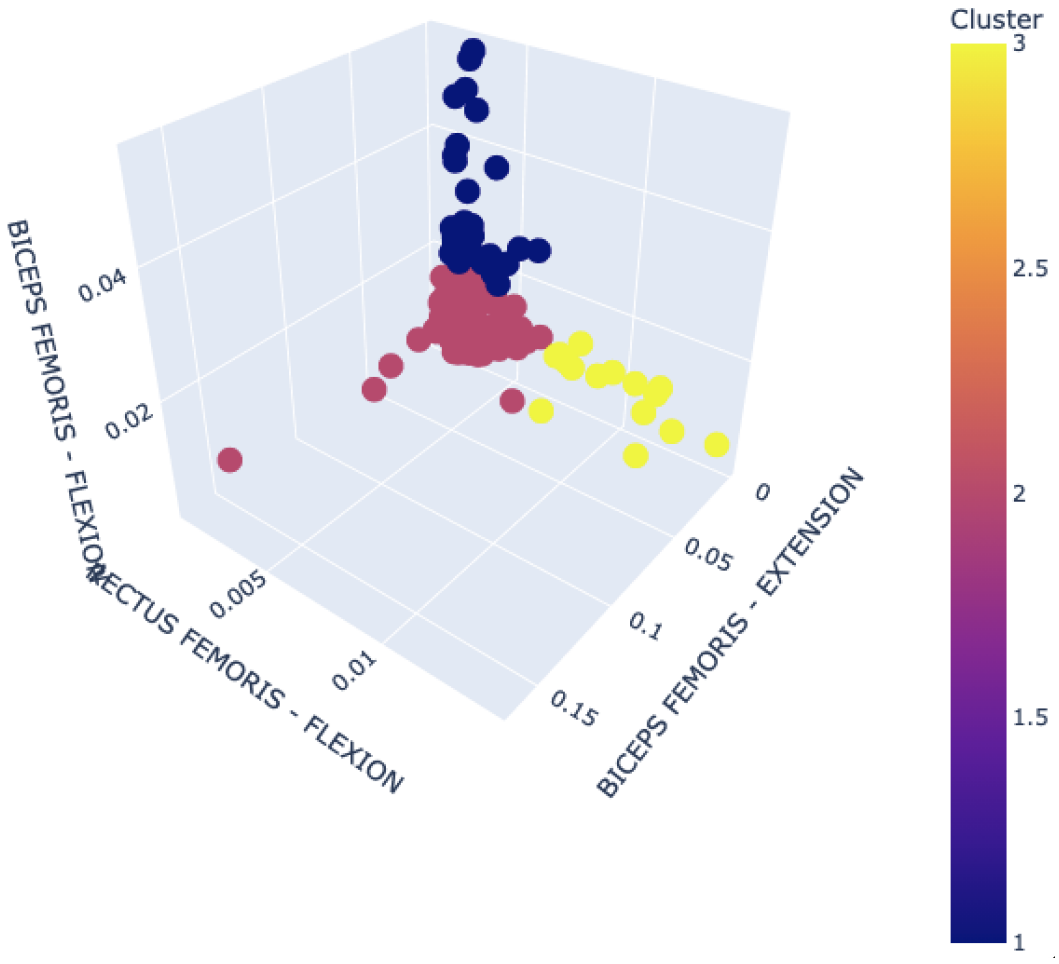

FIGURE 4.7 : Représentation tridimensionnelle des clusters obtenus.

Les clusters sont bien séparés, cela indique que les données forment des groupes distincts en fonction des variables mesurées, ce qui peut être révélateur de différences significatives

dans les caractéristiques des signaux EMG pour les différents états de santé des participants. La séparation claire entre les clusters suggère que l'algorithme *K-means* a été efficace pour trouver une structuration pertinente dans les données et que les caractéristiques choisies pour le *clustering* sont discriminantes.

Cependant, pour que cette discrimination soit cliniquement pertinente et pour une interprétation concrète, il est essentiel de mieux comprendre les clusters obtenus et la manière dont ils peuvent refléter différents niveaux de sévérité dans l'ARSACS. Il est également nécessaire d'établir un lien entre les clusters et les états de santé des participants. La Figure 4.8 illustre la répartition des différentes catégories d'état des patients (*healthy*, *non-walker*, *walker-1*, *walker-2*, *walker-3*) à travers les clusters identifiés.

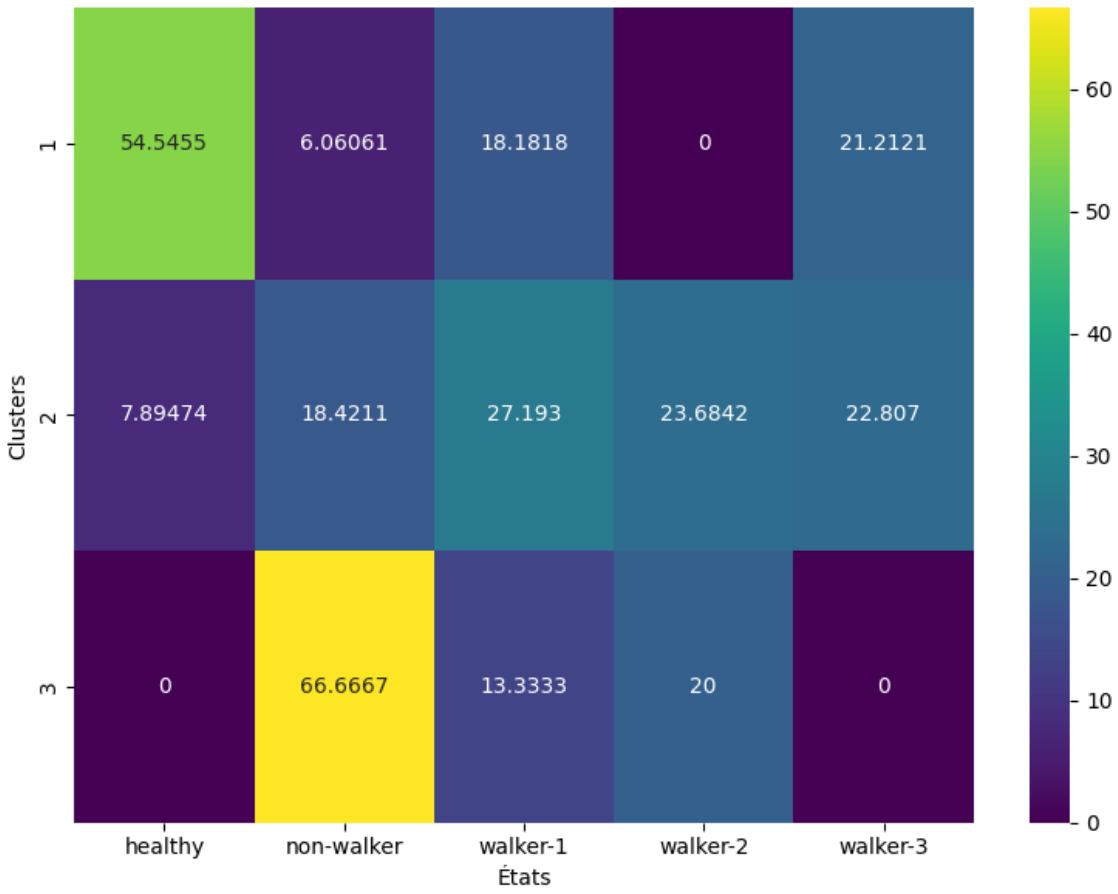

FIGURE 4.8 : Répartition des différentes catégories d'état des participants à travers les clusters identifiés.

Le Cluster 1 est principalement composé de participants *healthy*, constituant plus de la moitié de la population du cluster, avec une présence notable de participants *walker-1* et *walker-3* et une absence de *walker-2*. Il y a aussi une petite proportion de participants *non-walker*, ce qui suggère une certaine superposition des caractéristiques EMG entre les états de santé.

Le Cluster 3 est majoritairement constitué de participants *non-walker*, il comprend également des *walker-1* et *walker-2*, mais aucun participant *healthy* ou *walker-3*. Ceci suggère que ce cluster regroupe des états de santé plus sévères avec une limitation significative de la

mobilité. L'absence de participants *healthy* et de *walker-3* (catégorie de marcheurs avec une bonne mobilité) indique que les caractéristiques EMG de ce cluster sont probablement celles qui sont le plus altérées par l'ARSACS.

Le Cluster 2 rassemble des participants présentant divers états de santé, ce qui reflète une grande variété de conditions physiologiques, notamment en termes d'équilibre, au sein du groupe. Toutefois, les caractéristiques des signaux EMG associées à ces différents états de santé semblent subtiles et ne présentent pas de différences nettes. Cela suggère que, malgré la diversité des états de santé des participants, les variations des caractéristiques EMG sont relativement uniformes. Cette observation pourrait indiquer que les effets de l'ARSACS sur ces caractéristiques ne sont pas suffisamment prononcés pour permettre une distinction claire dans l'espace des caractéristiques sélectionnées. Ainsi, ce cluster pourrait illustrer un cas où les signaux EMG ne reflètent pas significativement les variations des états de santé, ou que les méthodes actuelles de caractérisation des signaux EMG ne sont pas suffisamment précises pour détecter ces variations de manière distincte.

4.4 CLUSTERING POUR PARTICIPANTS MARCHEURS SEULEMENT

Dans cette approche, l'ensembe de données utilisé concerne les données aux temps T3, T4 et T5. Comme indiqué dans le chapitre précédent, pour chaque participant, trois itérations sont enregistrées à chacun de ces trois temps. Ainsi, il est possible qu'un participant dispose de jusqu'à neuf itérations, réparties entre les trois temps, bien que certains participants n'aient pas été présents à tous les temps de mesure. Chaque itération constitue une instance distincte au sein de l'ensemble de données. L'ensemble de données contient les caractéristiques suivantes : la durée, l'amplitude, les données des aires d'activités musculaires des muscles biceps fémoral (*biceps femoris*), droit fémoral (*rectus femoris*), droit de l'abdomen (*rectus abdominis*), tenseur du fasia lata (*tensor fascia lata*), grand glutéal (*gluteus maximus*), long adducteur (*adductor longus*) et tibial antérieur (*tibialis anterior*), et ce pour la phase concentrique de l'extension et de la flexion. L'ensemble de données inclut des données cliniques, notamment l'âge, le sexe et le niveau de mobilité intérieure. Il inclut également les mesures d'évaluation suivantes : le score de Berg, l'amplitude active en extension et en flexion, la vitesse du participant lors du test de 10 mètres de marche à vitesse confortable et à vitesse maximale ainsi que l'indice de sévérité de la maladie.

L'ensemble de données contient initialement la caractéristique représentant l'état du participant, une variable catégorielle de trois catégories : *healthy* pour un participant sain, *walker* pour un participant atteint d'ARSACS marcheur et *non-walker* pour un participant atteint d'ARSACS non-marcheur. Dans le cas de cette approche, seulement les participants de la catégorie *walker* sont pris en considération. Les participants des deux autres catégories sont donc supprimés de l'ensemble de données. Les participants restants (marcheurs) sont divisés en quatre sous-catégories. Cette segmentation est fondée sur le score de Berg associé à chaque instance : la sous-catégorie *walker-1* est attribuée aux instances pour lesquelles le score de Berg est inférieur à 20, *walker-2* pour celles dont le score se situe entre 20 (inclus)

et 30 (inclus), *walker-3* pour celles dont le score se situe entre 30 et 45 (inclus), et *walker-4* pour celles avec un score supérieur à 45. La caractéristique représentant le score de Berg est ensuite supprimée afin de garantir l'objectivité de l'analyse de *clustering* et éviter tout biais potentiel. La caractéristique représentant les sous-catégories, soient les états des participants, est également séparée de l'ensemble de données et conservée. Ceci permettra d'évaluer ultérieurement la pertinence des clusters formés en établissant un lien entre ces derniers et les sous-catégories.

Dans cette approche, deux scénarios distincts sont explorés. Dans le premier scénario, toutes les caractéristiques de l'ensemble de données sont prises en en compte, à l'exception de la caractéristique relative au sexe, qui a été volontairement exclue pour cette expérimentation. Dans le second, seules certaines caractéristiques spécifiques sont analysées : les données des aires d'activités musculaires des muscles pendant la phase concentrique des mouvements d'extension et de flexion, ainsi que l'âge et le degré de sévérité de la maladie pour chaque participant. Comparer les résultats obtenus dans ces deux scénarios permettra d'évaluer l'impact de chaque ensemble de caractéristiques sur les résultats de l'analyse et de comprendre comment la réduction des dimensions affecte les motifs identifiés dans les données.

Les étapes de prétraitement de données appliquées dans cette approche sont identiques à celles appliquées dans les approches précédentes, et ce pour les deux scénarios. Un nettoyage de données a été effectué afin de gérer les valeurs manquantes. Les instances comportant des valeurs manquantes ont été supprimées afin de conserver la cohérence des données. Après cette étape, le premier scénario comporte 240 instances, tandis que le deuxième en comprend 246. Une normalisation des données a été appliquée ensuite dans le but que chaque caractéristique contribue de manière équivalente. Pour ce faire, la technique *Min-Max Scaling* a été utilisée du à la nature des données et à l'absence de valeurs abbérantes.

Les coefficients de corrélation entre chaque paire de variables sont calculés afin de déterminer les relations entre les variables. Un seuil de 0.8, en valeur absolue, a été fixé pour identifier les corrélations très fortes. Les paires de caractéristiques dépassant ce seuil seront considérées comme fortement corrélées. Pour chaque paire identifiée, la caractéristique correspondante est ajoutée à un ensemble, permettant ainsi de collecter uniquement les caractéristiques qui présentent une forte corrélation avec au moins une autre caractéristique, sans doublons. Les caractéristiques dans la liste sont ensuite supprimées de l'ensemble de données.

Dans le premier scénario, où l'ensemble de données inclut toutes les caractéristiques à l'exception de la caractéristique relative au sexe, une seule caractéristique présentant une forte corrélation a été sélectionnée : celle représentant le test de marche de 10 mètres à vitesse maximale. Cette dernière a été supprimée de l'ensemble de données. La matrice de corrélation de ce scénario est présentée dans la Figure 4.9, illustrant cette forte corrélation.

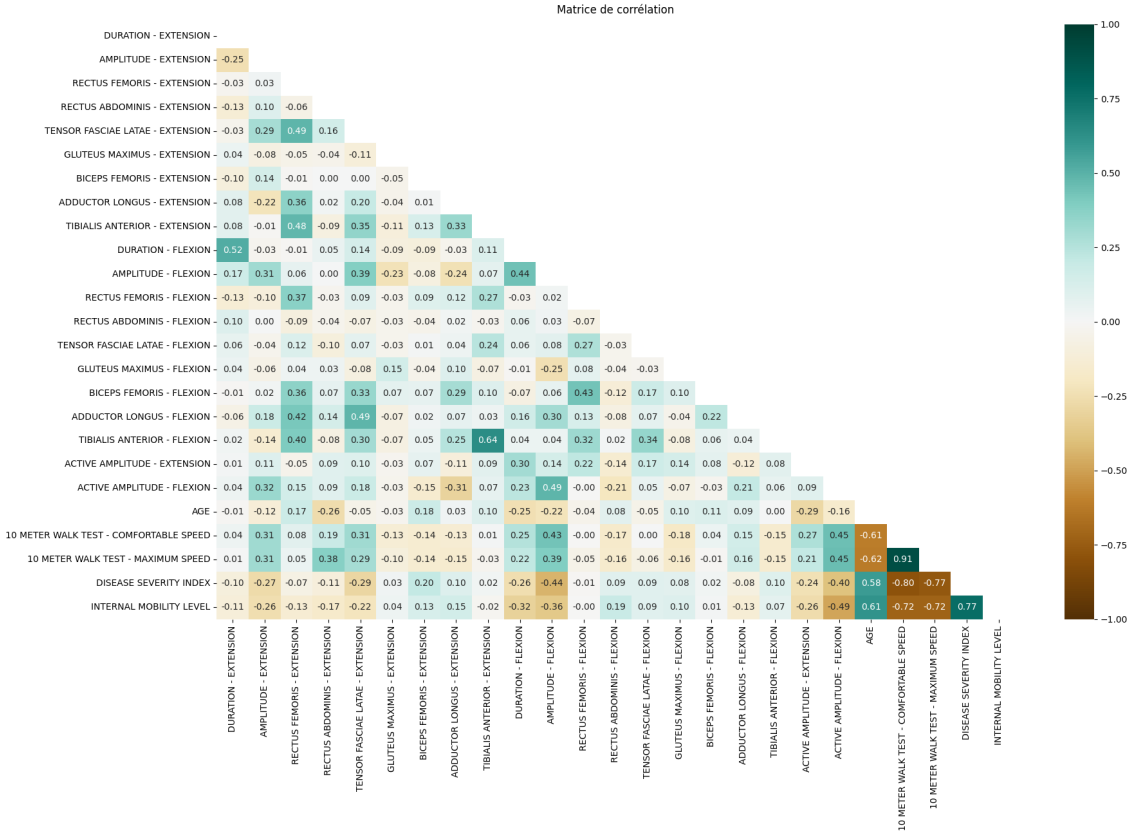

FIGURE 4.9 : Matrice de corrélation pour le premier scénario.

Dans le deuxième scénario, où l'ensemble de données inclut seulement une sélection de caractéristiques, aucune des caractéristiques n'est fortement corrélée. Cela était attendu puisque la caractéristique présentant une forte corrélation dans l'ensemble de données incluant toutes les caractéristiques, à l'exception de la caractéristique relative au sexe, n'est pas incluse dans cet ensemble réduit. La Figure 4.10 montre la matrice de corrélation pour ce deuxième scénario, confirmant l'absence de corrélation élevée entre les variables sélectionnées.

Cette démarche permet d'éviter le problème de colinéarité et de s'assurer que les variables incluses dans les analyses apportent des informations uniques, optimisant ainsi la validité et la fiabilité des résultats obtenus.

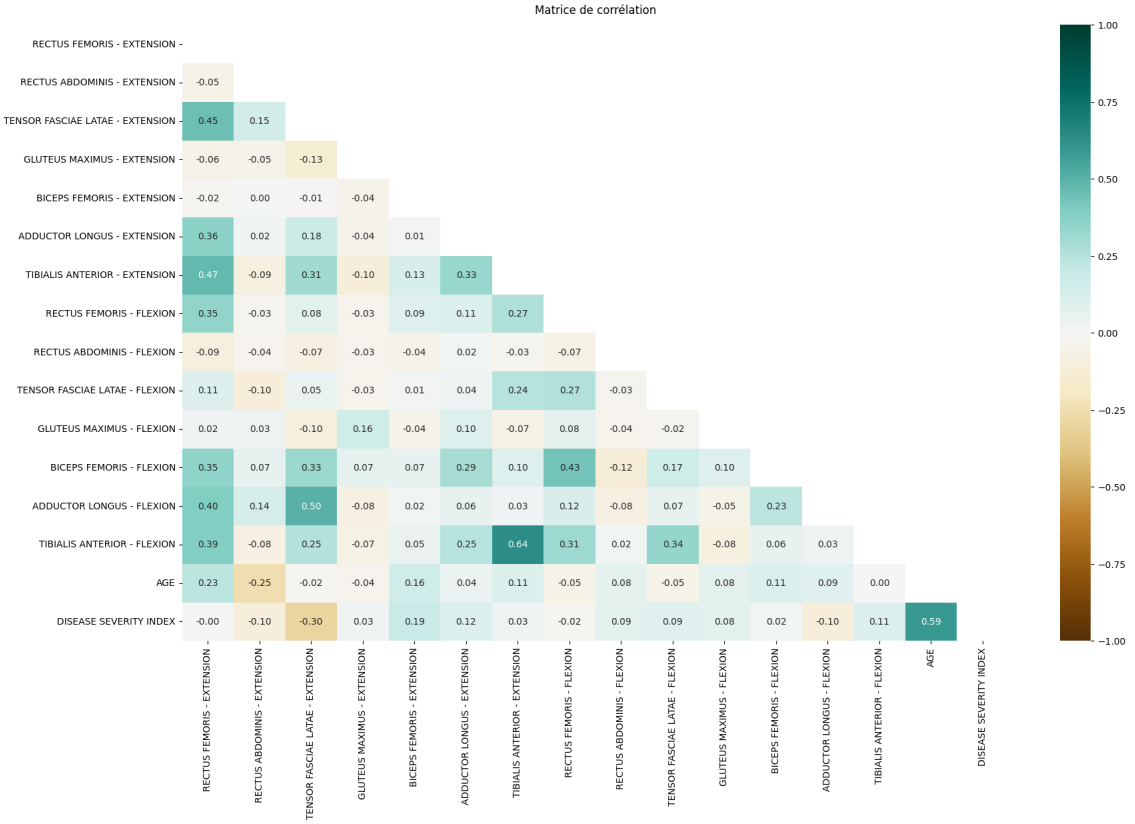

FIGURE 4.10 : Matrice de corrélation pour le deuxième scénario.

Afin de déterminer le nombre optimal de clusters dans le cas de cette expérimentation, la méthode du coude est utilisée. La courbe de cette dernière indique un nombre optimal de quatre clusters. L'algorithme *K-means* a été utilisé pour le *clustering* des données. Pour évaluer la qualité des groupements et faciliter la présentation des résultats, il a été nécessaire de procéder à une réduction de dimensionnalité. Deux techniques ont été employées : l'analyse en composantes principales et l'algorithme t-SNE.

La Figure 4.11 illustre la dispersion des clusters résultant de l'application de l'algorithme *K-means* avec quatre clusters. Les données sont réduites en deux dimensions via PCA, et sont représentées par deux axes correspondant aux deux premières composantes principales qui

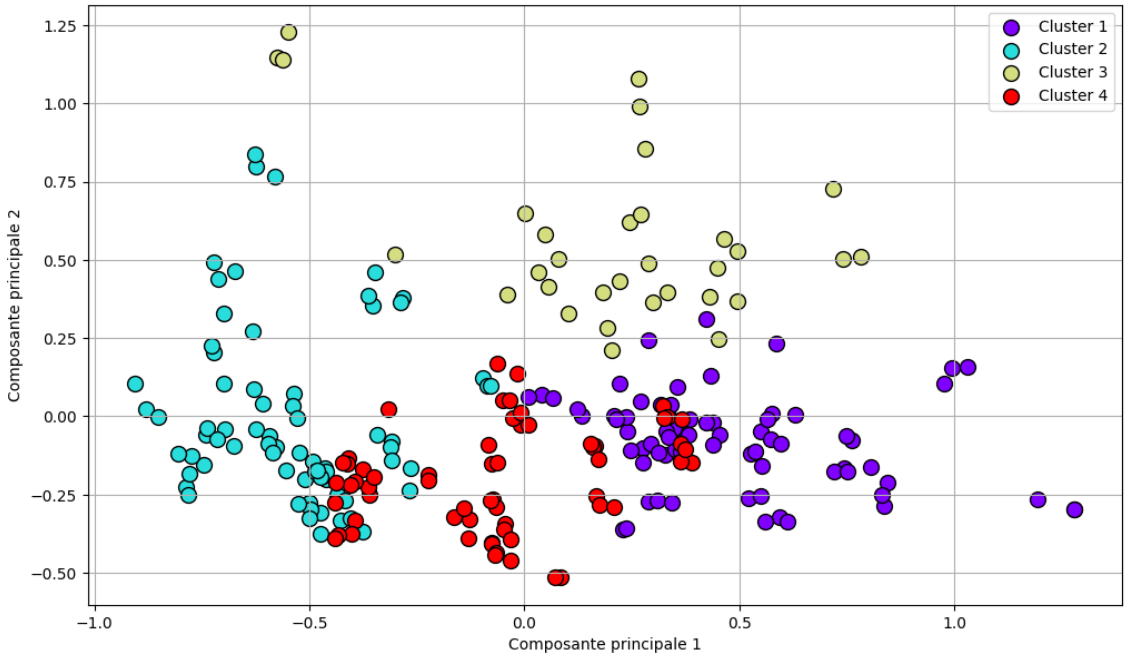

FIGURE 4.11 : Représentation bidimensionnelle des clusters obtenus en utilisant PCA pour le premier scénario.

capturent la plus grande variance des données. Les clusters sont colorés différemment pour faciliter la distinction.

La répartition des clusters indique une bonne distinction dans l'espace des composantes principales. Cela suggère que l'algorithme *K-means* a trouvé des groupements significatifs dans les données multidimensionnelles réduites à deux dimensions principales.

Les Clusters 1 et 3, de couleurs violet et jaune respectivement, semblent plus homogènes et bien définis, suggérant qu'ils regroupent des participants aux caractéristiques physiologiques et cliniques similaires. Cette similarité pourrait indiquer des états de santé qui sont plus uniformes. Par exemple, un des deux clusters pourrait représenter des participants avec une mobilité et un équilibre relativement préservée, tandis que l'autre cluster pourrait inclure ceux qui sont plus avancés dans leur condition, avec des limitations motrices plus sévères. À

l'opposé, les Clusters 2 et 4, colorés en cyan et rouge, sont plus dispersés, indiquant qu'ils peuvent regrouper des participants présentant une diversité plus large de conditions cliniques. Ces clusters pourraient inclure des participants en début ou dans un état intermédiaire de la maladie. Cette diversité pourrait également refléter des variations dans d'autres caractéristiques importantes telles que l'âge.

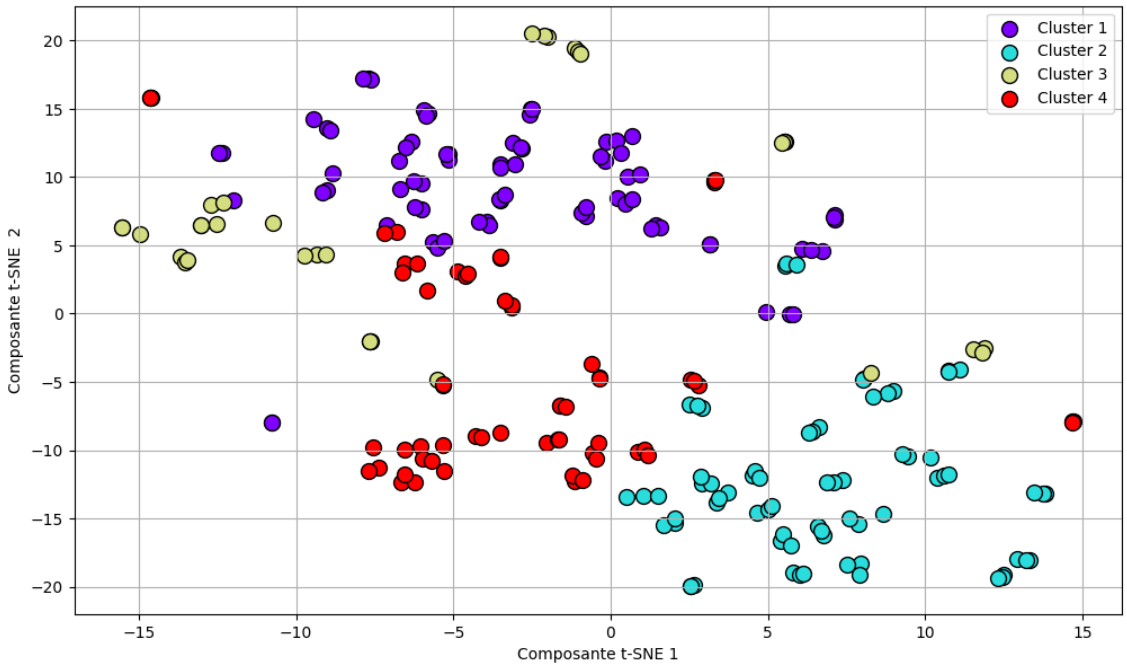

FIGURE 4.12 : Représentation bidimensionnelle des clusters obtenus en appliquant l'algorithme t-SNE pour le premier scénario.

La Figure 4.12 présente une visualisation 2D des clusters résultant de l'application de l'algorithme *K-means* avec quatre clusters, en utilisant l'algorithme t-SNE pour la réduction de dimension. Les clusters sont colorés de la même manière que dans la visualisation PCA précédente présentée dans la Figure 4.11, et ce pour faciliter la comparaison.

Les clusters sont plutôt bien définis et montrent une séparation claire, avec des groupes distincts et peu de chevauchement visible entre eux, bien qu'ils semblent plus étendus que

ceux observés dans la visualisation PCA présentée dans la Figure 4.11. Cela suggère que les différences entre les groupes dans l'ensemble de données sont marquées et que l'algorithme t-SNE a pu les capturer efficacement, reflétant des distinctions significatives dans les données de haute dimension. Cela peut être dû à des différences intrinsèques assez fortes entre les clusters qui se traduisent bien même après la réduction de dimension réalisée par t-SNE.

La visualisation t-SNE présente une répartition claire des clusters, avec une bonne séparation pour les clusters 2 (cyan) et 4 (rouge), bien que quelques chevauchements soient observés, notamment avec le cluster 1 (violet). Le Cluster 3 (jaune) présente une séparation moins nette et chevauche les clusters 1 et 4, ce qui suggère des similitudes entre certains individus de ces groupes. Cependant, le Cluster 1 montre une plus grande dispersion, couvrant une large gamme dans les deux dimensions t-SNE, ce qui pourrait indiquer une certaine hétérogénéité parmi les individus qui le composent. Cette dispersion pourrait expliquer le chevauchement observé avec les clusters 2 et 4. Le Cluster 2, quant à lui, semble être davantage regroupé dans la partie inférieure droite du graphique, ce qui pourrait suggérer que les individus de ce groupe partagent des similarités plus fortes par rapport aux autres clusters. Ces résultats montrent que les différences entre les clusters sont globalement bien représentées, bien qu'une analyse plus approfondie soit nécessaire pour comprendre les chevauchements observés.

La comparaison avec les résultats obtenus via PCA montre que ces deux méthodes sont complémentaires. Le PCA fournit une vue d'ensemble statistique plus globale, tandis que t-SNE est particulièrement utile pour capturer des séparations locales entre les clusters et révéler des relations non linéaires moins visibles avec une réduction de dimension linéaire.

Afin de mieux comprendre les clusters identifiés et d'assurer une interprétation concrète, il est crucial d'examiner les relations potentielles entre ces clusters et les différentes catégories de marcheurs définies par le score de Berg. La Figure 4.13 illustre la répartition des différentes

catégories de marcheurs (*walker-1*, *walker-2*, *walker-3* et *walker-4*) à travers les quatre clusters identifiés.

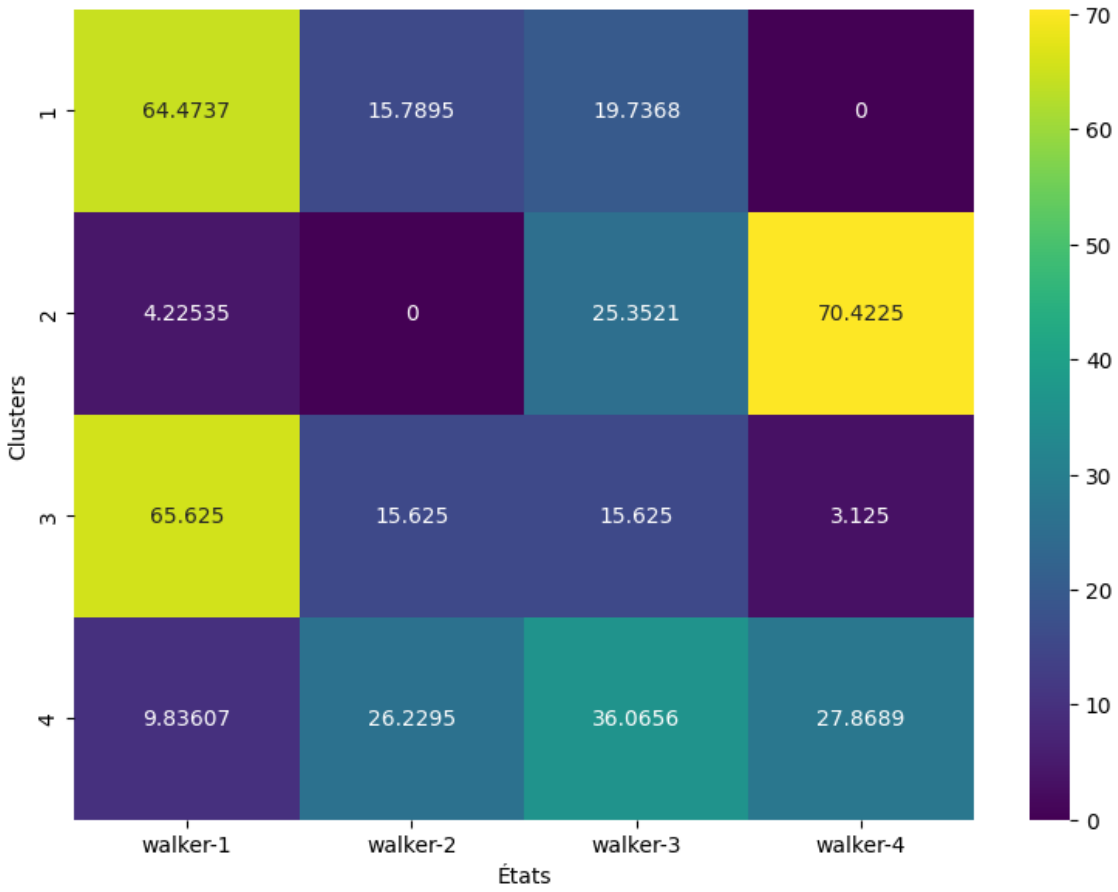

FIGURE 4.13 : Répartition des sous-catégories de marcheurs dans les clusters pour le premier scénario.

Le Cluster 1 est largement dominé par les marcheurs de type *walker-1*, avec une présence également notable de *walker-2* et *walker-3*, et une absence totale de *walker-4*. Cela qui indique que ce cluster regroupe principalement des participants ayant un équilibre relativement faible. En revanche, le Cluster 2, dominé par *walker-4* et inclut également *walker-3* avec, relativement, une très faible présence de *walker-1* et une absence totale de *walker-2*, suggère que ce cluster

inclut principalement des participants avec de bonnes à excellentes capacités motrices. Le Cluster 3, majoritairement constitué de participants de type *walker-1*, tout comme le Cluster 1. Cela indique que les participants de ce groupe ont également un équilibre faible et des limitations motrices importantes. Il inclut aussi des proportions égales de *walker-2* et *walker-3*, ce qui indique une certaine diversité parmi les participants, bien que *walker-4* soit faiblement représenté. Ce cluster regroupe donc des participants ayant principalement des limitations motrices sévères, mais certains d'entre eux ont des capacités légèrement meilleures. Enfin, le Cluster 4 est relativement diversifié, avec une proportion notable de *walker-3*, suivie de *walker-4* et *walker-2*. Ce cluster inclut également une petite proportion de *walker-1*. Cela suggère que les participants de ce groupe présentent une meilleure mobilité dans l'ensemble, bien que certains d'entre eux aient un équilibre plus faible. La diversité dans ce cluster montre que les capacités motrices des participants varient plus largement comparé aux Clusters 1 et 3, qui sont dominés par des marcheurs avec des limitations plus importantes. Globalement, cela démontre la complexité et l'hétérogénéité de la maladie, reflétant la grande variabilité des caractéristiques cliniques et des mesures EMG parmi les patients.

Dans le deuxième scénario, où seules certaines caractéristiques spécifiques, présentées dans la Figure 4.10, sont incluses dans l'ensemble de données, l'algorithme *K-means* a été utilisé pour le *clustering*, tout comme dans le premier scénario. De même, une réduction de dimensionnalité est nécessaire, réalisée à l'aide des techniques PCA et t-SNE.

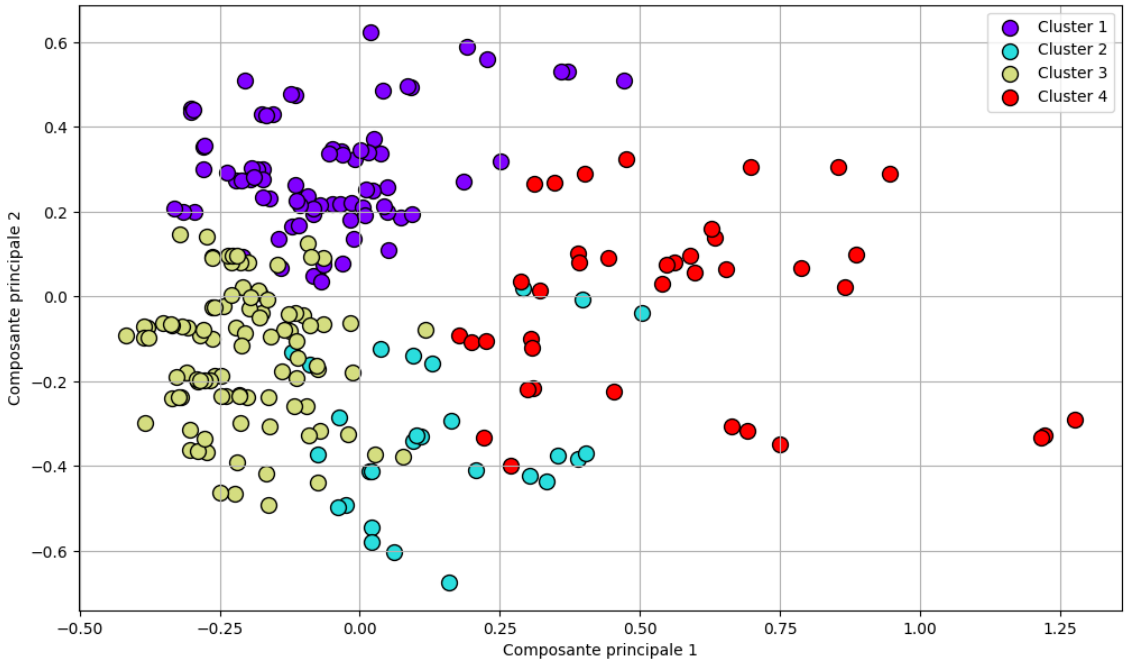

FIGURE 4.14 : Représentation bidimensionnelle des clusters obtenus en utilisant PCA pour le deuxième scénario.

La Figure 4.14 présente une visualisation 2D obtenue par l'analyse en composantes principales des clusters formés par l'algorithme *K-means*, appliquée au deuxième scénario de cette expérimentation. Les clusters sont colorés différement afin de faciliter leur distinction.

Le Cluster 1 présente une homogénéité modérée, avec une légère dispersion entre les participants, mais les points restent globalement concentrés. Cela indique que les participants de ce groupe partagent des caractéristiques cliniques et physiologiques relativement similaires, malgré quelques variations. Le Cluster 2 montre une plus grande dispersion des points, ce qui suggère une plus grande diversité parmi les participants. Cette dispersion pourrait refléter une diversité plus large des conditions cliniques des participants inclus dans ce groupe. Le Cluster 3, quant à lui, regroupe des points proches les uns des autres, concentrés dans une région spécifique de la représentation. Ce cluster présente une variabilité relativement faible, suggérant une homogénéité marquée des caractéristiques des participants qu'il regroupe.

Cette concentration pourrait indiquer des caractéristiques spécifiques associées à des formes cliniques moins sévères de l'ARSACS. Le Cluster 4 montre également une plus grande dispersion des points, ce qui suggère une plus grande diversité parmi les participants. Cette dispersion pourrait également refléter une diversité plus large des conditions cliniques des participants inclus dans ce groupe.

La séparation claire entre les clusters suggère que même un sous-ensemble restreint de caractéristiques peut capturer efficacement des différences significatives entre les groupes de participants. Cela peut indiquer que les caractéristiques choisies pour ce scénario sont particulièrement pertinentes pour discerner des variations cliniquement significatives.

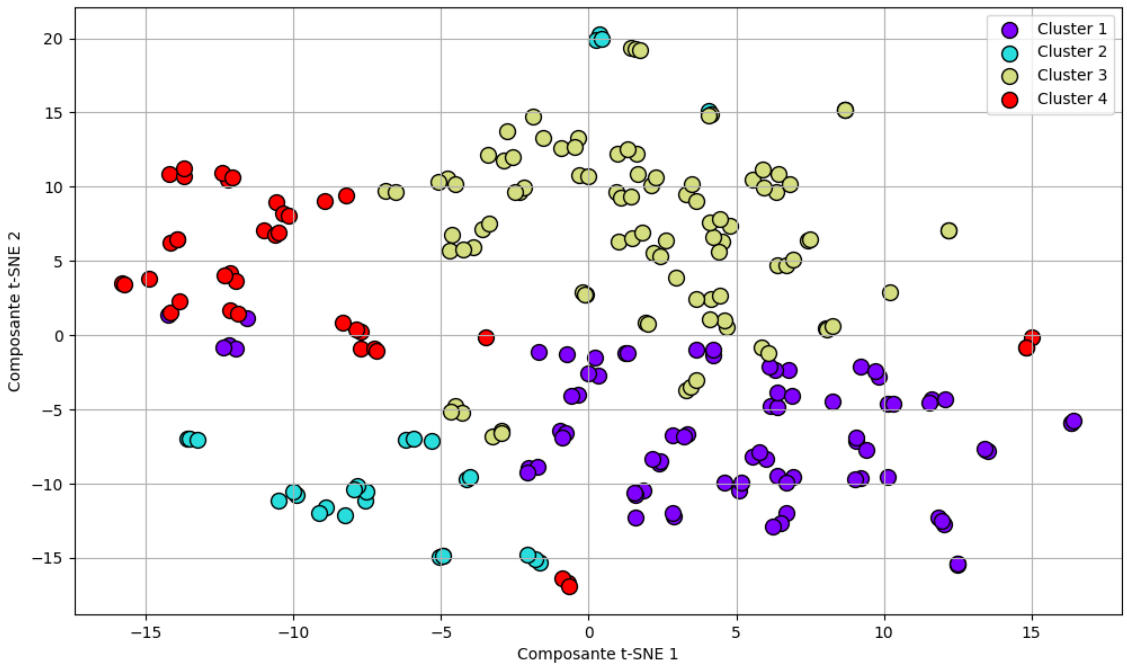

FIGURE 4.15 : Représentation bidimensionnelle des clusters obtenus en utlisant t-SNE pour le deuxième scénario.

La Figure 4.15 illustre une visualisation 2D des clusters formés par l'algorithme *K-means*, obtenue grâce à l'utilisation de l'algorithme t-SNE sur l'ensemble de données du

deuxième scénario de cette expérimentation. Cette visualisation montre une bonne séparation des clusters, capturant à la fois des groupes homogènes et d'autres plus dispersés, indiquant une hétérogénéité parmi certains participants du même groupe. Les points isolés suggèrent des profils atypiques ou des cas cliniques particuliers. Globalement, cette représentation révèle des relations non linéaires entre les données, avec des regroupements cohérents tout en mettant en évidence la diversité des profils cliniques dans certains clusters.

Les relations entre les clusters et les différentes catégories de marcheurs définies par le score de Berg ont été examinées comme dans le premier scénario de cette expérimentation, afin de pouvoir interpréter les résultats obtenus de façon concrète. La Figure 4.16 montre la répartition des différentes catégories de marcheurs à travers les quatre clusters identifiés dans le cadre de ce deuxième scénario.

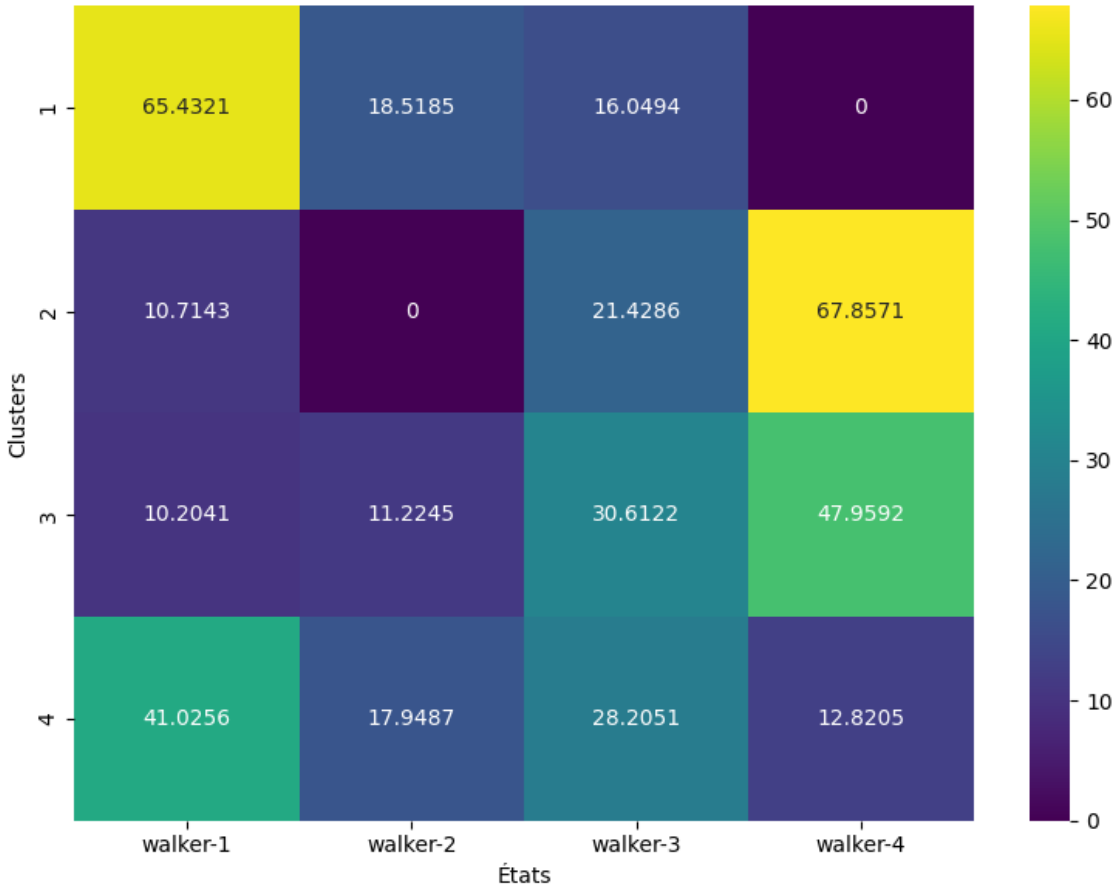

FIGURE 4.16 : Répartition des sous-catégories de marcheurs dans les clusters pour le deuxième scénario.

Dans le cas de ce deuxième scénario, le Cluster 1 est majoritairement composé de participants de la catégorie *walker-1* avec quelques représentants des catégories *walker-2* et *walker-3*. Il n'y a aucun participant de la catégorie *walker-4*, ce qui indique que ce cluster regroupe des participants ayant un équilibre relativement faible. Le Cluster 2 est dominé par les participants de la catégorie *walker-4* qui ont de bonnes à excellentes capacités motrices ainsi qu'une proportion notable de participants de la catégorie *walker-3*. Ce cluster inclut également une moindre présence de participants des catégories *walker-1* et une absence totale de *walker-2*, suggérant que ce cluster inclut principalement des participants avec de bonnes à excellentes capacités motrices. Le Cluster 3 est relativement diversifié, avec une proportion

notable de *walker-4*, suivie de *walker-3*. Ce cluster comprend également une petite proportion de *walker-1* et *walker-2*, ce qui suggère que les participants de ce groupe présentent une meilleure mobilité dans l'ensemble, bien que certains d'entre eux aient un équilibre plus faible. Le Cluster 4 est également hétérogène, avec une proportion significative de *walker-1*, suivie de *walker-3*. Ce cluster inclut également une moindre proportion de *walker-2* et *walker-4*, représentant ainsi un groupe de participants ayant des niveaux variés d'équilibre et de mobilité, bien que ceux avec des limitations soient majoritaires.

Dans le premier scénario, où toutes les caractéristiques étaient incluses à l'exception de la caractéristique relative au sexe, les clusters formés étaient nettement séparés, avec des frontières relativement claires. Cela illustre la capacité de l'ensemble complet des données à distinguer de manière significative entre différents profils de participants. En revanche, dans le deuxième scénario, utilisant un sous-ensemble restreint de caractéristiques spécifiques, certains clusters présentent une dispersion plus importante, avec une certaine confusion entre les niveaux d'équilibre dans certains clusters. Bien que les caractéristiques choisies pour ce scénario soient discriminantes, elles ne suffisent pas à distinguer clairement tous les niveaux d'équilibre des participants.

Cette comparaison souligne l'importance de sélectionner des caractéristiques qui sont non seulement représentatives des variations cliniques significatives, mais aussi suffisamment complètes pour assurer une discrimination efficace entre les différents états de santé. Dans le cas de l'ARSACS, cela signifie qu'il est essentiel d'inclure une variété de caractéristiques pertinentes pour capturer pleinement les distinctions entre les différents profils de participants.

4.5 CONCLUSION

Dans ce chapitre, nous avons présenté les différentes expérimentations de *clustering* menées. Les résultats obtenus de ces dernières montrent l'efficacité des approches de *clustering* a révéler des groupements inédits de profils de patients atteints d'ARSACS, permettant d'identifier des sous-groupes avec des caractéristiques cliniques distinctes. En limitant les caractéristiques étudiées à trois muscles, les clusters montrent une séparation claire, suggérant que même un sous-ensemble restreint de caractéristiques peut capturer des distinctions significatives, facilitant les visualisations tridimensionnelles. L'analyse des participants marcheurs a montré que l'inclusion de toutes les caractéristiques permet une meilleure discrimination des états de santé par rapport à un sous-ensemble de caractéristiques spécifiques. Les résultats soulignent l'importance d'une sélection complète et pertinente des caractéristiques pour capturer pleinement les distinctions entre les différents états de santé des patients atteints d'ARSACS. Les clusters obtenus montrent une certaine variabilité des états de santé au sein de chaque groupe, reflétant la complexité de la maladie d'ARSACS et la difficulté de distinguer les états de santé uniquement sur la base des caractéristiques EMG.

CHAPITRE V
EXPÉRIMENTATIONS ET RÉSULTATS DE RÉGRESSION

5.1 INTRODUCTION

La régression pourrait potentiellement jouer un rôle important dans la prédiction de l'évolution de l'ARSACS chez les patients. Dans cette étude, nous avons employé des modèles de régression pour tenter de prédire et d'estimer la progression de l'ARSACS, ainsi que pour évaluer la précision de ces prédictions. Cette approche inclut la prédiction ponctuelle de certains scores cliniques. Disposant de données recueillies à plusieurs moments, notre méthodologie permet également d'explorer des modèles afin d'évaluer la possibilité de prédire ces scores cliniques à des instants futurs (T+1), anticipant ainsi la progression de la maladie. Cette approche prédictive s'appuie sur des modèles de régression qui intègrent les données longitudinales des patients, en visant à obtenir des prévisions précises. Ces modèles permettent de tenir compte des variations individuelles et de l'évolution temporelle des symptômes. Les différentes expérimentations de régression menées et les résultats obtenus sont présentés dans ce chapitre.

5.2 PRÉDICTION DU SCORE DU TEST DE COORDINATION MOTRICE DES MEMBRES INFÉRIEURS

L'objectif principal de cette expérimentation est de prédire le score du test de coordination motrice des membres inférieurs des participants atteints d'ARSACS en utilisant divers modèles de régression pour identifier le plus performant. Prédire efficacement ce score pourrait aider à mieux comprendre la progression de l'ARSACS et à adapter les traitements et interventions pour les patients, en fonction de l'évolution prévue de leur condition.

L'ensemble de données utilisé dans cette expérimentation concerne les données aux temps T3, T4 et T5 en phase concentrique et inclut initialement les caractéristiques suivantes : la durée, l'amplitude, les données des aires d'activités musculaires des muscles biceps fémoral (*biceps femoris*), droit fémoral (*rectus femoris*), droit de l'abdomen (*rectus abdominis*), tenseur du fasia lata (*tensor fascia lata*), grand glutéal (*gluteus maximus*), long adducteur (*adductor longus*) et tibial antérieur (*tibialis anterior*), et ce pour la phase concentrique de l'extension et de la flexion. L'ensemble de données inclut des données cliniques, notamment l'âge, le sexe et le niveau de mobilité intérieure. Il inclut également les mesures d'évaluation suivantes : le score de Berg, l'amplitude active en extension et en flexion, le test de 10 mètres de marche à vitesse confortable et à vitesse maximale, l'indice de sévérité de la maladie ainsi que le test de coordination motrice des membres inférieurs droit et gauche. L'ensemble de données contient initialement les instances de tous les participants (sains, marcheurs et non-marcheurs).

Pour garantir la qualité et la pertinence des analyses de cette expérimentation, seules les instances des participants catégorisés comme *marcheurs* sont retenues. Cette sélection est justifiée par le fait que les caractéristiques de la motricité des marcheurs sont plus susceptibles de fournir des indices pertinents sur les effets de l'ARSACS sur la coordination motrice. Les variables considérées comprennent les données des aires d'activités musculaires des muscles pendant la phase concentrique des mouvements d'extension et de flexion, ainsi que l'âge, le sexe, et le score au test de coordination motrice des membres inférieurs droit pour chaque participant. De plus, les instances présentant des valeurs manquantes sont éliminées pour préserver l'intégrité des données. Ainsi, l'ensemble final de données utilisé pour cette expérimentation est composé de 326 instances.

Après le nettoyage des données, celles-ci sont normalisées à l'aide de la technique de *Min-Max Scaling* pour garantir que les variables sont à une échelle comparable. Cette étape est essentielle pour assurer que toutes les variables contribuent équitablement à l'ana-

lyse de régression, permettant une interprétation plus fiable des poids attribués à chaque caractéristique.

La variable cible, qui dans le cas de notre expérimentation représente le score du test de la coordination motrice du membre inférieur droit, a été soigneusement séparée du reste des variables. Cette démarche assure que les modèles prédictifs sont construits uniquement à partir des caractéristiques prédictives (variables indépendantes) sans être influencés par la variable à prédire (variable cible). Cette séparation est cruciale pour maintenir l'intégrité statistique de l'analyse et pour éviter tout risque de fuite d'informations entre les variables explicatives et la variable cible.

Avant de procéder à la construction d'un modèle de régression, il est essentiel d'examiner la corrélation entre les variables. À cette fin, une matrice de corrélation est calculée, comme illustré dans la Figure 5.1, permettant de visualiser les coefficients de corrélation entre chaque paire de variables. Aucun de ces coefficients ne dépasse le seuil fixé à 0.8 en valeur absolue, indiquant ainsi qu'aucune variable ne présente de forte corrélation avec une autre.

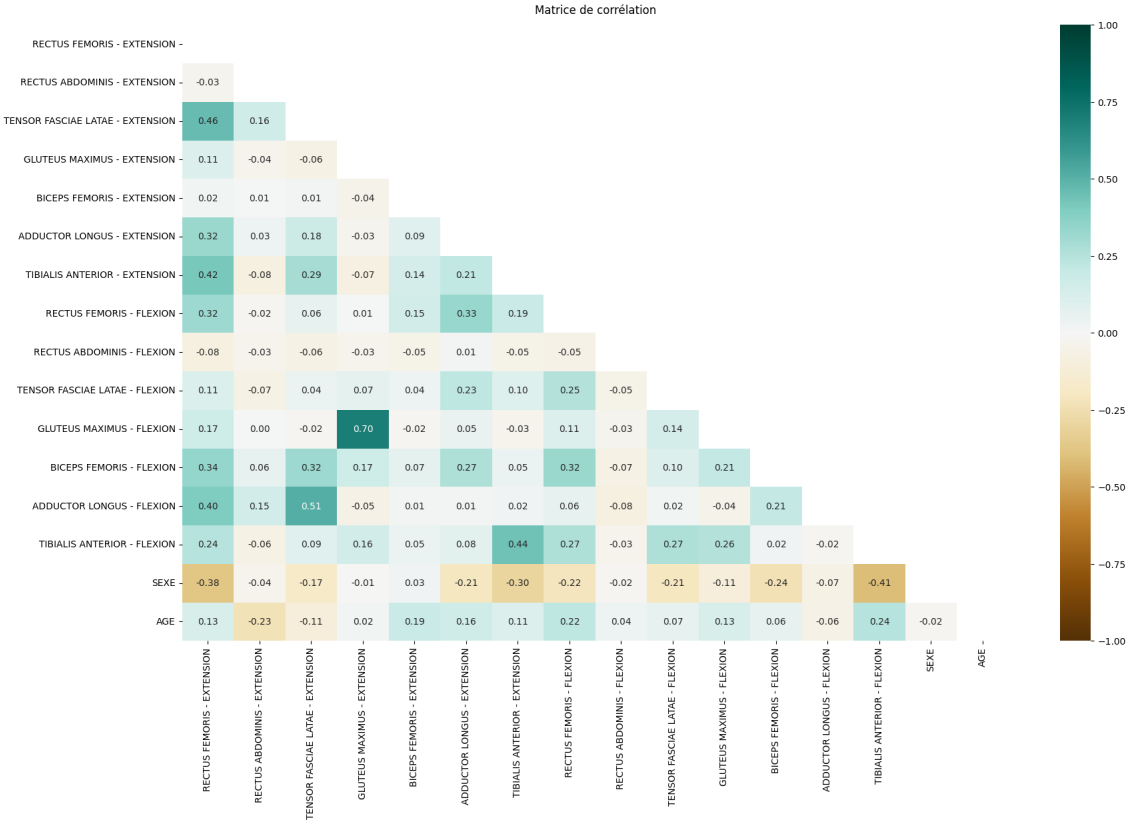

FIGURE 5.1 : Matrice de corrélation.

Pour une analyse plus complète et fiable de la multicolinéarité, le facteur d'inflation de la variance (*Variance Inflation Factor* ou VIF), une mesure statistique qui évalue le degré de multicolinéarité dans un modèle de régression, indiquant à quel point une variable indépendante est corrélée avec les autres variables indépendantes du modèle [38], est calculé pour chaque variable indépendante, comme illustré dans la Figure 5.2. Étant donné que tous les VIFs des variables indépendantes sont supérieurs à 1 et inférieurs à 5, cela indique une corrélation modérée avec les autres variables, mais qui ne nécessite pas une attention particulière.

	variables	VIF
0	RECTUS FEMORIS - EXTENSION	2.062046
1	RECTUS ABDOMINIS - EXTENSION	1.132510
2	TENSOR FASCIAE LATAE - EXTENSION	1.801740
3	GLUTEUS MAXIMUS - EXTENSION	2.052291
4	BICEPS FEMORIS - EXTENSION	1.094036
5	ADDUCTOR LONGUS - EXTENSION	1.339019
6	TIBIALIS ANTERIOR - EXTENSION	1.690360
7	RECTUS FEMORIS - FLEXION	1.426300
8	RECTUS ABDOMINIS - FLEXION	1.028907
9	TENSOR FASCIAE LATAE - FLEXION	1.186573
10	GLUTEUS MAXIMUS - FLEXION	2.148306
11	BICEPS FEMORIS - FLEXION	1.444448
12	ADDUCTOR LONGUS - FLEXION	1.606246
13	TIBIALIS ANTERIOR - FLEXION	1.762683
14	SEXE	1.471370
15	AGE	1.275880

FIGURE 5.2 : Facteur d'inflation de la variance pour chaque variable indépendante.

Dans le cadre de cette expérimentation, l'ensemble des données a été divisé en deux sous-ensembles : un ensemble d'entraînement et un ensemble de test. Cela a été réalisé à l'aide de la fonction *train_test_split* de la bibliothèque *sklearn.model_selection*. Les données avaient été préalablement séparées en variables explicatives et en variable cible. Pour la division, 50 % des données ont été allouées à l'ensemble d'entraînement et les 50 % restants à l'ensemble de test, garantissant ainsi une répartition équilibrée. Cette division équilibrée est cruciale pour contrebalancer la présence de répétitions multiples pour chaque participant

à différents moments. Chaque participant est représenté à trois instants, avec trois itérations par instant. Cette structure particulière de l'ensemble de données nécessite une approche de division qui réduit les risques de surapprentissage ou de sous-apprentissage durant les phases d'entraînement et de validation du modèle. En d'autres termes, cette méthode de division permet de s'assurer que les modèles peuvent généraliser non seulement entre différents participants mais aussi à travers différentes mesures pour un même participant.

Pour analyser les données complexes associées à l'ARSACS, plusieurs modèles de régression ont été mis en œuvre, chacun offrant des perspectives uniques pour comparer et évaluer leurs performances. Les modèles utilisés incluent la régression linéaire, qui sert de référence standard en raison de sa simplicité et de sa transparence, le *Decision Tree Regressor* [39] et le *k-NN Regression* [40], qui modélisent des relations non linéaires entre les variables, une combinaison de *k-NN* et *Decision Tree* pour exploiter les avantages complémentaires de ces deux approches, ainsi que des modèles plus complexes tels que le *Random Forest Regression* [41] et les régresseurs avancés basés sur le boosting, *CatBoost* [42] et *XGBoost* [43].

Chaque modèle a la capacité de révéler des aspects distincts des données, améliorant ainsi notre capacité à identifier les marqueurs prédictifs les plus pertinents et à comprendre les interactions entre variables. Cette diversité méthodologique enrichit non seulement notre analyse mais assure également l'optimisation des modèles en termes de performance prédictive.

L'évaluation de ces modèles est réalisée à travers des métriques rigoureuses telles que le Coefficient de Détermination (R^2) et l'Erreur Absolue Moyenne (MAE), permettant une comparaison équitable et une sélection basée sur leur efficacité et précision.

Le coefficient de détermination mesure la proportion de la variance de la variable dépendante qui est prévisible à partir des variables indépendantes dans un modèle de régression

[38]. Le R^2 est une statistique qui indique dans quelle mesure les entrées du modèle se rapportent aux observations réelles. Un R^2 de 1 indique que le modèle prévoit parfaitement les données observées, tandis qu'un R^2 de 0 indique que le modèle ne prévoit pas mieux que la moyenne simple des données.

Le R^2 est particulièrement pertinent dans le contexte de l'ARSACS car il mesure la proportion de variance dans les données relatives au stade de la maladie que notre modèle est capable d'expliquer. Un R^2 élevé indique que le modèle capte efficacement la complexité et les variations du stade de l'ARSACS, ce qui est essentiel pour une prédiction fiable et pertinente. D'autre part, le MAE offre une mesure directe de l'erreur moyenne des prédictions du modèle par rapport aux valeurs observées. Dans le contexte de notre étude, un MAE faible est crucial car il signifie que les prédictions du modèle sont proches des valeurs cliniques observées, ce qui renforce la pertinence et la précision du modèle.

Les performances des modèles testés sont résumées dans le Tableau 5.1 ci-dessous, qui présente les valeurs de R^2 et de MAE pour chaque modèle de régression utilisé.

TABLEAU 5.1 : Comparaison des performances des modèles de régression

Type de modèle	**R^2**	**MAE**
Régression linéaire	0.25	0.12
Arbre de décision	0.11	0.13
Algorithme k-NN	0.45	0.11
Combinaison de k-NN et Arbre de décision	0.49	0.10
Forêt Aléatoire	0.53	0.10
CatBoost	**0.67**	**0.08**
XGBoost	0.43	0.12

Bien que simple et souvent utilisée comme point de départ pour les analyses de régression, la *régression linéaire* montre un R^2 modeste de **0.25**. Cette performance suggère que le score Test de Coordination Motrice des Membres Inférieurs (LEMOCOT) est difficilement

prédictible par des relations linéaires, soulignant la nécessité de modèles qui peuvent capturer des interactions non linéaires complexes présentes dans les données cliniques. Cela met en évidence les limites des approches linéaires face à la complexité biologique et l'hétérogénéité des patients atteints d'ARSACS, où les interactions variables sont critiques.

Le modèle d'*arbre de décision* présente le R^2 le plus faible (**0.11**), ce qui peut indiquer une susceptibilité au surajustement ou une incapacité à généraliser à partir des données d'entraînement, en l'absence de techniques de régularisation efficaces. Cette performance souligne les défis de l'utilisation des modèles simples pour capturer la complexité des données cliniques de l'ARSACS.

L'algorithme *k-NN*, avec un R^2 de **0.45**, montre une amélioration significative par rapport à la régression linéaire. Sa capacité à modéliser des relations non linéaires en utilisant les informations locales des voisins les plus proches le rend plus adapté à des ensembles de données avec des interactions complexes. Toutefois, il reste inférieur aux méthodes de boosting, soulignant ses limites dans le traitement des données de haute dimensionnalité ou fortement corrélées.

La combinaison de *k-NN* et l'*arbre de décision* améliore encore les performances, atteignant un R^2 de **0.49**. Cela démontre que la combinaison de différentes méthodologies peut être bénéfique pour capturer divers aspects de la variabilité des données, offrant ainsi un modèle plus robuste et mieux adapté aux données complexes de l'ARSACS.

Le modèle de *forêt aléatoire* offre une bonne performance avec un R^2 de **0.53**. Sa capacité à réduire le surajustement tout en conservant la puissance des arbres de décision en fait un choix robuste pour les ensembles de données complexes comme ceux de l'ARSACS. Cela illustre l'efficacité des méthodes d'ensemble dans la modélisation des données médicales.

L'algorithme *CatBoost* se distingue avec le meilleur R^2 de **0.67** et le MAE le plus bas de **0.08**. Ce modèle démontre une excellente capacité à intégrer et à modéliser la complexité des données, optimisant le traitement des caractéristiques catégorielles et réduisant le surajustement, ce qui le rend particulièrement efficace pour les analyses cliniques où la précision est critique. *CatBoost* excelle dans la prédiction du score LEMOCOT.

L'algorithme *XGBoost*, avec un R^2 de **0.43**, reste inférieur au *CatBoost*, suggérant que certains paramètres ou techniques spécifiques de traitement des données pourraient nécessiter des ajustements pour maximiser son potentiel.

L'analyse révèle que les modèles basés sur les techniques de boosting et les méthodes d'ensemble surpassent nettement les approches plus traditionnelles. La capacité de ces modèles à gérer les caractéristiques complexes et à modéliser des interactions non linéaires est cruciale pour la prédiction précise des scores cliniques tels que le LEMOCOT. Ces résultats indiquent la nécessité d'utiliser des modèles avancés pour une prédiction efficace. L'algorithme *CatBoost* est particulièrement prometteur pour prédire le score LEMOCOT.

5.3 PRÉDICTION DU SCORE DU TEST DE COORDINATION MOTRICE DES MEMBRES INFÉRIEURS À UN INSTANT T+1

L'objectif de cette nouvelle expérimentation est de tenter de prédire le score du test de coordination motrice des membres inférieurs pour des participants atteints d'ARSACS à l'instant T+1. La principale innovation de cette étude par rapport à la précédente (Section 5.2) réside dans l'intégration des scores LEMOCOT à l'instant T comme variables prédictives. Cette approche ajoute une dimension longitudinale à notre analyse, permettant de saisir les évolutions temporelles des capacités motrices, une dimension absente dans nos études antérieures. L'utilisation de données temporelles enrichit le modèle prédictif et pourrait

améliorer la précision des prédictions. En analysant les tendances de progression ou de régression des scores au fil du temps, nous pouvons obtenir des indices précieux sur la dynamique individuelle de la maladie. Cela est particulièrement pertinent dans le contexte de l'ARSACS, où la trajectoire de progression peut varier considérablement d'un patient à l'autre.

Comme indiqué dans le chapitre précédent, l'ensemble de données initial contient des mesures prises à trois moments différents : T3, T4, et T5. Pour chaque instant, il y a trois instances par participant, chacune correspondant à une itération distincte. Les données cliniques et les mesures d'évaluation, y compris le score LEMOCOT, restent toutefois identiques pour les trois instances à chaque instant, tandis que les données des aires d'activités musculaires varient entre les instances. À partir de ces informations, une nouvelle variable est créée pour représenter le score LEMOCOT à l'instant T+1. Ainsi, pour les instances recueillies à l'instant T3 d'un participant, cette nouvelle variable correspond au score LEMOCOT à l'instant T4 pour ce même participant. Cette opération est applicable uniquement pour les instances des instants T3 et T4, en raison de la disponibilité des données subséquentes.

Les instances concernant les participants sains ont été exclues de cette expérimentation, concentrant l'analyse sur ceux affectés par l'ARSACS. Les variables analysées incluent les données des aires d'activités musculaires de plusieurs muscles lors de la phase concentrique des mouvements d'extension et de flexion. Les muscles pris en compte sont le biceps fémoral (*biceps femoris*), droit fémoral (*rectus femoris*), droit de l'abdomen (*rectus abdominis*), tenseur du fasia lata (*tensor fascia lata*), grand glutéal (*gluteus maximus*), long adducteur (*adductor longus*) et tibial antérieur (*tibialis anterior*). D'autres variables prises en compte incluent l'âge et le sexe des participants, ainsi que le score du test de coordination motrice du membre inférieur droit. Ces variables constituent les caractéristiques prédictives, avec la variable cible étant le score LEMOCOT à l'instant T+1. Toutes les instances contenant au moins une valeur

manquante, y compris toutes celles de l'instant T5, ont été supprimées de l'analyse. Ainsi, l'ensemble final de données utilisé pour cette expérimentation est composé de 184 instances.

De la même manière que dans l'expérimentation précédente, la matrice de corrélation a été utilisée pour examiner les interdépendances entre les variables indépendantes (voir Figure 5.5). Elle révèle que les coefficients de corrélation entre les paires de variables restent en dessous du seuil fixé à 0.8, indiquant qu'aucune variable ne présente de forte corrélation avec une autre.

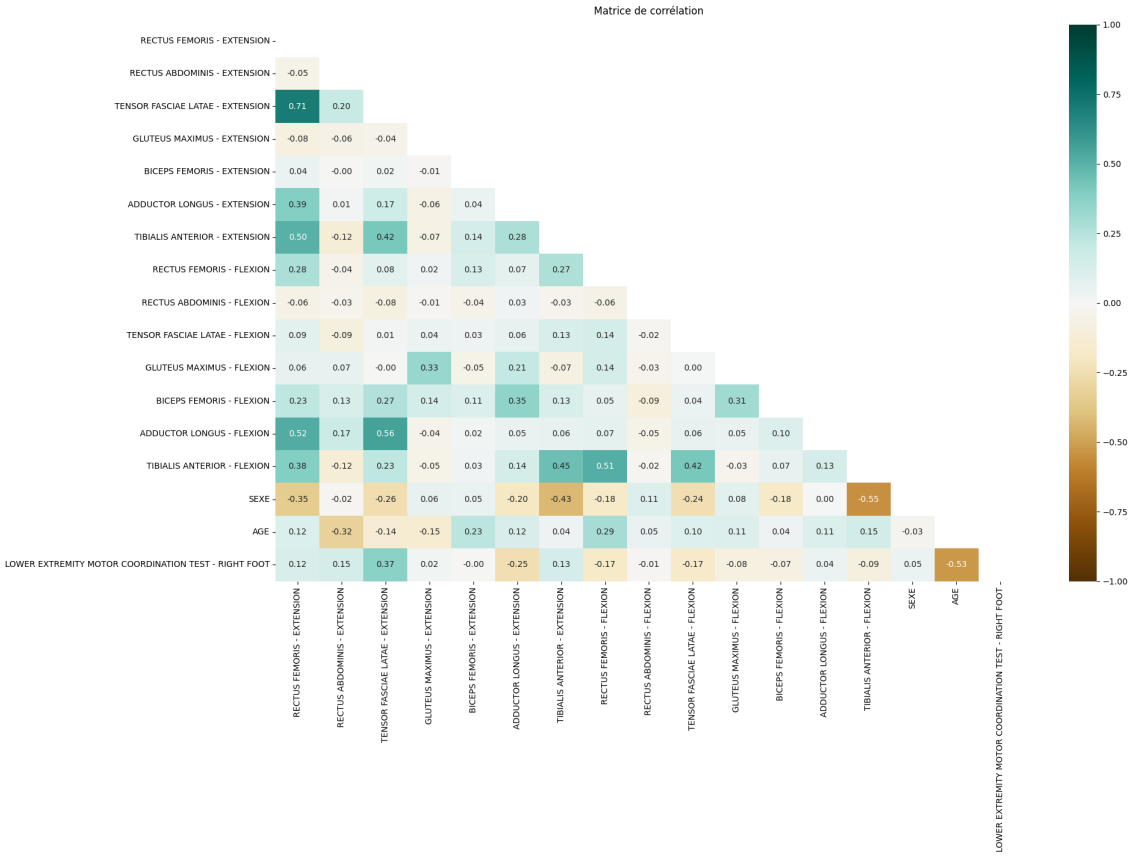

FIGURE 5.3 : Matrice de corrélation.

En complément, la multicolinéarité a été évaluée en calculant le facteur d'inflation de la variance pour chaque variable indépendante (voir Figure 5.6). Les valeurs des VIF indiquent une corrélation modérée, ce qui confirme les résultats de la matrice de corrélation.

	variables	VIF
0	RECTUS FEMORIS - EXTENSION	3.473584
1	RECTUS ABDOMINIS - EXTENSION	1.440678
2	TENSOR FASCIAE LATAE - EXTENSION	3.299563
3	GLUTEUS MAXIMUS - EXTENSION	1.266560
4	BICEPS FEMORIS - EXTENSION	1.180583
5	ADDUCTOR LONGUS - EXTENSION	1.662943
6	TIBIALIS ANTERIOR - EXTENSION	1.848572
7	RECTUS FEMORIS - FLEXION	1.679621
8	RECTUS ABDOMINIS - FLEXION	1.047616
9	TENSOR FASCIAE LATAE - FLEXION	1.272670
10	GLUTEUS MAXIMUS - FLEXION	1.409518
11	BICEPS FEMORIS - FLEXION	1.400765
12	ADDUCTOR LONGUS - FLEXION	2.039493
13	TIBIALIS ANTERIOR - FLEXION	2.427405
14	SEXE	1.824229
15	AGE	1.985988
16	LOWER EXTREMITY MOTOR COORDINATION TEST - RIGH...	1.993840

FIGURE 5.4 : Facteur d'inflation de la variance pour chaque variable indépendante.

Pour diviser l'ensemble des données en sous-ensembles d'apprentissage et de test, la fonction *train_test_split* de la bibliothèque *sklearn.model_selection* a été utilisée. Cette fonction partitionne aléatoirement les données en deux groupes : un pour l'apprentissage et l'autre pour les tests. Dans le cas de cette expérimentation, 60 % des données ont été allouées à l'ensemble d'apprentissage, tandis que les 40 % restants ont été réservés pour l'ensemble de test. Pour garantir la consistance des résultats entre différentes exécutions, la division

des données a été effectuée de manière à être reproductible. Cette approche assure que la séparation entre les ensembles d'apprentissage et de test reste constante, ce qui est essentiel pour la fiabilité et la validation des résultats expérimentaux.

Les modèles de régression utilisés dans cette expérimentation sont similaires à ceux de l'expérimentation précédente, à l'exception de *XGBoost* qui n'a pas été inclus cette fois-ci. L'évaluation de ces modèles s'est appuyée sur les mêmes métriques que celles utilisées précédemment, à savoir le coefficient de détermination et l'erreur absolue moyenne.

Le Tableau 5.2 présente une comparaison détaillée des performances de divers modèles de régression utilisés pour prédire le score du LEMOCOT à l'instant T+1 chez les participants atteints d'ARSACS. Il est essentiel de noter que les données utilisées n'ont pas été normalisées, ce qui peut influencer les valeurs absolues du MAE, reflétant les variations d'échelle parmi les variables prédictives.

TABLEAU 5.2 : Comparaison des performances des modèles de régression

Type de modèle	**R^2**	**MAE**
Régression linéaire	0.62	3.41
Arbre de décision	0.88	1.44
Algorithme k-NN	0.49	3.90
Combinaison de k-NN et Arbre de décision	0.88	2.15
Forêt Aléatoire	0.93	1.48
CatBoost	0.93	1.51

La régression linéaire, bien que fournissant une base pour la comparaison avec des méthodes plus complexes, montre un R^2 de 0.62 et un MAE de 3.41. Ces valeurs indiquent que, même si le modèle peut expliquer une part significative de la variance des scores LEMOCOT à l'instant T+1, l'erreur moyenne reste notable, suggérant que des prédictions plus précises pourraient être obtenues avec des modèles plus sophistiqués. L'arbre de décision se démarque avec un R^2 de 0.88 et un MAE de 1.44, montrant une capacité nettement améliorée à prédire

le score LEMOCOT par rapport à la régression linéaire. L'amélioration du R^2 et la réduction du MAE indiquent que ce modèle gère mieux les variations des données, ce qui en fait une option robuste pour cette analyse. Le modèle *k-NN* affiche un R^2 de 0.49 et un MAE de 3.90, ce qui reflète une performance limitée dans sa capacité à prédire le score LEMOCOT. Ce résultat indique que le modèle ne parvient pas à capturer efficacement la complexité des données, ce qui entraîne des prédictions moins fiables. La combinaison de *k-NN* et d'arbre de décision produit un R^2 de 0.88, identique à celui de l'arbre de décision, mais avec un MAE de 2.15, légèrement supérieur à celui de l'arbre de décision. Ce résultat suggère que l'ajout de *k-NN* n'apporte pas d'amélioration significative à la capacité prédictive du modèle et peut même introduire une certaine imprécision supplémentaire, comme l'indique l'augmentation du MAE. Les modèles ensemblistes, le modèle de forêt aléatoire et *CatBoost*, montrent les meilleures performances. Le modèle de forêt aléatoire atteint un R^2 de 0.93 et un MAE de 1.48, tandis que *CatBoost* réalise un R^2 de 0.93 et un MAE de 1.51. Ces résultats soulignent non seulement une excellente capacité de prédiction mais aussi une haute précision, faisant de ces modèles les choix les plus efficaces pour cette analyse. L'adoption de ces méthodes avancées pourrait significativement améliorer la précision des prédictions, ce qui serait crucial pour l'amélioration des stratégies de suivi.

Bien que ces scores puissent initialement suggérer une excellente capacité prédictive, ils soulèvent également des préoccupations potentielles concernant le surapprentissage, un phénomène où le modèle s'adapte trop spécifiquement aux données de l'échantillon d'entraînement au détriment de sa capacité à généraliser sur de nouvelles données. L'analyse des performances des modèles de régression révèle une possibilité de surapprentissage, particulièrement due à la structure unique de l'ensemble de données. Chaque participant est représenté à plusieurs instants, avec trois itérations par instant. Il existe donc un risque non négligeable que certaines de ces itérations se retrouvent dans l'ensemble d'entraînement tandis que d'autres, appartenant

au même participant à un instant similaire, soient placées dans l'ensemble de test. Cette situation pourrait permettre au modèle d'apprendre des détails spécifiques des données d'entraînement qui ne généralisent pas nécessairement à de nouvelles données, mais qui paraissent performantes lorsque des itérations apparentées sont présentes dans l'ensemble de test, bien que les itérations présentent des variations au niveau des aires d'activités musculaires.

Pour adresser cette problématique et assurer une évaluation rigoureuse de la capacité de généralisation des modèles, il était judicieux d'adopter une stratégie où les itérations d'un même participant sont exclusivement dans un seul des ensembles, soit tous en entraînement soit tous en test, à un instant donné. Pour ce faire, une approche garantissant ceci a été développée sur mesure pour cette expérimentation. De plus, cette approche permet de sélectionner 10 patients au hasard. Une fois les patients séléctionnés, une des trois instances est séléctionnée pour chaque patient séléctionné. Les instances séléctionnées constituent l'ensemble de test. Le reste des patients représentent l'ensemble d'entrainement. Cette approche assure que toute information potentiellement transposable d'un ensemble à l'autre via les itérations répétées ne fausse pas l'évaluation de la performance des modèles, ce qui est crucial pour éviter que ces derniers n'apprennent simplement à reconnaître les spécificités des participants plutôt que de capturer des tendances générales applicables à de nouvelles données.

Afin d'évaluer l'impact de notre nouvelle approche sur les performances des modèles, il était important de comparer les résultats obtenus avec et sans cette méthode de séparation par groupe. Un changement significatif dans les performances lors de l'utilisation de cette technique pourrait indiquer que les résultats initiaux étaient effectivement influencés par un biais dû à la répartition des itérations entre les ensembles.

Après l'intégration de la nouvelle approche, l'évaluation des performances des modèles de régression a révélé une variabilité significative des résultats à chaque exécution. Cette

variabilité est principalement attribuable au processus de sélection aléatoire des données, où chaque nouvelle sélection peut mener à des compositions différentes des ensembles de test et d'entraînement, influençant ainsi les performances des modèles en fonction des caractéristiques spécifiques des instances choisies.

Pour illustrer l'impact de la structure de données sur les performances des modèles, 10 exécutions distinctes ont été réalisées, chacune reflétant les variations dues aux différences de composition des ensembles de test et d'entraînement. Pour le modèle de régression linéaire, le R^2 a fluctué entre un minimum de -1.64 (avec un MAE de 9.45) et un maximum de 0.63 (avec un MAE de 4.94), tandis que le MAE a varié entre 2.86 (avec un R^2 de 0.41) et 9.45 (avec un R^2 de -1.64). Lors de l'exécution où le R^2 de la régression linéaire était à son minimum, le modèle *Random Forest* a affiché un R^2 de 0.61, avec un MAE de 3.06, et le modèle *CatBoost* a obtenu un R^2 de 0.58 et un MAE de 3.55. En revanche, lors de l'exécution où le R^2 de la régression linéaire était à son maximum, *Random Forest* a atteint un R^2 de 0.65 avec un MAE de 4.76, tandis que *CatBoost* a montré un R^2 de 0.53 et un MAE de 5.52. Pour le modèle *Random Forest*, le R^2 a varié entre -1.09 et 0.90, avec un MAE allant de 1.92 à 7.37. L'exécution avec le R^2 le plus faible est également celle avec le MAE le plus élevé, tandis que l'exécution avec le R^2 le plus élevé est celle avec le MAE le plus faible. Pour le modèle *CatBoost*, le R^2 minimum est de -0.25 (avec un MAE de 4.29), tandis que le R^2 maximum est de 0.88 (avec un MAE de 2.00). Le MAE varie entre un minimum de 1.86 (avec un R^2 de 0.54) et un maximum de 5.52 (avec un R^2 de 0.53). L'exécution ayant le R^2 maximal pour *Random Forest* est la même que celle ayant le R^2 maximal pour *CatBoost*.

Cette instabilité soulève plusieurs questions importantes concernant la fiabilité et la généralisabilité des modèles. Premièrement, les modèles montrent une sensibilité marquée aux données sur lesquelles ils sont entraînés, ce qui peut indiquer une tendance à capturer des détails spécifiques des données plutôt que des tendances générales applicables. Deuxièmement,

la variation des performances entre les exécutions met en question la reproductibilité des résultats, un pilier central pour notre étude où la fiabilité des prédictions est cruciale.

En outre, la petite taille de notre ensemble de données et son hétérogénéité posent des défis supplémentaires. La petite taille limite la capacité des modèles à apprendre des relations complexes sans surajustement. Dans les ensembles de données de petite taille, chaque donnée a un impact disproportionné sur la construction du modèle, ce qui peut entraîner des performances instables et des variations significatives lors de différentes exécutions. L'hétérogénéité des données, due à la variabilité des caractéristiques cliniques et des mesures EMG parmi les patients, introduit un autre défi. Cette variabilité peut induire des modèles à apprendre des spécificités qui ne sont pas généralisables à l'ensemble de la population cible.

5.4 PRÉDICTION DE LA VITESSE DE MARCHE SUR DIX MÈTRES

Dans le cadre de la prédiction du degré d'atteinte chez les personnes souffrant d'ARSACS, qui affecte principalement la motricité, il est crucial d'évaluer précisément la fonction motrice. La vitesse de marche sur 10 mètres, mesurée à un rythme confortable, est un indicateur largement utilisé de la performance motrice, reflétant l'autonomie et la capacité fonctionnelle des patients. L'objectif principal de cette expérimentation est de développer un modèle de régression qui utilise une combinaison de données d'aires d'activités musculaires et de données cliniques, pour prédire la vitesse de marche lors d'un 10 mètres de marche à une vitesse confortable chez les patients atteints d'ARSACS. Une prédiction fiable de cette vitesse permettrait potentiellement non seulement d'améliorer la compréhension de la progression de la maladie mais aussi de personnaliser et d'optimiser le suivi et la prise en charge. Par exemple, ajuster les protocoles de rééducation motrice pourrait être envisagé en fonction des résultats prédictifs, permettant ainsi une prise en charge plus ciblée et efficace.

L'ensemble de données inclut des mesures recueillies à trois moments différents : T3, T4 et T5, traitées comme des instances indépendantes afin d'augmenter la taille et la variabilité de l'ensemble de données. Bien que cette approche limite la capacité d'analyser des tendances longitudinales pour les individus, elle enrichit notre modèle en fournissant une large base de données transversales, ce qui est crucial pour développer un modèle de régression généralisable capable de prédire la vitesse de marche dans une population hétérogène de patients atteints d'ARSACS. Les variables considérées dans cette expérimentation comprennent les données d'aires d'activités musculaires des muscles biceps fémoral (*biceps femoris*), droit fémoral (*rectus femoris*), droit de l'abdomen (*rectus abdominis*), tenseur du fasia lata (*tensor fascia lata*), grand glutéal (*gluteus maximus*), long adducteur (*adductor longus*) et tibial antérieur (*tibialis anterior*), et ce pour la phase concentrique de l'extension et de la flexion. L'âge et le sexe sont également inclus en raison de leur influence potentielle sur la performance motrice et la progression de l'ARSACS.

Cette expérimentation, dédiée à la prédiction de la vitesse de marche sur dix mètres à une vitesse confortable, se concentre exclusivement sur les participants catégorisés comme marcheurs. La principale raison de cette sélection repose sur la pertinence clinique de l'évaluation de la vitesse de marche, qui ne peut être mesurée de manière fiable que chez les individus ayant la capacité physique de réaliser le test. L'inclusion de non-marcheurs introduirait un biais significatif, rendant les prédictions non seulement imprécises mais également inapplicables, car leurs données ne refléteraient pas une véritable performance de marche. De plus, cette sélection assure une uniformité au sein de l'ensemble des données analysées, car elle se limite aux individus qui peuvent effectivement marcher. Cela garantit que les variations observées dans la vitesse de marche sont réellement indicatives des variations de la condition motrice des participants, et non de leur incapacité à exécuter le test. En conséquence, les modèles développés sont finement adaptés aux particularités de cette population spécifique. Cette

précision améliore la qualité des prédictions et leur applicabilité à un suivi clinique plus précis et adapté, permettant ainsi un meilleur ajustement des stratégies de prise en charge pour les patients atteints d'ARSACS.

Dans le cadre de l'analyse préparatoire des données, les instances contenant au moins une valeur manquante ont été supprimées, représentant seulement 5,62 % de l'ensemble des données. Cette proportion est relativement faible. Comme observé dans les expérimentations précédentes, l'imputation des valeurs manquantes comporte plusieurs risques pour cette étude. Notamment, elle peut réduire la variabilité naturelle des données, diminuant ainsi la sensibilité du modèle aux nuances cruciales dans les performances motrices des patients. Ainsi, l'ensemble final de données utilisé pour cette expérimentation est composé de 252 instances. Bien que la suppression des données ait légèrement réduit le volume de l'échantillon, l'intégrité de l'analyse reste préservée. Cette décision simplifie la préparation des données et garantit que les analyses ultérieures reposent sur des informations fiables et complètes.

Dans le but d'examiner les interdépendances entre les variables indépendantes, une matrice de corrélation a été utilisée, comme illustré à la Figure 5.5. La multicolinéarité a également été évaluée en calculant le facteur d'inflation de la variance pour chaque variable indépendante, représenté à la Figure 5.6. Les résultats de cette analyse révèlent une absence de corrélations significatives entre les variables.

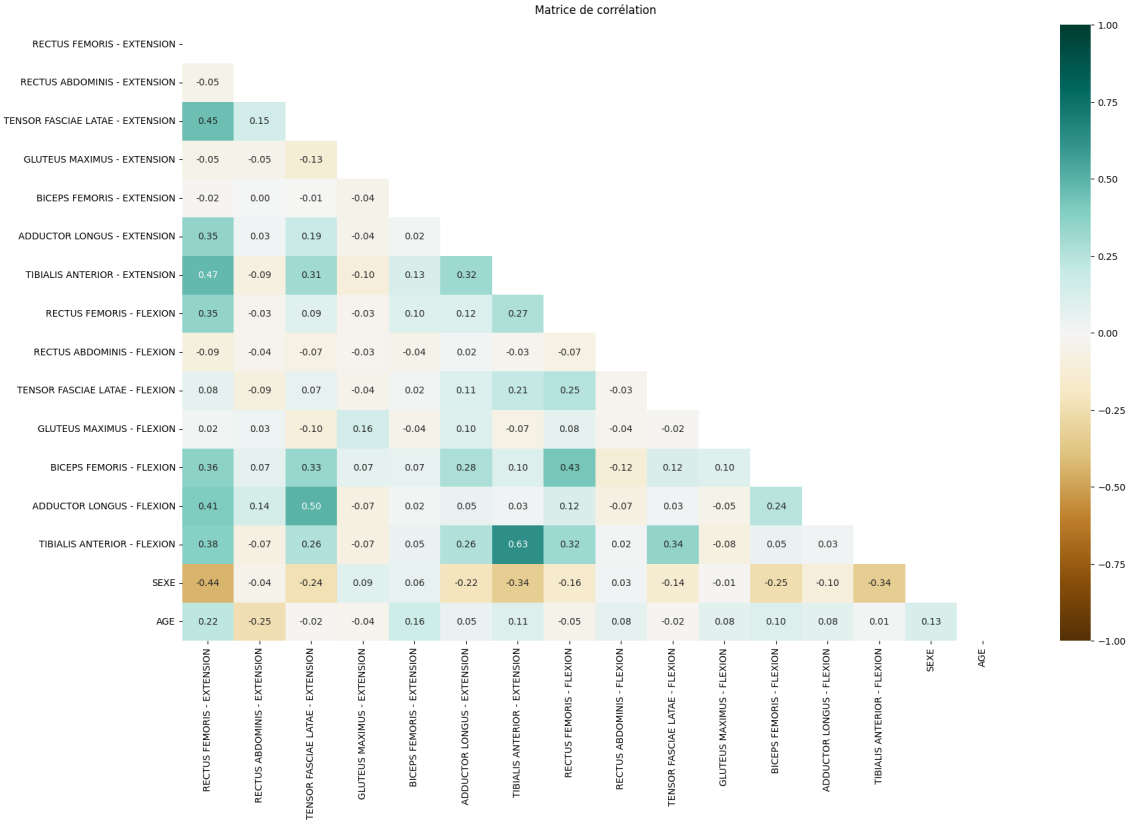

FIGURE 5.5 : Matrice de corrélation.

	variables	VIF
0	RECTUS FEMORIS - EXTENSION	2.394464
1	RECTUS ABDOMINIS - EXTENSION	1.154316
2	TENSOR FASCIAE LATAE - EXTENSION	1.834989
3	GLUTEUS MAXIMUS - EXTENSION	1.091257
4	BICEPS FEMORIS - EXTENSION	1.091485
5	ADDUCTOR LONGUS - EXTENSION	1.300749
6	TIBIALIS ANTERIOR - EXTENSION	2.085109
7	RECTUS FEMORIS - FLEXION	1.660198
8	RECTUS ABDOMINIS - FLEXION	1.045469
9	TENSOR FASCIAE LATAE - FLEXION	1.209020
10	GLUTEUS MAXIMUS - FLEXION	1.094307
11	BICEPS FEMORIS - FLEXION	1.739635
12	ADDUCTOR LONGUS - FLEXION	1.593153
13	TIBIALIS ANTERIOR - FLEXION	2.020971
14	SEXE	1.502999
15	AGE	1.379351

FIGURE 5.6 : Facteur d'inflation de la variance pour chaque variable indépendante.

Comme dans l'expérimentations précédente, la division de l'ensemble de données en ensembles d'entraînement et de test pour cette expérimentation a été effectuée en suivant la même méthode rigoureuse développée sur mesure pour prévenir le surapprentissage. Cette approche, détaillée dans la Section 5.3 de l'expérimentation précédente, implique la sélection aléatoire de 10 patients, avec l'attribution d'une des trois instances de chaque patient sélectionné à l'ensemble de test, tandis que les autres patients forment l'ensemble d'entraînement. Cette méthode garantit que toutes les itérations d'un même participant à un instant donné sont

exclusivement présentes dans un seul des ensembles, assurant ainsi une évaluation fiable de la capacité de généralisation des modèles développés.

Dans cette expérimentation, les performances des modèles de régression ont été évaluées à l'aide des mêmes mesures de performance utilisées dans les expérimentations précédentes, notamment, le coefficient de détermination et l'erreur absolue moyenne.

Les valeurs du coefficient de détermination R^2 varient considérablement d'une exécution à l'autre, reflétant une instabilité dans la capacité des modèles à expliquer la variance des données de test. Par exemple, le R^2 atteint parfois des valeurs élevées, indiquant une bonne adéquation du modèle. Cependant, dans d'autres cas, il chute à des valeurs très basses, voire négatives, ce qui suggère une performance prédictive médiocre. Similairement, le MAE a montré des variations notables, variant dans une plage significative, ce qui met en évidence des écarts dans la précision des prédictions. Ces variations pourraient être attribuées aux différences dans la composition des ensembles de test.

Afin d'illustrer les variations dues aux différences de composition des ensembles de test et d'entraînement, 10 exécutions distinctes ont été réalisées. Le modèle de régression linéaire a montré une grande amplitude dans ses performances, avec un R^2 oscillant entre -0.42 et 0.54, indiquant des fluctuations considérables dans sa capacité à expliquer la variance des données de test. Le MAE pour le modèle de régression linéaire a varié entre 0.20 et 0.28. Pour le modèle d'arbre de décision, le R^2 a varié entre -0.65 et 0.78 avec un MAE allant de 0.15 à 0.33. Le modèle *k-NN* a affiché un R^2 variant de -0.53 à 0.73 et un MAE variant de 0.15 à 0.30. Pour la combinaison de *k-NN* et d'arbre de décision, le R^2 a varié entre -0.18 et 0.68 avec un MAE allant de 0.14 à 0.25. Pour les modèles ensemblistes, le modèle de forêt aléatoire a affiché un R^2 variant de -0.23 à 0.73 et un MAE variant de 0.13 à 0.26. Quant au modèle CatBoost, le R^2 a varié entre -0.05 et 0.81, tandis que le MAE a varié entre 0.12 et 0.23.

Pour illustrer concrètement ces variations, des cas spécifiques de prédictions peuvent être examinés. Par exemple, dans un cas où le modèle performait bien, les prédictions de vitesse de marche se rapprochaient étroitement des valeurs réelles observées, avec des erreurs minimes. En revanche, dans des cas de faible performance, les prédictions déviaient significativement de la réalité, ce qui pourrait indiquer des problèmes de surajustement ou de sous-spécification des modèles.

5.5 CONCLUSION

Dans ce chapitre, nous avons présenté les différentes expérimentations de régression menées, visant à prédire et à estimer la progression de l'ARSACS. Les résultats initiaux montrent une capacité prédictive prometteuse, avec une nette supériorité des modèles ensemblistes, en particulier *CatBoost*, pour modéliser la complexité des données cliniques et EMG. Ces modèles capturent des relations non linéaires importantes, que les méthodes plus simples, comme la régression linéaire, ne parviennent pas à identifier.

Cependant, l'excellence des résultats a soulevé des préoccupations quant au risque de surapprentissage des modèles. En effet, la structure particulière des données, avec plusieurs instances par participant à différents moments, expose les modèles à ce risque. Afin d'évaluer la capacité de généralisation des modèles prédictifs, une approche, présentée dans la section 5.3, a été mise en œuvre pour éviter que des itérations d'un même participant se retrouvent simultanément dans les ensembles d'entraînement et de test. L'évaluation des modèles de régression après l'adoption de cette nouvelle approche a révélé une forte variabilité des résultats, principalement due à la sélection aléatoire des données. Cette instabilité remet en question la fiabilité des modèles, soulignant leur sensibilité aux données d'entraînement et leur manque de reproductibilité.

L'ARSACS est reconnue pour l'hétérogénéité marquée de ses patients, tant en termes de symptômes que de progression de la maladie. Cette grande variabilité nécessite des ensembles de données conséquents afin de capturer les nombreuses nuances et interactions présentes. Cependant, en raison de la rareté de l'ARSACS, il est particulièrement difficile d'obtenir un ensemble de données suffisamment large, ce qui limite la portée des analyses et la performance des modèles de régression.

Nous avions initialement l'espoir que, malgré cette limitation de la taille de l'échantillon, les modèles de régression pourraient tout de même fournir des prédictions satisfaisantes. Malheureusement, les résultats obtenus n'ont pas répondu à nos attentes. Toutefois, il était crucial de mener ces explorations pour confirmer ou infirmer nos hypothèses, ce qui nous a permis de mieux comprendre les limites inhérentes à notre ensemble de données et aux techniques appliquées.

CONCLUSION

Dans cette recherche, nous avons exposé et analysé les différentes expérimentations menées et les résultats obtenus, visant à explorer l'utilisation des techniques d'apprentissage automatique, notamment le *clustering* et la régression, pour identifier des marqueurs prédictifs du degré d'atteinte des personnes souffrant d'ARSACS et comprendre son évolution au fil des années.

L'application du *clustering* a permis d'investiguer si des groupements inédits de profils de patients atteints d'ARSACS étaient présents afin de mettre en évidence des sous-groupes avec des caractéristiques cliniques distinctes. Les sous-groupes identifiés montrent des patterns cliniques variés. En limitant les caractéristiques étudiées à trois muscles, les *clusters* obtenus montrent une séparation claire, suggérant que même avec un sous-ensemble restreint de caractéristiques, des distinctions significatives peuvent être capturées. Cette approche a permis de faciliter les visualisations tridimensionnelles directes et a montré que les caractéristiques choisies étaient particulièrement pertinentes pour discerner des variations cliniques significatives. Bien que les *clusters* soient similaires à ceux obtenus avec l'expérimentation incluant quatre muscles, cette méthode offre une perspective simplifiée et directe des données, sans recours à des techniques de réduction de dimensionnalité. En se concentrant uniquement sur les participants marcheurs et en utilisant deux scénarios distincts (inclusion de toutes les caractéristiques et inclusion de caractéristiques spécifiques), les résultats montrent que dans le premier scénario, incluant toutes les caractéristiques, les *clusters* formés étaient distincts avec des frontières bien définies, démontrant ainsi la capacité de l'ensemble des données à différencier de manière significative les divers profils de participants. En revanche, dans le deuxième scénario, incluant des caractéristiques spécifiques, les *clusters* sont moins distincts, avec une certaine confusion entre les différents degrés d'atteinte au sein de certains *clusters*.

Les caractéristiques sélectionnées pour ce deuxième scénario, bien que discriminantes, ne permettent pas de distinguer clairement tous les états de santé des participants. Cette comparaison souligne l'importance de choisir des caractéristiques qui reflètent non seulement les variations cliniques significatives, mais qui sont aussi suffisamment complètes pour discriminer efficacement les différents degrés d'atteinte. Pour l'ARSACS, il est crucial d'inclure une variété de caractéristiques pertinentes afin de capturer pleinement les distinctions entre les profils des participants.

Les *clusters* obtenus ont été analysés en examinant la répartion des participants selon leurs états de santé au sein des clusters. Malgré la bonne séparation des *clusters* obtenus, ces derniers regroupent plusieurs participants avec des états de santé différents. Certains *clusters* regroupents des participants avec une variabilité importante dans les caractéristiques. Les manifestations cliniques, la progression de la maladie et la prédisposition génétique varient souvent au sein d'individus diagnostiqués avec la même maladie neurodégénérative. Ce mélange des états de santé au sein des clusters pourrait indiquer une similarité des profils EMG parmi des patients de différents états de santé, suggérant que les marqueurs EMG ne sont pas suffisamment discriminants. Cette situation peut être attribuée à la nature de la maladie d'ARSACS, caractérisée par une variabilité intrinsèque où les différences entre les états de santé ne sont pas toujours claires ni linéairement séparables. Cela souligne la complexité qui dépasse la simple catégorisation des états de santé basée sur les caractéristiques EMG, et met en évidence la difficulté de distinguer l'état de santé et le degré d'atteinte des personnes souffrant d'ARSACS. L'hétérogénéité des maladies neurodégénératives complique la compréhension de leurs causes, car des mécanismes variés peuvent être à l'origine de la maladie chez différents individus. Il devient donc de plus en plus courant de regrouper les participants en se basant sur des critères plus précis que ceux d'une simple catégorie,

permettant ainsi de créer des sous-groupes plus homogènes et d'améliorer la précision des analyses de l'étude.

Les modèles de régression testés pour la prédiction du score LEMOCOT incluent la régression linéaire, les arbres de décision, *k-NN*, la combinaison *k-NN*-arbre de décision, la forêt aléatoire, *CatBoost*, et *XGBoost*. Le modèle *CatBoost* a démontré les meilleures performances avec un R^2 de 0.67 et un MAE de 0.08, indiquant une bonne capacité prédictive. Les résultats montrent que les modèles plus complexes et non linéaires comme *CatBoost* et la forêt aléatoire sont plus aptes à capturer les interactions complexes entre les variables EMG et cliniques. La régression linéaire, bien qu'utilisée comme référence, ne peut pas modéliser adéquatement la complexité de l'évolution de l'ARSACS. Les modèles de *boosting* et de forêt aléatoire surpassent nettement les approches plus traditionnelles, mettant en évidence la nécessité d'utiliser des modèles avancés pour une prédiction efficace. *CatBoost*, en particulier, a montré une capacité remarquable à prédire les scores LEMOCOT en tenant compte des non-linéarités et des interactions complexes entre les variables.

L'intégration des scores LEMOCOT à l'instant T comme variables prédictives afin de prédire les scores LEMOCOT à l'instant T+1 a permis d'ajouter une dimension temporelle à l'analyse, ce qui améliore la capacité du modèle à anticiper la progression de la maladie. Les modèles de régression ont montré des performances variées, avec *CatBoost* atteignant un R^2 de 0.92 et un MAE de 1.83. Bien que les scores prédictifs puissent sembler excellents, ils soulèvent des préoccupations d'*overfitting*, où le modèle s'adapte trop précisément aux données d'entraînement, réduisant ainsi sa capacité à généraliser sur de nouvelles données. Ceci est dû à la structure des données, où chaque participant a plusieurs itérations à différents moments. Pour éviter cela, une stratégie spécifique a été développée : à un moment donné, toutes les itérations d'un même participant sont placées soit dans l'ensemble d'entraînement, soit dans l'ensemble de test. Dix patients sont choisis au hasard, et une instance de chaque est

mise dans l'ensemble de test, le reste formant l'ensemble d'entraînement. Après l'intégration de la nouvelle approche, l'évaluation des performances des modèles de régression a montré une grande variabilité des résultats à chaque exécution, principalement due à la sélection aléatoire des données. Cette instabilité remet en question la fiabilité et la généralisabilité des modèles, soulignant leur sensibilité aux données d'entraînement et posant des problèmes de reproductibilité des résultats.

La grande variabilité entre les individus souffrant d'ARSACS, que ce soit en termes de présentation clinique, de gravité des symptômes ou de l'évolution de la maladie, explique en partie les différences de performance entre les modèles. L'ARSACS se manifeste de manière très hétérogène, ce qui complique la généralisation des modèles prédictifs. Chaque patient peut présenter une combinaison unique de symptômes et une progression différente de la maladie, ce qui complexifie la tâche des modèles de régression pour capturer ces variations.

De plus, cette variabilité interindividuelle pourrait également se traduire par une variabilité des données EMG. Les signaux EMG, représentant la dénervation des muscles disatux, essentiels pour les prédictions, peuvent différer considérablement d'un patient à l'autre en raison des différences de fonctionnement musculaire et de la réponse neuromusculaire. Cette hétérogénéité intrapatient et interpatient représente un défi majeur pour le développement de modèles prédictifs robustes et précis. Les variations individuelles doivent être prises en compte dans l'élaboration des algorithmes pour améliorer leur capacité à généraliser et à fournir des prédictions fiables pour différents patients.

Pour les modèles de prédiction de la vitesse de marche sur 10 mètres, les valeurs du R^2 et de MAE varient fortement d'une exécution à l'autre, révélant une instabilité des modèles. Parfois, le R^2 est élevé, indiquant une bonne adéquation du modèle, mais il peut aussi chuter à des valeurs très basses ou négatives, montrant une mauvaise performance prédictive. Ces

variations sont dues aux différences dans la composition des ensembles de test. Lors de bonnes performances, les prédictions de vitesse de marche sont proches des valeurs réelles, tandis que lors de mauvaises performances, les prédictions s'écartent considérablement, suggérant des problèmes de surajustement ou de sous-spécification des modèles.

La variabilité des résultats pour la prédiction de la vitesse de marche peut également être attribuée à la grande variabilité interindividuelle observée chez les patients ARSACS. La différence dans la capacité motrice et l'évolution de la maladie contribuent à la difficulté de prédire de manière uniforme la vitesse de marche.

Un défi majeur de cette étude est la taille limitée de l'échantillon. Un petit nombre de participants peut réduire la robustesse des modèles et la généralisation des résultats, limitant leur capacité à apprendre et à identifier des tendances fiables, ce qui peut aggraver les effets de la variabilité individuelle. De plus, les données ne couvrent qu'un nombre restreint de temps (T3, T4, T5), ce qui limite la possibilité de réaliser des analyses longitudinales robustes.

Les recherches futures devraient viser à augmenter la taille de l'échantillon et à inclure des données recueillies sur des périodes plus longues pour améliorer la robustesse et la précision des modèles prédictifs. De plus, l'intégration de techniques d'apprentissage automatique plus avancées, telles que les réseaux de neurones profonds, pourrait améliorer les capacités de prédiction, comme l'ont démontré Rezaee et al. dans leur étude sur l'utilisation de structures d'apprentissage profond par transfert (*deep transfer learning*), pré-entraînées et combinées à des modèles d'apprentissage automatique conventionnels pour diagnostiquer la maladie de Parkinson à partir de signaux électromyographiques de surface des membres supérieurs, atteignant une justesse de plus de 99%, et ce, avec un traitement minimal des caractéristiques.

Pour enrichir les analyses, il serait également bénéfique d'explorer l'intégration de données supplémentaires, telles que des données d'imagerie, incluant des vidéos des exercices

effectués par les patients. En ce qui concerne les vidéos des exercices, il serait pertinent de se concentrer sur les mêmes exercices pour lesquels les données EMG sont enregistrées. De cette manière, les données seraient synchronisées, permettant de créer un ensemble de données multimodal. Cette approche offrirait une vue d'ensemble plus complète des mouvements et de la coordination des patients, tout en permettant de corréler les données visuelles avec les signaux EMG pour une analyse plus approfondie et une modélisation plus précise.

En conclusion, cette étude explore le potentiel des techniques d'apprentissage automatique et met en évidence ses limites actuelles pour prédire la progression de l'ARSACS, en raison de la taille restreinte de l'échantillon et de l'hétérogénéité clinique des patients. Bien que des progrès aient été réalisés, il est clair que des études futures nécessiteront un nombre de données beaucoup plus important afin de surmonter ces défis et d'améliorer la robustesse des modèles. La contribution majeure de cette recherche réside dans l'application novatrice de l'apprentissage automatique aux données EMG pour une maladie encore peu étudiée dans ce contexte, ouvrant ainsi de nouvelles perspectives pour la recherche et la pratique clinique.

BIBLIOGRAPHIE

[1] J. Goodgold et A. Eberstein, *Electrodiagnosis of Neuromuscular Disease*, 2nd éd. Baltimore : Williams & Wilkins, 1978.

[2] M. De Braekeleer, F. Giasson, J. Mathieu, M. Roy, J.-P. Bouchard, et K. Morgan, "Genetic epidemiology of autosomal recessive spastic ataxia of charlevoix-saguenay in northeastern quebec," *Genetic Epidemiology*, vol. 10, n° 1, pp. 17–25, 1993.

[3] J. C. Engert, P. Bérubé, J. Mercier, C. Doré, P. Lepage, B. Ge, J. P. Bouchard, J. Mathieu, S. B. Melançon, M. Schalling, E. S. Lander, K. Morgan, T. J. Hudson, et A. Richter, "Arsacs, a spastic ataxia common in northeastern québec, is caused by mutations in a new gene encoding an 11.5-kb orf," *Nature Genetics*, vol. 24, n° 2, pp. 120–125, Feb 2000.

[4] J. P. Bouchard, A. Barbeau, R. Bouchard, et R. W. Bouchard, "Electromyography and nerve conduction studies in friedreich's ataxia and autosomal recessive spastic ataxia of charlevoix-saguenay (arsacs)," *Canadian Journal of Neurological Sciences*, vol. 6, n° 2, pp. 185–189, May 1979.

[5] J. P. Bouchard, A. Richter, J. Mathieu, D. Brunet, T. J. Hudson, K. Morgan, et S. B. Melançon, "Autosomal recessive spastic ataxia of charlevoix-saguenay," *Neuromuscular Disorders*, vol. 8, n° 7, pp. 474–479, October 1998.

[6] C. Gagnon, I. Lessard, C. Lavoie, I. Côté, R. St-Gelais, J. Mathieu, et B. Brais, "An exploratory natural history of ataxia of charlevoix-saguenay : A 2-year follow-up," *Neurology*, vol. 91, n° 14, pp. e1307–e1311, 2018, epub 2018 Aug 29.

[7] R. Merletti et P. Parker, *Electromyography : Physiology, Engineering, and Non-invasive Applications*. John Wiley & Sons, 2004.

[8] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, n° 3, pp. 210–229, 1959.

[9] M. A. Myszczynska, P. N. Ojamies, A. M. B. Lacoste, D. Neil, A. Saffari, R. Mead, G. M. Hautbergue, J. D. Holbrook, et L. Ferraiuolo, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nature Reviews Neurology*, vol. 16, n° 8, pp. 440–456, August 2020, epub 2020 Jul 15.

[10] H. J. Hermens, B. Freriks, C. Disselhorst-Klug, et G. Rau, “Development of recommendations for semg sensors and sensor placement procedures,” *Journal of Electromyography and Kinesiology*, vol. 10, n° 5, p. 361–374, Oct 2000.

[11] J. V. Basmajian et C. J. De Luca, *Muscles Alive : Their Functions Revealed by Electromyography*, 5e éd. Baltimore : Williams & Wilkins, 1985.

[12] S. Ödman et P. Öberg, “Movement-induced potentials in surface electrodes,” *Medical and Biological Engineering and Computing*, vol. 20, pp. 159–166, 1982.

[13] L. A. Geddes et L. E. Baker, *Principles of Applied Biomedical Instrumentation*. New York : Wiley, 1968.

[14] H. J. Hermens, B. Freriks, R. Merletti, D. F. Stegeman, J. H. Blok, G. Rau, C. Disselhorst-Klug, G. Hägg, I. H. J. B. Hermens, et Freriks, “European recommendations for surface electromyography : Results of the seniam project,” 1999. [En ligne]. Repéré à : https://api.semanticscholar.org/CorpusID:114598925

[15] D. Farina, R. Merletti, et R. M. Enoka, “The extraction of neural strategies from the surface emg,” *Journal of Applied Physiology*, vol. 96, n° 4, pp. 1486–1495, 2004.

[16] C. J. De Luca, “The use of surface electromyography in biomechanics,” *Journal of Applied Biomechanics*, vol. 13, n° 2, pp. 135–163, 1997.

[17] C. Richards, J. P. Bouchard, R. Bouchard, et H. Barbeau, “A preliminary study of dynamic muscle function in hereditary ataxia,” *Canadian Journal of Neurological Sciences*, vol. 7, n° 4, pp. 367–377, 1980.

[18] Y. Jiang, C. Chen, X. Zhang, C. Chen, Y. Zhou, G. Ni, S. Muh, et S. Lemos, “Shoulder muscle activation pattern recognition based on semg and machine learning algorithms,” *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105721, 2020. [En ligne]. Repéré à : https://doi.org/10.1016/j.cmpb.2020.105721

[19] A. Sadiq, S. G. Khawaja, M. U. Akram, N. S. Alghamdi, A. Khan, et A. Shaukat, “Machine learning and signal processing based analysis of semg signals for daily action classification,” *IEEE Access*, vol. 10, 2022. [En ligne]. Repéré à :

https://doi.org/10.1109/ACCESS.2022.3166885

[20] U. Kleinholdermann, M. Wullstein, et D. Pedrosa, “Prediction of motor unified parkinson’s disease rating scale scores in patients with parkinson’s disease using surface electromyography,” *Clin Neurophysiol*, vol. 132, n° 7, pp. 1708–1713, Jul 2021, epub 2021 Mar 13.

[21] M. Ferreira, F. Barbieri, V. Moreno, T. Penedo, et J. Tavares, “Machine learning models for parkinson’s disease detection and stage classification based on spatial-temporal gait parameters,” *Gait & Posture*, vol. 98, pp. 49–55, Oct 2022, epub 2022 Aug 20.

[22] F. H. M. Oliveira, A. R. P. Machado, et A. O. Andrade, “On the use of t-distributed stochastic neighbor embedding for data visualization and classification of individuals with parkinson’s disease,” *Computational and Mathematical Methods in Medicine*, vol. 2018, p. 8019232, November 2018. [En ligne]. Repéré à : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6247646/

[23] T. Schmitz-Hübsch, S. Du Montcel, L. Baliko, J. Berciano, S. Boesch, C. Depondt, P. Giunti, C. Globas, J. Infante, J.-S. Kang *et al.*, “Scale for the assessment and rating of ataxia : development of a new clinical scale,” *Neurology*, vol. 66, n° 11, pp. 1717–1720, 2006.

[24] L. L. Abeysekara, B. Kashyap, C. Kolambahewage, P. N. Pathirana, M. Horne, et D. J. Szmulewicz, “A study of upper-limb motion using kinematic measures for clinical assessment of cerebellar ataxia,” dans *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2023, pp. 1–5.

[25] I. K. A. Purnawan, A. Dharma Wibawa, W. Caesarendra, et M. H. Purnomo, “Enhancing neuromuscular disease diagnosis through pca-svm analysis of emg signals : A classification approach,” dans *IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society*, 2023, pp. 1–6.

[26] E. Anselmino, A. Mazzoni, et S. Micera, “Emg-based prediction of step direction for a better control of lower limb wearable devices,” *Comput Methods Programs Biomed*, vol. 254, p. 108305, 2024, epub ahead of print.

[27] F. Castelli Gattinara Di Zubiena, G. Menna, I. Mileti, A. Zampogna, F. Asci, M. Paoloni,

A. Suppa, Z. Del Prete, et E. Palermo, “Machine learning and wearable sensors for the early detection of balance disorders in parkinson’s disease,” *Sensors (Basel)*, vol. 22, n° 24, p. 9903, Dec 2022.

[28] S. Butterworth, “On the theory of filter amplifiers,” *Wireless Engineer*, vol. 7, n° 6, pp. 536–541, 1930.

[29] P. Konrad, *The ABC of EMG : A Practical Introduction to Kinesiological Electromyography*. Scottsdale, AZ, USA : Noraxon, 2005.

[30] A. Savitzky et M. J. Golay, “Smoothing and differentiation of data by simplified least squares procedures,” *Analytical Chemistry*, vol. 36, n° 8, pp. 1627–1639, 1964.

[31] R. L. Drake, W. Vogl, et M. A. W. M., *Gray’s Anatomie pour les étudiants*. Elsevier, 2015.

[32] I. Lessard, R. St-Gelais, L. J. Hébert, I. Côté, J. Mathieu, B. Brais, et C. Gagnon, “Functional mobility in walking adult population with ataxia of charlevoix-saguenay,” *Orphanet Journal of Rare Diseases*, vol. 16, n° 1, 2021.

[33] K. Berg, “Measuring balance in the elderly : Development and validation of an instrument,” Ph.D. Thesis, McGill University, Montreal, Canada, 1992.

[34] R. W. Bohannon et A. Williams Andrews, “Normal walking speed : A descriptive meta-analysis,” *Physiotherapy*, vol. 97, n° 3, p. 182–189, 2011.

[35] J. Desrosiers, A. Rochette, et H. Corriveau, “Validation of a new lower-extremity motor coordination test,” *Archives of Physical Medicine and Rehabilitation*, vol. 86, n° 5, p. 993–998, 2005.

[36] C. Gagnon, B. Brais, I. Lessard, C. Lavoie, I. Côté, et J. Mathieu, “Development and validation of a disease severity index for ataxia of charlevoix-saguenay,” *Neurology*, vol. 93, n° 16, 2019.

[37] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow*, 2nd éd.

Sebastopol, CA : O'Reilly Media, 2019.

[38] M. H. Kutner, C. J. Nachtsheim, J. Neter, et W. Li, *Applied Linear Statistical Models*, 5e éd. McGraw-Hill Irwin, 2005.

[39] L. Breiman, J. Friedman, C. Stone, et R. Olshen, *Classification and Regression Trees*. Taylor & Francis, 1984. [En ligne]. Repéré à : https://books.google.ca/books?id=JwQx-WOmSyQC

[40] T. Hastie, R. Tibshirani, et J. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, 2nd éd. Springer, 2009.

[41] L. Breiman, "Random forests," *Machine learning*, vol. 45, n° 1, pp. 5–32, 2001.

[42] A. V. Dorogush, V. Ershov, et A. Gulin, "Catboost : gradient boosting with categorical features support," 2017.

[43] T. Chen et C. Guestrin, "Xgboost : A scalable tree boosting system," dans *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA : Association for Computing Machinery, 2016, p. 785–794. [En ligne]. Repéré à : https://doi.org/10.1145/2939672.2939785

[44] K. Rezaee, S. Savarkar, X. Yu, et J. Zhang, "A hybrid deep transfer learning-based approach for parkinson's disease classification in surface electromyography signals," *Biomedical Signal Processing and Control*, vol. 71, p. 103161, 2022. [En ligne]. Repéré à : https://www.sciencedirect.com/science/article/pii/S1746809421007588

CONSIDÉRATION ÉTHIQUE

Ce mémoire a fait l'objet d'une certification éthique auprès du CER-UQAC. Le numéro du certificat est 2024-1638.